General-to-Specific Approaches for Evaluating Multi-Step System Forecasts



Andrew B. Martinez
Magdalen College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Economics
Trinity 2019

Contents

Acknowledgements **iii**

Abstract **iv**

Statements of Publication **v**

1 Introduction with an Application to Evaluating Fiscal Forecasts **1**

1 Introduction 2
2 Literature Review 3
3 Data 4
3.1 Description 5
3.2 Graphical Analysis 7
4 Alternative Metrics for Comparing Forecasts 8
5 Structure of the Thesis 19

2 Evaluating Multi-Step System Forecasts with Relatively Few Observations **23**

1 Introduction 24
2 Existing Methods to Evaluate Forecast Accuracy 25
2.1 An Invariant Measure of Forecast Accuracy 27
2.2 Limitations of the GFESM 31
3 Methodological Approach 33
4 Simulations 39
4.1 Eigenvalues 41
4.2 The GFESM 42
5 Comparing Performance Across Forecast Models 44
5.1 Simulation Results 47
5.2 Ranking the Forecast Models 48
6 Allowing for Structural Breaks 50
7 Evaluating a Vector of Forecasts for the US Economy 53
8 Conclusions 55
A Appendix: Estimating the VEqCM 57
A.1 Data Sources and Description 57
A.2 A Forecast Model of the US Economy 59
A.3 An Initial Assessment of Forecast Performance 64

3 Testing for Differences in Path Forecast Accuracy **67**
1 Introduction . . . 68
2 Measuring Path Forecast Accuracy . . . 71
2.1 The Joint Density as a General Loss Function . . . 72
3 A General Path Forecast Accuracy Test . . . 74
3.1 Comparisons with Other Joint Multi-Horizon / Multivariate Approaches . . . 78
3.2 Some Comments on Nested Path Forecasts . . . 79
4 A Special Case of the Path Forecast Accuracy Test . . . 80
5 Simulations . . . 83
5.1 The Path Forecast-Error Generating Process . . . 83
5.2 Null Rejection Frequency . . . 87
5.3 Non-Null Rejection Frequency . . . 88
5.4 Extensions to Large H and K . . . 89
6 Evaluating FRB Greenbook and DSGE Model Path Forecasts . . . 90
7 Evaluating Hurricane Irma's Path Forecast Accuracy . . . 99
8 Conclusions . . . 105
A Technical Appendix . . . 107
A.1 Proof of Theorem 2 . . . 107
A.2 Proof of Corollary 2 . . . 112

4 The Impact of Forecast Accuracy on Hurricane Damages **113**
1 Introduction . . . 114
2 Natural Disasters: Damages and Imperfect Forecasts . . . 117
3 Theoretical Framework . . . 119
4 Application: Short-term Hurricane Forecasts and Damages . . . 124
4.1 Methods . . . 126
4.2 Data . . . 129
5 A Model of Hurricane Damages . . . 135
6 Robustness of the Results . . . 139
6.1 Model Misspecification . . . 140
6.2 Alternative Measures of Forecast Accuracy and Uncertainty . . . 143
6.3 Controlling for Potentially Omitted Variables . . . 144
6.4 Out-of-Sample Fit . . . 147
7 The Value of Improved Hurricane Forecast Accuracy . . . 149
8 Conclusions . . . 152
A Appendix: Additional Figures and Tables . . . 154
B Technical Appendix . . . 159
B.1 Approximation of the Normal Distribution . . . 159
B.2 Counterfactual Policy Analysis . . . 160

5 Conclusions and Future Work **162**

References **167**

Acknowledgements

My deepest gratitude goes to my supervisors David F. Hendry and Jennifer L. Castle. Their unwavering support kept me going despite repeated failures. Their comments and feedback also helped push my work to a level higher than I could have ever hoped to achieve on my own. Special thanks also to Neil R. Ericsson for introducing me to the world of forecasting, for encouraging me to apply to and to take a leap of faith and come to Oxford, and for additional guidance, feedback, and support throughout my DPhil.

The DPhil would have been impossible without funding from Climate Econometrics, the Robertson Foundation, and the Department of Economics. My research was also supported in part through work that I did as a Research Assistant for Professor Sir David F. Hendry and Professor Sophocles Mavroeidis as well as through my teaching in Oxford, Havana, London and at various summer schools.

My consulting work for the International Monetary Fund in Monrovia as well as my summers at the Federal Reserve Bank of Cleveland and Goldman Sachs International in London provided additional financial support and were another source of feedback for my research. Thanks in particular to Todd Clark for making my stay at the Federal Reserve Bank of Cleveland such a valuable and productive experience.

Thanks also to the participants of the Gorman workshops, econometrics lunch seminars, Federal Reserve Bank of Cleveland brownbag seminars, VU Amsterdam lunch seminars, GWU brownbag seminar on forecasting, and conferences including: Dynamic Econometrics (a.k.a. OxMetrics), International Symposium on Forecasting, Royal Economic Society Meetings, Computational and Financial Econometrics, Econometric Models of Climate Change, and the International Summit on Hurricanes and Climate Change.

At Oxford I benefited from close interactions with fellow classmates in particular Seungyoon, Maria, Gregorio, Jonas, Xiyu, Kevin, Patrick, Eileen, and Alessio. Living in Magdalen College and walking around the deer park and gardens proved to be an ideal location to work through thoughts and problems and to be inspired. Coming to Nuffield College each day provided much needed consistency, a great source of nutrition, intellectual stimulation, and motivation. Despite my regular grumblings I am grateful to my office mates Felix, Angela, and Luke for putting up with me.

This dissertation is dedicated to my family: Jack, Susannah, Sebastian and Colin and to my closest friends: Familie Kruse, Chuck, Jennifer, Julie and Joe. Your encouragement, patience, and understanding over the years made this project possible and for the most part enjoyable. Thank you all. I can confidently predict that without you my path would have been very different.

Abstract

Multi-horizon forecasts from large scale models play important roles in ongoing policy debates across a wide range of disciplines. Although there are many ways to evaluate imperfect forecasts, we show the importance of taking a more general approach and considering the joint distributions over variables and horizons. Once a multi-step system context is envisaged, potential singularities need to be addressed, as do tests of significant differences between the forecasts and assessments of the value of improved forecasts. These important issues are the focus of the thesis and are illustrated using applications in the fields of macroeconomics and climate.

Statement of Publication and Co-Authorship

Some chapters have been published or are currently under review at peer-reviewed journals. Most of the work has also been presented at seminars and conferences. Some chapters were written as co-authored papers. The following section provides an overview of publication and co-authorship.

Chapter 1

Chapter 1 extends a forthcoming paper co-written with Neil R. Ericsson (Federal Reserve Board and The George Washington University, Department of Economics):

Ericsson, N. R. and Martinez, A. B. "Evaluation of Government Budget Forecasts", forthcoming in *Handbook of Government Budget Forecasting*. D. Williams and T. Calabrese eds. Palgrave Macmillan.

I collected the data and the forecasts and conducted the literature review. The work associated with the loss functions and the illustrations is completely my own and does not overlap with our joint work. I am grateful to Jennifer L. Castle, Michael P. Clements, David F. Hendry, Neil R. Ericsson, Jaime Marquez and Dan Williams for helpful comments and suggestions.

Chapter 2

Chapter 2 has been published and originated out of my MPhil Thesis for the Degree of Master of Philosophy in Economics at the University of Oxford (awarded in Trinity Term 2015). The published paper is co-authored with David F. Hendry and has appeared as:

Hendry, D.F. and Martinez, A.B. "Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations", *International Journal of Forecasting* Vol. 33, 2 (2017), pp. 359-372.

David Hendry provided the original research design as well as considerable input at all stages of the process. I was responsible for the writing, the simulations and the empirical application. Early versions of this paper were presented at the 16th OxMetrics Users Conference and the 36th International Symposium on Forecasting. I am grateful to Karim M. Abadir, Jennifer L. Castle, Jurgen A. Doornik, David F. Hendry, Søren Johansen, Michael McCracken, Bent Nielsen, Felix Pretis and two anonymous reviewers for helpful comments and suggestions.

Chapter 3

This chapter is single-authored. An earlier version of the chapter is available as a Federal Reserve Bank of Cleveland Working Paper:

Martinez, A.B. "Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter", Federal Reserve Bank of Cleveland Working Paper No. 17-17, November 2017.

Early versions of which were presented at the 17th OxMetrics Users Conference, the Oxford Econometrics Lunch seminar, the 37th International Symposium on Forecasting, multiple Federal Reserve

Bank of Cleveland brown bag seminars, the VU Amsterdam Econometrics Lunch Seminar and as a poster at the 2nd Conference on Forecasting at Central Banks. I am grateful to Jennifer L. Castle, Todd E. Clark, Michael P. Clements, David F. Hendry, Ryoko Ito, Xiyu Jiao, Bent Nielsen, Felix Pretis, Kevin Sheppard, the participants at the 17th OxMetrics conference, the Oxford Econometrics Lunch Seminar, and multiple Federal Reserve Bank of Cleveland brown bag seminars for helpful comments and suggestions.

Chapter 4

This chapter is single-authored. An earlier version of the chapter is available as an Oxford University, Department of Economics Discussion Paper:

Martinez, A.B. "A False Sense of Security: The Impact of Forecast Uncertainty on Hurricane Damages", University of Oxford Department of Economics Discussion Paper No. 831, October 2018.

Early versions of which were presented at the 1st Conference on Econometric Models of Climate Change, the 6th International Summer on Hurricanes and Climate Change, Computational and Financial Econometrics, the 2017 RES PhD meetings, the VU Amsterdam Institute for Environmental Studies lunch seminar, the Gorman workshop, the DPhil / Postdoc lunch seminar, 38th International Symposium on Forecasting, the 20th OxMetrics Users Conference, the Oxford Econometrics lunch seminar and the GWU brown bag seminar on forecasting. I am grateful to Laura Bakkensen, Maria Balgova, Nadia Bloemendaal, Jennifer L. Castle, Todd E. Clark, Tatyana Deryugina, Jurgen Doornik, Neil R. Ericsson, David F. Hendry, Luke P. Jackson, Felix Pretis and Robert G. Williams for helpful comments and suggestions.

Chapter 5

This chapter is single-authored.

Additional Material

Not all work conducted during my DPhil has been included in the final thesis. These projects, however, are closely linked to the work presented here. They include work on testing for invariance, which resulted in two papers:

Castle, J.L and Hendry, D.F. and Martinez, A.B. "Evaluating Forecasts, Narratives and Policy using a Test of Invariance", Econometrics, 2017, 5(3), 39.

Martinez, A.B. "A Note of Caution on Instrumental Variables, Exogeneity, and Causal Inference in Time Series", submitted.

As well as a variety of other, as yet unpublished, papers on a diverse range of topics loosely connected to this thesis.

Word Count

This thesis comprises 190 pages (including forematter and references) with an average page containing around 418 words, resulting in an approximate total word count of 79,002 words. Alternative software word counts put the total at around 55,000 words excluding the references.

Chapter 1

Introduction with an Application to Evaluating Fiscal Forecasts

This chapter explains the background to the thesis which concerns multi-horizon forecasts from large scale models. These play important roles in policy debates across a number of disciplines. To illustrate the issues and the current evaluation approaches, we first review the literature on the evaluation of government budget forecasts, and illustrate forecast evaluation with an empirical analysis of different U.S. government agencies' forecasts of the U.S. gross federal debt and deficit. We then describe several forecast evaluation metrics as well as their strengths and weaknesses in this context. The chapter concludes with an overview of the rest of the thesis.

This chapter builds on the following publication:

Ericsson, N. R. and Martinez, A. B. "Evaluation of Government Budget Forecasts", forthcoming in *Handbook of Government Budget Forecasting*. D. Williams and T. Calabrese eds. Palgrave Macmillan.

1 Introduction

Multi-horizon forecasts from large scale models play important roles in ongoing policy debates. Macroeconomic policy decisions generally rely on forecasts from large scale dynamic stochastic general equilibrium models and vector autoregression models, which are used by central banks; see Bårdsen et al. (2012) and Burgess et al. (2013) among others. Decisions about natural hazards and emergency responses to weather disasters rely on predictions from dynamical multi-level physics models, whereas longer-term climate policy is guided by predictions from large scale General Circulation Models; see Pretis and Roser (2017). The forecasts produced by these models influence the decisions made by governments, businesses, and individuals.

The widespread proliferation and influence of large scale and multi-step forecasting systems in ongoing policy decisions across a range of disciplines illustrates the importance of econometric methods and approaches that can evaluate, improve, and assess the impacts of these forecasts. However, to date much of the literature has focused on evaluating univariate forecasts at individual horizons. As a result, despite their importance, there are far fewer methods for evaluating forecasting systems across multiple horizons. This implies that the amount that we can learn from and improve upon these forecasts is relatively limited to the methods that are available.

In this thesis I seek to extend the econometric tools that are available for evaluating large scale forecasting systems across multiple horizons. In doing so, I follow a general-to-specific approach. In this sense, I start by thinking about multi-step and multivariate accuracy from a systems perspective. This is most clearly embodied in the general measure of forecast accuracy, originally proposed by Clements and Hendry (1993), known as the general forecast-error second-moment. This provides an initial overall view of the system's forecast accuracy across all variables and horizons. It serves as a basis for understanding how all variables and horizons contribute to the overall performance of the system. It also allows us to then assess where improvements in individual variables or horizons might be most needed or would be the most beneficial for the overall forecasting system.

It is useful to start by considering an application of multi-step system forecasts in order to motivate the methods and the analysis in this thesis. This chapter focuses on U. S. government agency forecasts of the federal debt and deficit. Government budgets have attracted considerable attention, especially with federal debt limits, sequestration, and federal government shut-downs in the United States and continuing discussions about national debt limits in the Euro Area.

For example, in the United States, increasing government debt despite a lasting economic expansion has raised questions about the role of fiscal policy. In 2018, the U.S. gross federal government debt outstanding was 104 percent of GDP while the budget deficit was 3.8 percent of GDP. This has prompted concerns that the U.S. debt burden will become unsustainable. Furthermore, the actual debt and its forecasts have spurred a debate over the role of the debt and deficits during periods of economic growth as embodied in the recent discussions of Modern Monetary Theory. These discussions imply that it is important for policy makers to understand how well we are able to predict the long-term evolution of both the debt and deficits.

This chapter starts by reviewing the literature on the evaluation of government budget agency forecasts. Section 3 describes the data and forecasts used in the empirical illustrations. Section 4 considers various metrics for comparing alternative forecasts and applies them to evaluate forecasts of the debt and deficit. Finally, section 5 lays out the structure of the rest of the thesis.

2 Literature Review

A large body of literature evaluates government budget forecasts. The current section focuses exclusively on forecasts from the U.S. federal budget agencies and includes both economic and budget forecasts. Existing studies can be divided into two types. The first compares agencies' forecasts by looking at the forecast errors directly or with summary statistics such as the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE). The second type uses regression results and tests derived therefrom to compare the forecasts. Both types provide valuable information about the forecasts.

Studies of the first type date back to Kamlet et al. (1987), who use measures of bias and MAPE to evaluate forecasts from the Congressional Budget Office (CBO), the Office of Management and Budget (OMB), their own model, and the ASA/NBER survey. McNees (1995) relies on measures of the mean absolute errors and the root mean squared errors for forecasts from the Federal Reserve Board (FRB), the CBO, the Council of Economic Advisors (CEA), and several private forecasters. Frendreis and Tatalovich (2000) examine forecast bias and its magnitude and direction in CBO, OMB, and FRB economic forecasts. CBO conducts its own semi-annual comparison of its economic forecasts, along with OMB and Blue Chip Consensus forecasts, most recently in CBO (2017a). CBO (2015b) and CBO (2017b) evaluate the fiscal forecasts.

Studies of the second type use regressions to evaluate and compare the forecasts. For example, Howard (1987) regresses the OMB's forecast errors on an intercept and the CBO's forecast errors to understand how the two forecasts are related to one another. Belongia (1988) regresses the actual growth rate of economic variables on an intercept and the CBO, CEA, and private-sector predicted growth rates in order to see which forecast outperforms the others. Cohen and Follette (2003) regress the actual budget deficit on OMB, CBO, and FRB forecasts in order to test which forecast contains the most information content. Krause and Douglas (2005) run similar tests on CBO, OMB, and FRB economic and budget forecasts. In a related vein, Corder (2005) uses regression tests to evaluate forecasts from the Social Security Administration, CBO, and OMB, and asks whether the forecasts could be improved if they incorporate information from each other.

Both types provide information on the relative performance of the forecasts across different samples, variables, and metrics. Table 2.1 provides an overview of these studies.

3 Data

Empirical examples illustrate the different forecast evaluation methods considered below. Section 3.1 describes the forecasts in those examples, which are all of the gross U.S. federal debt and deficit. Section 3.2 provides a graphical perspective as a prelude to the numerical illustrations.

Table 2.1: Previous Evaluations of U.S. Fiscal Agency Forecasts

Study	Forecaster			Forecast	Variable forecast					Forecast
	CBO	OMB	Other	horizon	B	Y	Δp	U	X	Period
CBO (various)	•	•	•	2, 5		•	•		•	various
Kamlet et al. (1987)	•	•	•	1-6		•	•	•		1976-1984
Howard (1987)	•	•		1		•	•	•	•	1976-1985
Plesko (1988)	•	•		1-5	•	•				1974-1988
Belongia (1988)	•	•	•	1		•	•	•		1976-1987
Miller (1991)	•			1	•	•				1980-1987
Blackley and DeBoer (1993)		•		1	•	•	•	•	•	1963-1989
Auerbach (1995)		•		1	•					1982-1993
Campbell and Ghysels (1995)		•		1	•					1967-1990
McNees (1995)	•	•	•	1-4		•	•	•		1972-1994
Auerbach (1999)	•	•	•	1-11	•					1986-1999
Frendreis and Tatalovich (2000)	•	•	•	1		•	•	•		1979-1997
Kliesen and Thornton (2001)	•			1, 5	•					1981-2000
Lipford (2001)	•	•		1-5	•	•	•	•		1980-1999
Penner (2001, 2002)	•			1, 5	•					1980-2000
Kitchen (2003)	•	•		1-5	•					1982-2001
Cohen and Follette (2003)	•	•	•	1	•					1977-2003
Krause and Douglas (2005)	•	•	•	1	•	•	•	•		1976-2001
Corder (2005)	•	•	•	1-5		•	•	•	•	1976-2003
Krause and Douglas (2006)	•	•		1	•					1947-2001
Penner (2008)	•			1, 2, 5	•					1983-2005
Huntley and Miller (2009)	•		•	1-5		•	•	•	•	1993-2003
Kliesen and Thornton (2012)	•		•	1, 5	•					1976-2010
Krol (2014)	•	•	•	2, 5		•				1976-2008
CBO (2015b)	•	•		1-6	•	•				1982-2014
Martinez (2011, 2015)	•	•	•	1,5	•					1984-2013
Tsuchiya (2016)	•	•	•	1-4		•			•	1984-2012
CBO (2017b)	•	•		1, 6	•				•	1982-2014
Ericsson (2017)	•	•	•	1	•					1984-2012
Croushore and Van Norden (2017)	•		•	1	•	•		•		1967-2010
Croushore and Van Norden (2018)	•		•	1	•			•		1981-2010

Notes: "CBO (various)" denotes CBO (2002), CBO (2004), CBO (2005), CBO (2006), CBO (2007), CBO (2009), CBO (2010), CBO (2013), CBO (2015a), and CBO (2017a), which examine forecast periods from 1976 through (respectively) 2000, 2003, 2004, 2005, 2006, 2008, 2009, 2010, 2012, and 2014. "Other" forecasters include the FRB, CEA, SSA, APB, SPF, Blue Chip Consensus, the ASA/NBER survey, and various private sector sources. Forecast horizons are in years. Variables forecast are the budget and related items (B), gross domestic product (Y), inflation (Δp), unemployment (U), and other economic variables (X).

3.1 Description

The primary variables of interest are the U.S. gross federal debt outstanding and the U.S. federal deficit in billions of dollars from 1984 to 2018. The data are published by the U.S. Department of the Treasury's Bureau of the Fiscal Service in the December issue of the Treasury Bulletin and in its Monthly Treasury Statement, and are measured on a fiscal year basis ending September 30th. In public discourse, discussion of the deficit often overshadows discussion of the debt, since the deficit is commonly thought of as equaling the change in debt. However, the deficit and the change in debt are not equal. The deficit excludes certain items that contribute to the change in debt: the crisis-

era Troubled Asset Relief Program, changes in cash balances held by the Treasury, and other direct lending programs by the federal government.

The forecasts are released at the beginning of the calendar year–several months apart between January and March–and are of the U.S. gross federal debt and deficit at the end of the (then) current fiscal year ($h = 1$) and of future fiscal years out to ten-years-ahead ($h = 10$). They are denoted as:

- CBO (Congressional Budget Office) from its Budget and Economic Outlook,
- OMB (Office of Management and Budget) from its Budget of the US Government, and
- APB (Analysis of the President's Budget).

The Congressional Budget Office and the Office of Management and Budget are different agencies within the US federal government. The Analysis of the President's Budget is published by the Congressional Budget Office, but the policy assumptions embedded in the Analysis of the President's Budget differ from those in the forecasts from the Budget and Economic Outlook. They are referred to as the "APB forecast" and the "CBO forecast" respectively, despite both coming from the Congressional Budget Office. For expositional convenience, these three forecasts are referred to as "agency forecasts", even though only two agencies are involved.

The agency debt forecasts and their relative merits depend on which measure is used. For example, the difference between the CBO and OMB gross federal debt forecasts in 2009 was largely due to differences in forecasts of the change in financial assets and liabilities in response to the financial crisis; see Martinez (2015, Table 2). The gross federal debt is also a closer measure of the debt subject to the debt ceiling, which featured prominently in policy debates about the debt following the financial crisis. Therefore, it is important to examine the forecasts of both the federal debt and deficit.

Comparison of government agency forecasts is complicated because they condition on different economic and policy assumptions. The CBO assumes that current law will be unchanged over the forecast period, whereas the OMB and APB assume that the policy changes proposed in the president's budget will be implemented. That being said, analyzing these different forecasts relative to actual outcomes can give a sense of the degree of usefulness of the agencies' forecasts. In this sense, the forecasts can be viewed as different policy scenarios rather than true unconditional forecasts. It

is of interest to determine how useful or accurate these different scenarios are, especially with the prominence they are given in policy debates and decisions.[1]

Forecast performance has both economic significance and statistical significance. That being said, the specific meaning of "poor forecast performance" depends in part on whether an agency's forecasts are interpreted as "forecasts" or as "projections", where "projections" are taken in the sense of being policy simulations conditional upon a certain set of assumptions. If the agency's forecasts are interpreted qua forecasts, then poor forecast performance implies potential room for improvement in terms of these performance measures. If the forecasts are interpreted qua projections, then poor forecast performance implies a limited usefulness of the forecasts as representing hypothetical paths for economic policy. In this context the overarching question that we seek to address is: which forecast represents the most useful path for economic policy at various forecast horizons?

3.2 Graphical Analysis

Graphs provide a useful preliminary to a more formal comparison and evaluation of the forecasts. This subsection presents the actual debt and deficit, the forecasts, and their forecast errors.

Figure 3.1 plots the debt and deficit and each of the agency forecasts to provide an initial view of how they have evolved over the past 30 years. The first and third columns present the "takeoff hedgehog plots" of each forecast path of the gross federal debt in logs and the federal deficit in billions of dollars across all horizons at the time of the initial forecast. Each "spine" on a takeoff graph plots a path of forecasts that were made on a given date, across multiple forecast horizons. The spines illustrate that longer-horizon forecasts perform particularly poorly around turning points. Forecasts made in 2001 and 2008 (both beginnings of recessions) tend to under-predict future outcomes. That said, debt forecasts in the late 1990s (an expansionary period) tend to over-predict somewhat: the economy grew faster than expected, with tax receipts bringing in more revenue than anticipated.

[1] As a linguistic matter, CBO refers to its forecasts of deficits and debt as "projections", while OMB refers to them as "estimates", and the APB publications typically call them "re-estimates". We use the term "forecasts" throughout, drawing on the more general usage given in Clements and Hendry (2002, p.2): "A forecast is any statement about the future".

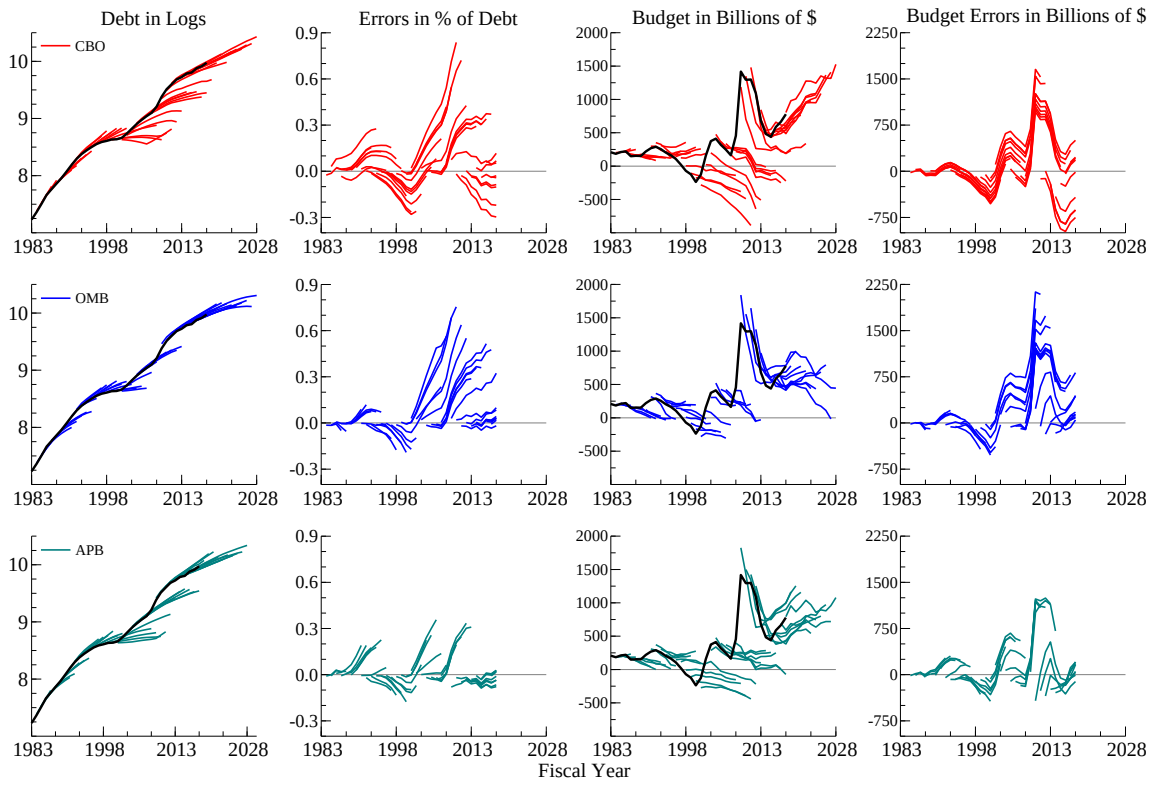


Figure 3.1: Forecasts and Forecast Errors of the U. S. Federal Debt and Deficit

More generally, takeoff graphs portray how "optimistic" or "pessimistic" the forecasts were, relative to outcomes, and how that optimism or pessimism evolved across horizons and over time.

The second and final columns of Figure 3.1 plot the path forecast errors, which were generated by the deviations of the forecasts from their actual values. There are prolonged periods in which the forecast errors, particularly those at longer horizons, are positive, indicating an optimistic forecast bias. The largest forecast errors were in 1990, 2001, 2002, 2008, 2009, 2011, and 2013. In each of these years, the United States was entering a recession or expansion (as dated by the National Bureau of Economic Research), or there were major policy changes.

4 Alternative Metrics for Comparing Forecasts

In many frameworks, good forecasts produce small expected losses, while bad forecasts produce large expected losses. One very common loss function is squared error loss, also known as quadratic mean squared error (MSE) loss. The MSE satisfies the requirements laid out by Granger (1999) that

a loss function (i) has a minimum of zero for a zero forecast error, (ii) is greater than or equal to zero for all forecast errors, and (iii) is non-decreasing in the magnitude of the error. Additionally, the MSE is symmetric, and its quadratic nature penalizes larger forecast errors more than proportionately.

In this vein, a forecast can be empirically evaluated by estimating its expected loss with the sample average of the loss. For the MSE, the sample average of the squared forecast errors is:

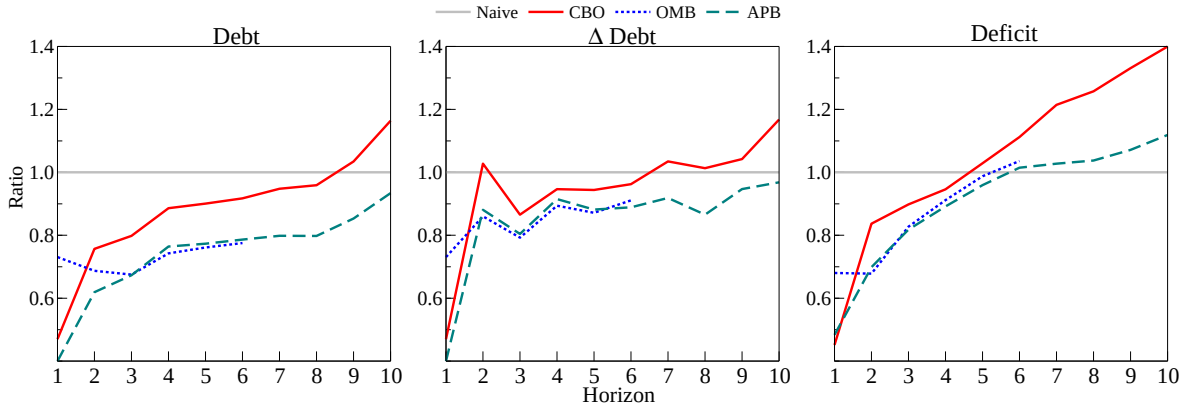
$$MSE_h = \hat{\mathbb{E}}\left[\tilde{e}^2_{t+h|t}\right] = \frac{1}{T}\sum_{t=1}^{T}\tilde{e}^2_{t+h|t} = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t+h} - \tilde{y}_{t+h|t}\right)^2, \tag{4.1}$$

where $\hat{\mathbb{E}}[\cdot]$ is the empirical expectations operator, y_{t+h} is the outcome at time $t+h$ (i.e., the variable being forecast), $\tilde{y}_{t+h|t}$ is its forecast produced using information available at time t, and the forecasts are made for T observations $\{t = 1,\ldots,T\}$.

In practice, the square root of the MSE is typically reported, rather than the MSE itself, noting that the root mean square error (RMSE) is the out-of-sample equivalent to the in-sample residual standard error. From the properties of the RMSE, a smaller value indicates a better forecast performance. Hence, it is common to compare forecast performance by comparing the RMSEs across forecasts, with smaller RMSEs indicating better performance.

Illustration. Figure 4.1 presents the RMSEs for each of the three agencies' debt, change in debt, and deficit forecasts. It reports the debt forecast performance relative to the RMSE for forecasts from a simple double-differenced device (DDD), which is a 'robust' naive forecast device that is calculated as the previous year's debt plus the change in the previous year's debt; see Hendry (2006). It reports the deficit and change in debt forecast performance relative to the RMSE for forecasts from a simple random walk (RW), which is calculated as the previous years deficit or change in debt.

Looking first at debt, all three agency forecasts perform better than the naive forecast through to at least eight-years-ahead. APB has the smallest RMSE followed by the CBO and then OMB. APB continues to do best as the forecast horizon increases. However, by three-years-ahead there is little difference in performance between the APB and OMB. By nine-years-ahead, the CBO no longer



Notes: Computed using different numbers of observations at each horizon and so are not entirely consistent across horizons. The maximum number of observations is 35 and the minimum is 12.

Figure 4.1: Relative RMSE Rankings by Horizon and Variable

outperforms a naive forecast and does not do so for any of the longer forecast horizons thereafter.

Now consider the rankings for the change in debt. At the first forecast horizon they are identical to forecasts of the level of debt. However, the rankings quickly change when considering multiple horizons. While APB continues to dominate across all ten horizons, CBO does not outperform the naive forecast at two-years-ahead or for horizons greater than seven-years-ahead. Overall the differences between the forecasts are much smaller at longer horizons.

Finally consider the rankings for the deficit. The rankings in the third panel of figure 4.1 illustrate that deficit forecast performance is different from debt and change in debt forecast performance. Contrary to the other forecasts, CBO starts out with the smallest RMSE at the one-year-ahead forecast horizon. However, it quickly becomes the worst performing forecast so that by five-years-ahead it no longer outperforms the naive forecast. Furthermore, unlike the debt and change in debt forecasts, APB and OMB also do not outperform the naive forecast beyond five-years-ahead.

A Lack of Invariance

Despite its popularity, the MSE does not satisfy a further requirement, originally proposed by Clements and Hendry (1993), that a loss function should allow for comparisons between different but isomorphic representations of the same forecasting system where the outcome does not depend on which representation is chosen; see Clements and Hendry (1998, p. 68). This requirement is motivated by

the fact that in practice there is often little consideration given as to whether the series of interest is in levels or growth rates of a particular variable. Therefore, the loss function should be invariant to this choice so as to indicate which forecast has the best performance regardless.

We can illustrate this in the context of fiscal forecasts with a relatively simple data generation process (DGP) for debt

$$y_t = y_{t-1} + \beta \Delta y_{t-1} + \epsilon_t \;\; \epsilon_t \sim \mathrm{IN}\left(0, \sigma^2\right), \tag{4.2}$$

where y_t denotes the debt at period t, y_{t-1} the debt in the previous period, Δy_{t-1} denotes the change in debt over the previous period, and ϵ_t denotes a random normal shock experienced at period t. The deficit is often thought to be equal to the change in debt so that by subtracting y_{t-1} from both sides the same DGP can be written in terms of the deficit or change in debt as

$$\Delta y_t = \beta \Delta y_{t-1} + \epsilon_t, \tag{4.3}$$

which provides two representations of the same DGP, about which policy makers may be ambivalent.

Consider forecasts of the debt. For simplicity, assume that the true model and parameters are known. Then the one-step-ahead forecast for the debt based on the conditional expectations using information through period $t-1$ is

$$\tilde{y}_{t|t-1} = y_{t-1} + \beta \Delta y_{t-1}. \tag{4.4}$$

The forecast error is then computed by subtracting (4.4) from (4.2) to get

$$\tilde{e}_{t|t-1} = y_t - \tilde{y}_{t|t-1} = \epsilon_t, \tag{4.5}$$

so that for a large sample of forecast errors the MSE will be

$$\mathbb{E}\left[\tilde{e}^2_{t|t-1}\right] = \sigma^2. \tag{4.6}$$

Now consider forecasts of the deficit where the one step-ahead-forecast is

$$\Delta\tilde{y}_{t|t-1} = \beta\Delta y_{t-1}, \tag{4.7}$$

and so the forecast error is

$$\Delta\tilde{e}_{t|t-1} = \Delta y_t - \Delta\tilde{y}_{t|t-1} = \epsilon_t, \tag{4.8}$$

which is identical to the forecast error in (4.5). As a result, the MSE will also be the same

$$\mathbb{E}\left[\Delta\tilde{e}^2_{t|t-1}\right] = \sigma^2. \tag{4.9}$$

Thus, when the deficit is a simple transformation of the debt and is forecast using the same model, then the 1-step-ahead MSE's will be identical. This is also seen in the panels for debt and the change in debt of figure 4.1 where the relative RMSE rankings are identical for the one-year-ahead forecast. Thus, the MSE loss function satisfies the 'invariance requirement' for one-step-ahead forecasts.

Now consider the two-step-ahead forecasts for both the debt and the deficit. Under the same assumptions as before and relying only on information available through period $t-1$, the two-step-ahead forecasts for the debt and change in debt are

$$\begin{aligned}\tilde{y}_{t+1|t-1} = \tilde{y}_{t|t-1} + \beta\Delta\tilde{y}_{t|t-1} = y_{t-1} + \left(\beta + \beta^2\right)\Delta y_{t-1} \\ \text{and } \Delta\tilde{y}_{t+1|t-1} = \beta\Delta\tilde{y}_{t|t-1} = \beta^2\Delta y_{t-1}.\end{aligned} \tag{4.10}$$

Given this, the forecast errors for these forecasts will be

$$\begin{aligned}\tilde{e}_{t+1|t-1} = y_{t+1} - \tilde{y}_{t+1|t-1} = (1+\beta)\epsilon_t + \epsilon_{t+1} \text{ and} \\ \Delta\tilde{e}_{t+1|t-1} = \Delta y_{t+1} - \Delta\tilde{y}_{t+1|t-1} = \beta\epsilon_t + \epsilon_{t+1}.\end{aligned} \tag{4.11}$$

Therefore, the MSE's are

$$\mathbb{E}\left[\tilde{e}^2_{t+1|t-1}\right] = (1+\beta)^2\sigma^2 + \sigma^2 \text{ and}$$
$$\mathbb{E}\left[\Delta\tilde{e}^2_{t+1|t-1}\right] = \beta^2\sigma^2 + \sigma^2, \tag{4.12}$$

which are not equal except for the special case where $\beta = -0.5$. Furthermore, this result generalizes for all horizons $h > 1$. See the panels for debt and the change in debt of figure 4.1 where the relative rankings are different for the multi-year-ahead forecasts. Thus, the MSE does not satisfy the 'invariance requirement' for multi-step-ahead forecasts and so these rankings are not invariant to isomorphic transformations of the data, the forecasts, or the forecast errors.

The General Forecast-Error Second-Moment

Clements and Hendry (1993) propose an alternative loss function which does satisfy this requirement: the general forecast-error second-moment (GFESM). Rather than evaluating forecasts based on their expected squared loss at a single horizon, the GFESM considers the joint expected squared loss across multiple horizons. For the GFESM, the sample average or $\hat{\mathbb{E}}(\cdot)$, of the joint squared loss is:

$$GFESM_h = \left|\hat{\mathbb{E}}\left[\tilde{E}_{h|t-1}\tilde{E}'_{h|t-1}\right]\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\tilde{E}_{h|t-1}\tilde{E}'_{h|t-1}\right|, \tag{4.13}$$

where $\tilde{E}_{h|t-1} = \left\{\tilde{e}_{t|t-1},\ldots,\tilde{e}_{t+h-1|t-1}\right\}'$ is a stacked vector of the forecast errors for all horizons through h and $|\cdot|$ is the determinant. In practice, it is common to standardize the GFESM by raising it to the $1/2h$ power, noting that this produces a geometric average of the RMSE across horizons. It is also common to take the natural log which relates the GFESM to other multivariate measures of forecast accuracy; see Doan et al. (1984). In either case, a smaller value indicates a better forecast performance.

We can illustrate that the GFESM satisfies the invariance requirement by considering the same example as above. Starting with the debt forecasts and considering horizons through the two-step-

ahead forecasts, then for a large sample of forecast errors the GFESM will be:

$$\left|\mathbb{E}\left[\tilde{E}_{2|t-1}\tilde{E}'_{2|t-1}\right]\right| = \begin{vmatrix} \sigma^2 & (1+\beta)\sigma^2 \\ (1+\beta)\sigma^2 & (1+\beta)^2\sigma^2+\sigma^2 \end{vmatrix} = \sigma^4(1+\beta)^2+\sigma^4-\sigma^4(1+\beta)^2 = \sigma^4, \quad (4.14)$$

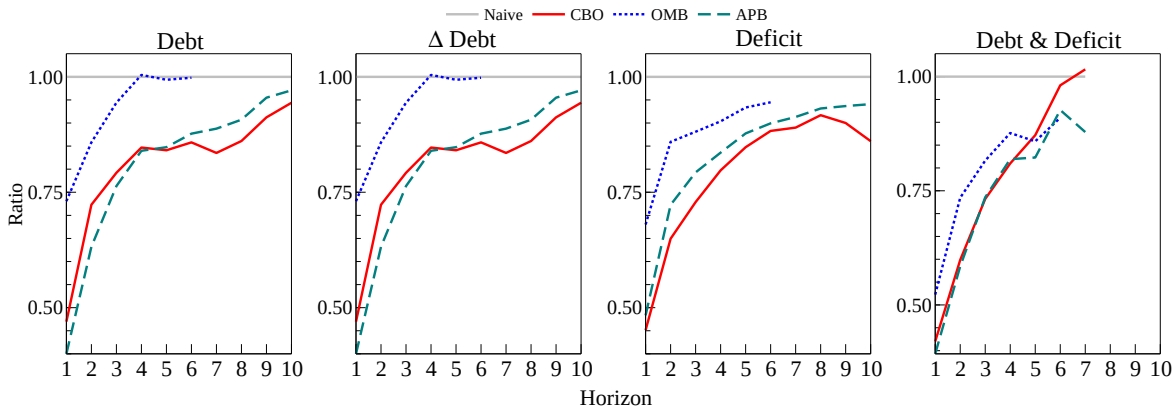
where if we standardize this measure, we get the exact same result as the one-step-ahead RMSE in the case where there is no parameter estimation uncertainty and the true model is known. Now consider the deficit forecasts where for a large sample of forecast errors the GFESM will be:

$$\left|\mathbb{E}\left[\Delta\tilde{E}_{2|t-1}\Delta\tilde{E}'_{2|t-1}\right]\right| = \begin{vmatrix} \sigma^2 & \beta\sigma^2 \\ \beta\sigma^2 & \beta^2\sigma^2+\sigma^2 \end{vmatrix} = \sigma^4\beta^2+\sigma^4-\sigma^4\beta^2 = \sigma^4, \quad (4.15)$$

where $\Delta\tilde{E}_{2|t-1}$ is the stacked vector of deficit forecast errors. This illustrates that the GFESM is identical regardless of whether the forecasts for the debt or deficit are considered. It is possible to represent the transformation from the debt to the deficit as $\Delta\tilde{E}_{2|t-1} = M_2\tilde{E}_{2|t-1}$ where $M_2 = \left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$. This implies that the invariance requirement will be satisfied for all forecast horizons and for any transformation such that $|M_h| = 1$.

Illustration. Figure 4.2 presents the relative GFESM rankings of the three agency forecasts for debt, change in debt, deficit as well as the joint debt and deficit forecasts. The rankings are again presented relative to naive forecasts. The GFESM is standardized so that the one-year-ahead rankings are identical to the RMSE rankings. The relative rankings in the panels for debt and change in debt are identical across all horizons which illustrates the invariance properties of the GFESM. In both cases APB starts out with the smallest GFESM up through four-years-ahead. CBO has the lowest GFESM for horizons longer than four-years-ahead while OMB does not outperform the naive forecast for horizons beyond four-years-ahead.

The rankings of the deficit forecasts indicate that CBO performs best across all horizons. The third panel in figure 4.2 shows that CBO has the lowest GFESM followed by APB and then OMB across all horizons. Furthermore, the agency deficit forecasts all outperform the naive forecast which



Notes: Computed using different numbers of observations at each horizon and so are not entirely consistent across horizons. The maximum number of observations is 35 and the minimum is 12.

Figure 4.2: Relative GFESM Rankings by Horizon and Variable

differs considerably from the RMSE rankings.

We also use the GFESM to examine the joint forecast accuracy of the debt and deficit across horizons. This is presented in the fourth panel of figure 4.2. The measure only extends out through seven horizons due to the relatively small number of observations. However, it provides a joint perspective on the relative accuracy of the forecasts across variables and horizons. Note that if the deficit were purely a transformation of the debt then this analysis would not be possible since the forecast errors would be perfectly correlated. CBO and APB's joint forecast performance is almost identical up through five-years-ahead, after which, CBO experiences a relative deterioration. By seven-years-ahead CBO no longer performs differently from the naive forecasts while APB and OMB continue to do slightly better. This illustrates that the rankings for individual variables can differ from the rankings for the entire system which accounts for dependencies across variables and horizons.

The Conditional MSE

While the GFESM measures the joint loss across multiple horizons and variables, there is also a single horizon loss function that satisfies the same requirements as the GFESM. We refer to this loss function as the Conditional MSE (CMSE) since it effectively measures the expected squared loss at a particular horizon, conditional on the errors at all previous horizons. The CMSE, the sample average

of the squared conditional loss is:

$$CMSE_h = \hat{\mathbb{E}}\left[\tilde{e}^2_{t+h-1|t-1}|\tilde{E}_{h-1|t-1}\right] = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{e}_{t+h-1|t-1} - \hat{\theta}_h\tilde{E}_{h-1|t-1}\right)^2, \tag{4.16}$$

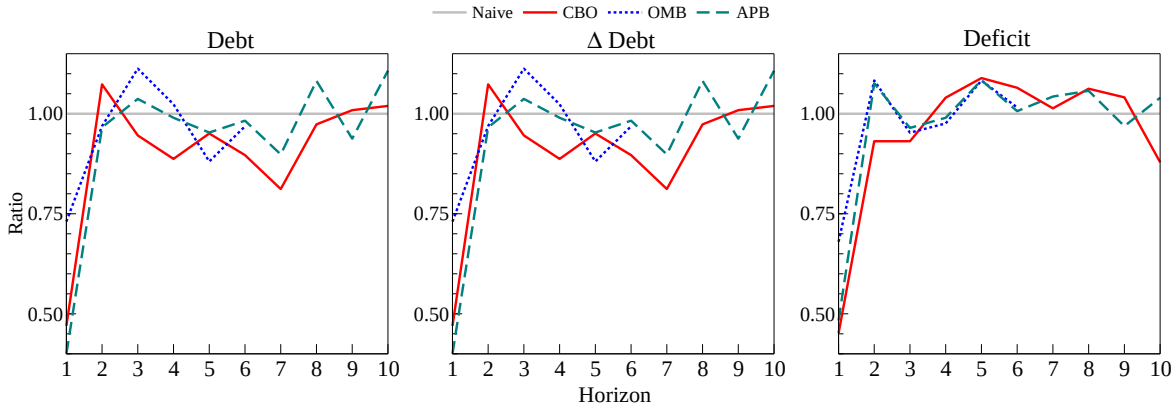
where $\hat{\theta}_h$ is a $(1 \times h-1)$ parameter vector of the estimated dependence between the forecast error at horizon h and the forecast errors at earlier horizons. Therefore, this measure effectively captures the unique contribution of horizon h to the squared loss by conditioning on the forecast errors at earlier horizons.[2] We can illustrate the invariance properties of the CMSE by evaluating the debt and deficit forecasts from our example. For a large sample of forecast errors the CMSE for the two-step-ahead forecasts of the debt and deficit will be:

$$\begin{aligned}\mathbb{E}\left[\tilde{e}^2_{t+1|t-1}|\tilde{e}_{t|t-1}\right] &= \sigma^2 \text{ and} \\ \mathbb{E}\left[\Delta\tilde{e}^2_{t+1|t-1}|\Delta\tilde{e}_{t|t-1}\right] &= \sigma^2,\end{aligned} \tag{4.17}$$

which illustrates that the CMSE has the same invariance properties as the GFESM despite effectively only measuring the expected loss of one forecast horizon at a time.

Illustration. Figure 4.3 presents the relative CMSE rankings of the three agency forecasts for debt, change in debt, and the deficit. The rankings are again presented relative to naive forecasts and are standardized by taking the square root so that the one-year-ahead rankings are identical to the RMSE rankings. As with the GFESM rankings, the rankings for the debt and the change in debt are identical. Unlike the RMSE and GFESM rankings, the CMSE rankings are very unstable with no clear dominant forecast across horizons. Although APB starts out withd the smallest CMSE, by three-years-ahead it performs worse than the naive forecast with only slight fluctuations around it. CBO does somewhat better after two-years-ahead but it does not outperform the naive forecast beyond eight-years-ahead. Similar dynamics are observed for the deficit forecasts with the exception

[2]This is also essentially what the determinant measures for multi-horizon forecasts. In fact it is possible to derive the CMSE from the GFESM as $CMSE_h = GFESM_h/GFESM_{h-1}$.



Notes: Computed using different numbers of observations at each horizon and so are not entirely consistent across horizons. The maximum number of observations is 35 and the minimum is 12.

Figure 4.3: Relative Conditional RMSE Forecast Rankings by Horizon and Variable

that each of the agency forecasts consistently perform slightly worse than the naive forecast after three-years-ahead with CBO only doing well in the ten-years-ahead forecasts.

The Relative MSE

In some instances it may also be useful to compare between different loss functions. For example, a useful comparison is what we refer to as the relative MSE which can be computed by dividing the CMSE by the MSE. This measure always has a value of less than or equal to one and presents the relative share of the unique contribution of the forecast errors at a horizon relative to the overall forecast performance at that horizon. In the context of our running example, the relative MSE's for a large sample of debt and deficit forecast errors are:

$$\begin{aligned} \mathbb{E}\left[\tilde{e}^2_{t+1|t-1}|\tilde{e}_{t|t-1}\right]/\mathbb{E}\left[\tilde{e}^2_{t+1|t-1}\right] &= \frac{1}{(1+\beta)^2+1} \text{ and} \\ \mathbb{E}\left[\Delta\tilde{e}^2_{t+1|t-1}|\Delta\tilde{e}_{t|t-1}\right]/\mathbb{E}\left[\Delta\tilde{e}^2_{t+1|t-1}\right] &= \frac{1}{\beta^2+1}. \end{aligned} \tag{4.18}$$

These measures have a direct link to the commonly used forecast error variance decomposition (FEVD) measure discussed in Hamilton (1994, p. 323) and Lütkepohl (2005, p. 63-66). Reformulating the FEVD in a multi-horizon context implies that it effectively measures the amount of the overall forecast error variance at horizon h that is attributable to earlier horizons. Thus, the relative

MSE is equal to $1 - FEVD_h$. The main difference between the multi-horizon formulation of $FEVD_h$ as opposed to the standard multivariate interpretation is that there is a unique temporal ordering across horizons as opposed to variables in a system.

General Remarks

Overall the different forecast accuracy metrics illustrate that if the debt and deficit forecasts are simple isomorphic transformations of one-another, then while MSE can give different rankings for multi-horizon forecasts, the GFESM and CMSE rankings will be identical regardless of which transformation is considered. Furthermore, if the debt and deficit forecasts were perfectly collinear but evaluated jointly across horizons, the GFESM would be singular. Since this is not the case, then the debt and deficit are not simple transformations of one-another and there is value in examining them jointly. Thus, looking at multiple loss functions can provide a more complete perspective and additional insight into the accuracy of and relationship between forecasts across horizons and variables.

Across each of the metrics, we find that the agency forecasts perform better than naive forecasts at short and medium-term horizons. At the one-year-ahead forecast horizon, the APB forecasts outperform the CBO and OMB forecasts in terms of the RMSFEs (for debt but not the deficit). This is not surprising given APB's somewhat later release date and therefore potential information advantage which has a larger impact at shorter horizons. The evidence is more mixed for medium and longer-term horizons. At the five-year-ahead forecast horizon where information advantages should matter less, the naive forecasts achieve the lowest RMSE for the deficit while APB has the lowest RMSE for the debt. Both APB and CBO perform well at longer horizons for the debt and deficit in terms of the GFESM. Overall the results suggest that, while sometimes the different rankings are aligned, there can be substantial differences in the rankings across metrics. While the APB forecasts dominate across most horizons in terms of the RMSE measure for both the debt and deficit, CBO's forecasts generally perform best as measured by the GFESM and particularly at longer horizons. This illustrates the importance of using different metrics. The RMSE captures total forecast performance at

an individual horizon, the CMSE captures the unique contribution to forecast performance at a given horizon and the the GFESM provides a joint measure of forecast performance across all horizons.

5 Structure of the Thesis

Chapter 2. Evaluating Multi-Step System Forecasts with Relatively Few Observations

Despite the importance of analyzing forecasts jointly across horizons and variables, analyses of large systems and/or long horizons are often limited by a shortage of forecast-error observations. Estimates of second-moment matrices allow for the joint evaluation of forecast errors across variables and horizons, but deteriorate rapidly when there are more variables times horizons than observations. This chapter solves this problem by producing design-free and model-free estimates of forecast-error second-moment matrices when there are more variables times forecast horizons than observations.

The chapter seeks to answer the following questions: How well does the standard procedure for calculating the GFESM perform as the number of variables (K) times horizons (H) gets closer to the number of observations (N)? Is it possible to improve on this approach in settings where KH is larger than N? This chapter compares several different approaches for extending estimates of second-moment matrices to high dimensional cases where $KH > N$. It yields several important findings. First, the standard procedure is increasingly biased and imprecise when there are relatively few forecast-error observations, which can distort forecast rankings. Second, the proposed design-free and model-free method outperforms the standard procedure and other more restricted methods across a variety of forecast models and data generation processes. The proposed method typically produces the correct forecast ranking even when there are relatively few observations.

Chapter 3. Testing for Differences in Path Forecast Accuracy

Multi-horizon forecast paths contain information about both the individual point forecasts at each horizon as well as their relationship with and dependence on the point forecasts at all other horizons.

This chapter constructs a test for differences in path forecast accuracy. Testing for differences in the path allows us to capture the expected dynamics of individual variables across horizons as well as the dynamic relationships between variables in the system. Despite their prominence, tests of point forecasts at individual horizons do not capture these important aspects of the future trajectory.

We demonstrate the link between general forecast error metrics and joint densities. Building on this insight, we draw on methods from both the point and density forecast accuracy literature to construct a general test of differences in path forecast accuracy. We show that this general path forecast accuracy test nests existing multi-horizon point forecast tests while also being able to capture differences in the forecast dynamics across models. We also derive the asymptotic distribution of this test for a special case where both forecasts are well specified.

Monte Carlo simulations illustrate the trade-offs associated with this test. Path forecast accuracy tests are less able to detect differences in biases. This is due to the fact that bias tends to be highly correlated across horizons and so is offset by higher forecast-error covariances. However, path forecast accuracy tests have a much greater ability to capture differences in variances, covariances, and dynamics across forecast models. This is particularly true for differences in forecast dynamics across horizons, which other tests are unable to capture. Thus, our path forecast accuracy test captures a new aspect of differences in forecast accuracy. We also demonstrate how path tests could be extended for use in higher dimensional settings.

As an application of the methods, we compare the Federal Reserve Board's Greenbook path forecasts with forecasts from four DSGE models. We decompose the differences in the path forecasts to show that they are primarily driven by the differences in forecast dynamics across horizons. This illustrates the importance of accounting for differences across the entire path when assessing forecast accuracy. Next, we evaluate the accuracy of path forecasts of Hurricane Irma in 2017 across many models and using different path weighting schemes to illustrate the flexibility of these methods.

Chapter 4. The Impact of Forecast Accuracy on Hurricane Damages

This chapter quantifies how the accuracy of forecasts leading up to a natural disaster affect the resulting destruction. While a considerable amount of research examines changes in the natural hazards and vulnerabilities associated with natural disasters, little attention has been paid to the effectiveness of short-term damage mitigation decisions just prior to a disaster's occurrence. Even less research focuses on how early warning systems based on forecasts of the event affects these decisions. Quantifying the extent to which short-term forecasts of natural disasters help individuals make better damage mitigation decisions is important for understanding the effectiveness of short-term adaptation efforts, for illuminating the rising costs of natural disasters, and for understanding the speed of the recovery.

We start by showing that, in an expected utility framework, errors in the forecasts of a natural disaster can affect damages if individuals place too much faith in the accuracy of these forecasts when formulating their beliefs about the benefits of damage mitigation. Next, using our theoretical framework as a guide we test for evidence of these effects using a newly constructed database of all hurricanes to strike the continental United States from 1955-2014. We formulate an empirical model of damages with many determinants including the 12-hour-ahead landfall-forecast errors and estimate it using ordinary least squares. We use model selection methods to determine the best model specification and also whether forecast errors are among the most important determinants of damages. The robustness of our results is confirmed using a variety of selection procedures and model specifications. Finally, we conduct a counterfactual exercise in order to quantify how much improvements in short-term forecast accuracy since 1970 have altered hurricane damages.

Chapter 5. Conclusions

This final chapter summarizes the findings from each of the chapters and provides a potential path forward for future research avenues that extend the multi-step system forecast evaluation framework set forth in this thesis. There are many ways to evaluate imperfect forecasts. We show the importance of considering the joint distributions over variables and horizons. Once a multi-step system context is

envisaged, potential singularities need to be addressed, as do tests of significant differences between forecasts and assessments of the value of improved forecasts. These important issues are the focus of the thesis and are illustrated using a variety of applications in macroeconomics and climate.

Chapter 2

Evaluating Multi-Step System Forecasts with Relatively Few Observations

This chapter develops a new approach for evaluating multi-step system forecasts with relatively few forecast-error observations. It extends the work of Clements and Hendry (1993) by using that of Abadir et al. (2014) to generate "design-free" estimates of the general matrix of the forecast-error second-moment when there are relatively few forecast-error observations. Simulations show that the usefulness of alternative methods deteriorates when their assumptions are violated. The new approach compares well against these methods and provides correct forecast rankings.

This chapter has been published as:

Hendry, D.F. and Martinez, A.B. "Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations", *International Journal of Forecasting* Vol. 33, 2 (2017), pp. 359-372.

1 Introduction

Both short- and long-term forecasts are important for policy debates. For example, short-term debt forecasts are linked to concerns about growth, while long-term debt forecasts are tied to debt sustainability (e.g. Martinez, 2015).[1] The joint importance of short- and long-term forecasts is not limited to macroeconomics, but is found in many areas, including climate change (e.g. Pretis and Roser, 2017). While forecasts are usually evaluated separately across horizons, Clements and Hendry (1993) illustrate that it is important to jointly evaluate forecasts across horizons to capture any dependence. This allows for the assessment of forecast performance in both the short- and the long-term.

While jointly evaluating multiple forecast horizons is useful, so too is the evaluation of large forecasting systems. Large macroeconomic forecasting models which try to capture the dependency across variables are widely used by central banks.[2] Although individual variables from large forecasting systems are evaluated regularly, relatively little work has been done to evaluate the accuracy of the whole system. An exception is the work of Sinclair et al. (2012, 2015), who evaluate a vector of forecasts simultaneously at a single horizon. However, there are limitations in terms of how easily these methods can be extended to evaluate multiple horizons jointly.

Despite the importance of analyzing forecasts jointly across horizons and variables, the available methods are limited, due largely to a shortage of forecast-error observations. For example, Clements and Hendry (1993) propose the use of the general matrix of the forecast-error second-moment and its determinant (GFESM) as an invariant measure of forecast accuracy. While the GFESM allows for the joint evaluation of forecast errors across variables and horizons, it deteriorates rapidly in relatively small samples. Thus, an improved approach for evaluating multi-step system forecasts with relatively few forecast-error observations is required.

This chapter presents a solution to this problem by extending estimates of the GFESM to relatively small samples. It combines the work of Abadir et al. (2014) and Clements and Hendry (1993)

[1] See Standard & Poor's, "United States of America Long-Term Rating Lowered to 'AA+' on Political Risks and Rising Debt Burden: Outlook Negative," August 2, 2011.

[2] For example, see Bårdsen et al. (2012) and Burgess et al. (2013).

to allow for estimation of the GFESM when there are more variables (K) times forecast horizons (H) than forecast-error observations (N). This extends the GFESM to large forecasting systems with long horizons even when there are relatively few forecast-error observations.

The chapter seeks to answer the following questions: How well does the standard approach for calculating the GFESM perform as KH gets closer to N? Is it possible to improve estimates of the forecast-error second-moment matrix in these cases? Comparing several approaches, this chapter extends estimates of the GFESM beyond the non-singular case to the singular case where $KH > N$. It yields several important findings. First, the standard approach is increasingly biased and imprecise when there are relatively few forecast-error observations, which can distort forecast rankings. Second, the proposed method outperforms the standard approach across a variety of forecast models and data generation processes (DGPs). Third, the proposed method typically produces the correct forecast ranking even when there are relatively few observations.

The rest of the chapter is structured as follows. The next section reviews different forecast accuracy evaluation methods, with a focus on the GFESM. Section 3 lays out a new analytical approach for estimating the GFESM when $KH > N$. Section 4 conducts Monte Carlo experiments for a known-parameters model to determine how well the proposed method performs relative to the standard approach in small-sample settings and across various DGPs. Section 5 extends the Monte Carlo experiments to alternative forecast models. Section 6 allows for structural breaks and section 7 applies the methods to a vector of forecasts of the US economy. Section 8 concludes.

2 Existing Methods to Evaluate Forecast Accuracy

This section provides a brief introduction to the theory of forecasting as well as a common approach to evaluating forecasts, i.e. the mean square forecast error (MSFE). It then goes on to introduce an alternative approach, the general matrix of the forecast-error second-moment and its determinant (GFESM). The advantages of the GFESM over MSFE measures are illustrated, together with its limitations and the ways in which it has been applied in the literature.

Consider a DGP defined by a stationary, pth-order vector autoregressive process (e.g. VAR(p)) for a vector of K variables $\mathbf{Y_t}$:

$$\mathbf{Y}_t = \boldsymbol{\theta} + \sum_{j=0}^{p-1} \boldsymbol{\Pi}_j \mathbf{Y}_{t-1-j} + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim IN_K\left[\mathbf{0}, \boldsymbol{\Omega}\right]. \tag{2.1}$$

Bold terms are vectors, and $\mathbf{v}_t$ is a $(K \times 1)$ vector of independent normal residuals with mean $\mathbf{0}$, positive definite variance $\boldsymbol{\Omega}$ and all the eigenvalues of the polynomial matrix $\left(\mathbf{I}_K - \sum_{j=0}^{p-1} \boldsymbol{\Pi}_j L^{j+1}\right)$ where L denotes the lag operator, are inside the unit circle. Assuming the initial value is equal to its long-run mean, $\boldsymbol{\Theta} = \left(\mathbf{I}_K - \sum_{j=0}^{p-1} \boldsymbol{\Pi}_j\right)^{-1} \boldsymbol{\theta}$, then (2.1) can be rewritten as

$$\mathbf{X}_t = \sum_{j=0}^{p-1} \boldsymbol{\Pi}_j \mathbf{X}_{t-1-j} + \mathbf{v}_t, \tag{2.2}$$

where $\mathbf{X}_t = (\mathbf{Y}_t - \boldsymbol{\Theta})$ is demeaned by its long-run mean. Let $\boldsymbol{\Gamma}_{p,i} = \sum_{j=0}^{p-1} \boldsymbol{\Gamma}_{p,i-1-j} \boldsymbol{\Pi}_j$, $\boldsymbol{\Gamma}_{p,0} = \mathbf{I}_K$, $\boldsymbol{\Gamma}_{p,i} = \mathbf{0}$ when $i < 0$, and $\boldsymbol{\Pi}_j = \mathbf{0}$ when $j \geq p$. Then the best possible h-step-ahead forecast at time T is the conditional expectation: $\mathbf{X}_{T+h|T} = \mathbb{E}_T\left[\mathbf{X}_{T+h} | \mathbf{X}_T\right] = \sum_{j=0}^{p-1} \boldsymbol{\Gamma}_{p,max(0,h-j)} \boldsymbol{\Pi}_j^{min(1,j)} \mathbf{X}_{T-j}$, where $h \in [1,...,H]$. Therefore given estimates of $\left\{\widehat{\boldsymbol{\Pi}}_j\right\}$, the forecast error is

$$\widetilde{\mathbf{u}}_{T+h|T} = \left(\mathbf{X}_{T+h} - \widetilde{\mathbf{X}}_{T+h|T}\right) = \sum_{j=0}^{p-1} \left(\boldsymbol{\Gamma}_{p,max(0,h-j)} \boldsymbol{\Pi}_j^{min(1,j)} - \widehat{\boldsymbol{\Gamma}}_{p,max(0,h-j)} \widehat{\boldsymbol{\Pi}}_j^{min(1,j)}\right) \mathbf{X}_{T-j} + \sum_{i=0}^{h-1} \boldsymbol{\Gamma}_{p,i} \mathbf{v}_{T+h-i}. \tag{2.3}$$

When the true parameters are known, (2.3) delivers unbiased forecast errors that have a variance of $\sum_{i=0}^{h-1} \boldsymbol{\Gamma}_{p,i} \boldsymbol{\Omega} \boldsymbol{\Gamma}'_{p,i}$. Thus, even the smallest possible multi-step forecast errors from a dynamic model are a moving average process.

The MSFE, which assumes a quadratic loss function, is commonly used to evaluate forecasts. In multivariate systems, the MSFE becomes the mean square forecast error matrix (or the matrix of the forecast-error second-moment, MFESM):

$$\mathbb{E}_T\left[\widetilde{\mathbf{u}}_{T+h|T} \widetilde{\mathbf{u}}'_{T+h|T}\right] = \mathbf{V}_h = \sum_{i=0}^{h-1} \boldsymbol{\Gamma}_{p,i} \boldsymbol{\Omega} \boldsymbol{\Gamma}'_{p,i}, \tag{2.4}$$

where the last equality holds when the true parameters are known and $\mathbf{v}_t$ is IID. Multivariate forecasts are often evaluated using the trace of the MFESM: $tr(\mathbf{V}_h)$.

2.1 An Invariant Measure of Forecast Accuracy

Clements and Hendry (1993, 1998) propose a more general and invariant measure of forecast accuracy termed the general matrix of the forecast-error second-moment (GMFESM) $\mathbf{\Phi}_H$ and its determinant $|\mathbf{\Phi}_H|$, the GFESM. The GMFESM is calculated by multiplying the stacked forecast errors across all horizons and variables. Following from (2.3) the forecast error, $\widetilde{\mathbf{u}}_{T+h+n|T+n}$ is a $(K \times 1)$ vector at horizons $h = 1,...,H$ from origin $T+n$, where $n = 0,...,N-1$ represents the number of forecast-error observations.[3] Stack the forecast errors as

$$\widetilde{\mathbf{W}}_{H,N|T} = \begin{pmatrix} \widetilde{\mathbf{u}}_{T+1|T} & \cdots & \widetilde{\mathbf{u}}_{T+N|T+N-1} \\ \vdots & \ddots & \vdots \\ \widetilde{\mathbf{u}}_{T+H|T} & \cdots & \widetilde{\mathbf{u}}_{T+H+N-1|T+N-1} \end{pmatrix} \tag{2.5}$$

which is a $KH \times N$ matrix. Then the GMFESM is the $KH \times KH$ matrix

$$\widehat{\mathbf{\Phi}}_H = \tfrac{1}{N}\widetilde{\mathbf{W}}_{H,N|T}\widetilde{\mathbf{W}}'_{H,N|T} = \begin{pmatrix} \frac{1}{N}\sum_{n=0}^{N-1}\widetilde{\mathbf{u}}_{T+1+n|T+n}\widetilde{\mathbf{u}}'_{T+1+n|T+n} & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\widetilde{\mathbf{u}}_{T+1+n|T+n}\widetilde{\mathbf{u}}'_{T+H+n|T+n} \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1}\widetilde{\mathbf{u}}_{T+H+n|T+n}\widetilde{\mathbf{u}}'_{T+1+n|T+n} & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\widetilde{\mathbf{u}}_{T+H+n|T+n}\widetilde{\mathbf{u}}'_{T+H+n|T+n} \end{pmatrix}. \tag{2.6}$$

The GFESM is the determinant of this matrix. (2.6) illustrates that each element along the main diagonal is the MSFE across variables and horizons. Thus, the trace of this matrix is the trace MSFE over all K and H (Trace). This illustrates the connection between the GFESM and other forecast evaluation methods. It also shows that the GFESM includes all of the information available in the Trace while also incorporating information from the covariances across variables and horizons.

The known-parameters case provides additional insight. When the true parameters are known,

[3]Note that in the literature forecast origins and forecast-error observations are used interchangeably. Here the latter term is used.

but the error variance is estimated as $\widehat{\mathbf{\Omega}}_{N+h} = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{v}_{T+n+h}\mathbf{v}'_{T+n+h}$, then plugging (2.3) into (2.6), simplifying, and then taking the determinant gives

$$\begin{aligned}
\left|\widehat{\mathbf{\Phi}}_H\right| &= \begin{vmatrix} \mathbf{\Gamma}_{p,0}\widehat{\mathbf{\Omega}}_{N+1}\mathbf{\Gamma}'_{p,0} & \cdots & \mathbf{\Gamma}_{p,0}\widehat{\mathbf{\Omega}}_{N+1}\mathbf{\Gamma}'_{p,H-1} \\ \vdots & \ddots & \vdots \\ \mathbf{\Gamma}_{p,H-1}\widehat{\mathbf{\Omega}}_{N+1}\mathbf{\Gamma}'_{p,0} & \cdots & \sum_{i=0}^{H-1}\mathbf{\Gamma}_{p,i}\widehat{\mathbf{\Omega}}_{N+H-i}\mathbf{\Gamma}'_{p,i} \end{vmatrix} \\
&= \begin{vmatrix} \mathbf{\Gamma}_{p,0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Gamma}_{p,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{\Gamma}_{p,H-1} & \cdots & \mathbf{\Gamma}_{p,1} & \mathbf{\Gamma}_{p,0} \end{vmatrix} \begin{vmatrix} \widehat{\mathbf{\Omega}}_{N+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\mathbf{\Omega}}_{N+H} \end{vmatrix} \begin{vmatrix} \mathbf{\Gamma}_{p,0} & \mathbf{\Gamma}_{p,1} & \cdots & \mathbf{\Gamma}_{p,H-1} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{\Gamma}_1 \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Gamma}_{p,0} \end{vmatrix} = \prod_{h=1}^{H}\left|\widehat{\mathbf{\Omega}}_{N+h}\right|
\end{aligned} \tag{2.7}$$

The second equality follows from the fact that the determinant of a product is the product of the determinants. The third equality follows from the fact that $\mathbf{\Gamma}_{p,0} = \mathbf{I}_K$. Thus, in this special case the GFESM is the product of the determinant of the error variances across horizons.

In finite samples when the true error variance is known and unchanging, or in large samples when the error variance converges to the same value such that $\mathbf{\Omega} = \widehat{\mathbf{\Omega}}_{N+1} = \ldots = \widehat{\mathbf{\Omega}}_{N+H}$, then the GFESM simplifies further. Standardizing by the forecast horizon H gives

$$|\mathbf{\Phi}_H|^{1/H} = \left(|\mathbf{\Omega}|^H\right)^{1/H} = |\mathbf{\Omega}|. \tag{2.8}$$

Clements and Hendry (1995, p. 135) note that this transformation "stabilizes estimates of the GFESM for large H when $|\mathbf{\Phi}_1|$ differs from unity" without any impact on forecast comparisons. However, it may also "downplay differences when parameters are estimated" (p. 139).

This illustrates that the one-step-ahead forecast errors specify the complete ranking across all horizons when *the true parameters are known*. Several studies, notably De Gooijer and Vidiella-i-Anguera (2004) and Engle (1993), are critical of this result. They argue that the GFESM attributes more weight to the one-step forecast errors than other methods. However, as Schmidt (1993) and Castle and Hendry (2008) note, although forecast uncertainty can increase monotonically, this does

not hold if the error variance changes, if parameters are unknown in finite samples, or if there are other types of mis-specification. Further, this result is not unique to the GFESM. The Trace also relies heavily on the one-step forecast errors when the parameters are constant and known.

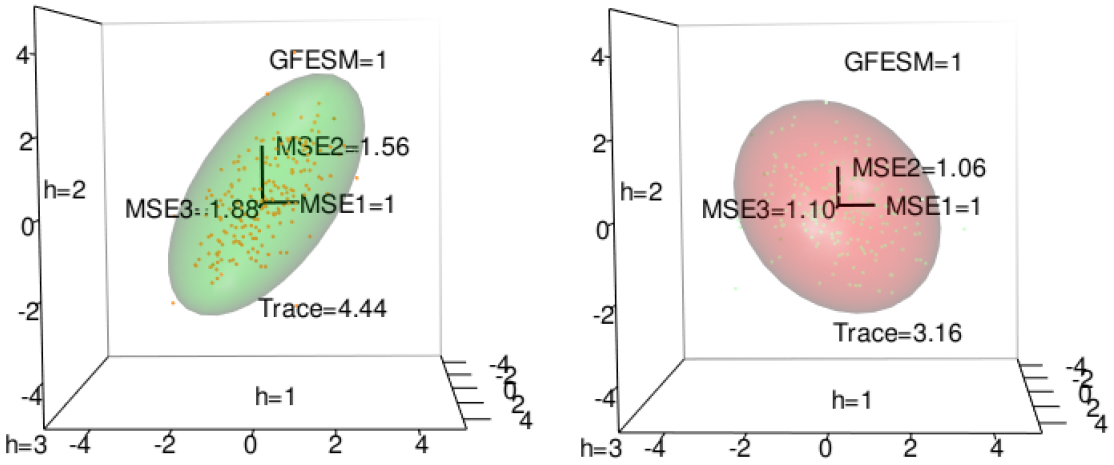
While the GFESM is not unique in terms of the relative importance of the one-step forecast errors, it is unique in terms of its invariance properties under certain widely used transformations. Consider transforming the forecast errors by the matrix $\boldsymbol{M}_{K,h}$ such that the vector of forecast errors becomes $\boldsymbol{M}_{K,h}\widetilde{\mathbf{u}}_{T+h+n|T+n}$. Generalizing Clements and Hendry (1998, p. 73) to allow for transformations of variables and forecast horizons, the transformed matrix of stacked forecast errors in (2.5) becomes $\widetilde{\mathbf{W}}^{*}_{H,N|T} = (\boldsymbol{M}_H \otimes \boldsymbol{M}_K)\widetilde{\mathbf{W}}_{H,N|T}$ where $\boldsymbol{M}_H$ is an alternative weighting matrix across horizons while $\boldsymbol{M}_K$ is a scaling matrix across variables and $\otimes$ is the Kronecker product. Estimating the GFESM with the transformed stacked forecast errors gives

$$\left|\widehat{\boldsymbol{\Phi}}^{*}_{H}\right| = \left|\frac{1}{N}\widetilde{\mathbf{W}}^{*}_{H,N|T}\widetilde{\mathbf{W}}^{*\prime}_{H,N|T}\right| = \left|\widehat{\boldsymbol{\Phi}}_{H}\right| \times |\boldsymbol{M}_H|^{2H} \times |\boldsymbol{M}_K|^{2H}. \tag{2.9}$$

First, consider the case where $abs(|\boldsymbol{M}_H|) = 1$ and $abs(|\boldsymbol{M}_K|) = 1$ such that transformations are scale preserving and non-singular. Included in this class of transformations are cointegrating combinations, differences, differentials and substitution of identities among others. Equation (2.9) shows that the GFESM is invariant to this class of transformation, $\left|\widehat{\boldsymbol{\Phi}}^{*}_{H}\right| = \left|\widehat{\boldsymbol{\Phi}}_{H}\right|$, and so the ranking of forecasts is unchanged. This result is not generally true for the Trace, where $tr\left(\widehat{\boldsymbol{\Phi}}^{*}_{H}\right) \neq tr\left(\widehat{\boldsymbol{\Phi}}_{H}\right)$. For example, the choice between evaluating forecasts in levels or in differences is irrelevant for the GFESM but not for the Trace.[4]

Now consider the case where $|\boldsymbol{M}_K|$ (and/or $|\boldsymbol{M}_H|$) $\neq 1$; for example, when variables are scaled differently. The GFESM is not invariant to this transformation. However, as (2.9) illustrates, this will rescale the original GFESM by the determinant of the respective weight matrix. Thus, transformations that are applied uniformly across all forecast models have no impact on the relative rankings.

[4]See Clements and Hendry (1995, 1997). Note that estimation methods that rely on non-linear transformations, such as those discussed below, generate finite sample estimates of the GFESM that are not invariant.



Panel A: Original Forecast Errors

Panel B: Transformed Forecast Errors

Notes: The ellipsoids are generated based on 300 forecast errors from a multivariate normal distribution when H=3 and K=1. The forecast errors are generated by a MA(h-1) process where the persistence parameter $\pi = 0.75$ and the variance of the underlying shocks is $\sigma^2 = 1$.

Figure 2.1: An illustration of Invariance

The interpretation as to whether one forecast dominates the other in terms of the GFESM is similar to the MFESM and MSFE criteria. For example, consider the forecasts from two different models (A and B). Model B's forecasts dominate model A's forecasts in terms of the mean square forecast error if the MSFE for model B is smaller than that of model A. This holds for the GMFESM, where differences between the GMFESM for A and B gives a positive definite matrix. It follows that GMFESM dominance is a sufficient but not a necessary condition for GFESM dominance where

$$\widehat{\mathbf{\Phi}}_H^A - \widehat{\mathbf{\Phi}}_H^B > \mathbf{0} \text{ implies } \left|\widehat{\mathbf{\Phi}}_H^A\right| > \left|\widehat{\mathbf{\Phi}}_H^B\right|. \tag{2.10}$$

Thus, GFESM dominance is weaker than GMFESM dominance. Taking logs, GFESM dominance can be rewritten as $ln\left(\left|\widehat{\mathbf{\Phi}}_H^A\right| / \left|\widehat{\mathbf{\Phi}}_H^B\right|\right) > 0$. The log GFESM is related to the profile predictive likelihood when the residuals are normally distributed as illustrated by Clements and Hendry (1993, 1998).

The link to the predictive likelihood illustrates that the GFESM can be interpreted as a measure of the volume of space needed by a KH-dimensional hyper-ellipsoid around the origin to enclose the forecast errors, where a larger value denotes a poorer set. Further, as the hyper-ellipsoid may be 'elongated' more in some directions than others, the lack of invariance to MSFE-type measures is easily seen to be due to the direction measured. We can illustrate this with a simple example

where $H = 3$, $K = 1$, $\pi = 0.75$ and $\sigma_v^2 = 1$. The 3-dimensional ellipsoid surrounding the forecast errors at all three horizons is shown in Panel A of Figure 2.1 where $|\mathbf{\Phi}_3| = 1$ and $tr(\mathbf{\Phi}_3) = 4.44$. We can transform the forecast errors by differencing them at longer horizons so that $M_3 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$. The 3-dimensional ellipsoid around the transformed forecast errors is shown in Panel B. While the mean square errors shrink at longer horizons so that $tr\left(\mathbf{\Phi}_3^*\right) = 3.16$, the total volume around the errors across all three horizons remains the same: $\left|\mathbf{\Phi}_3^*\right| = 1$. In fact, for any value of $\pi \neq 0.5$, then $tr(\mathbf{\Phi}_3) \neq tr\left(\mathbf{\Phi}_3^*\right)$ and $|\mathbf{\Phi}_3| = \left|\mathbf{\Phi}_3^*\right| = 1$ for this transformation. This illustrates the invariance of the GFESM to transformations of the forecasts and the lack thereof in other MSFE measures.

2.2 Limitations of the GFESM

Despite its advantages relative to MSFE measures, the GFESM has one important limitation which hampers its applicability in practice. While the GMFESM is a positive definite moment matrix of the stacked forecast errors, when variables are collinear or when there are more variables times horizons (KH) than forecast-error observations (N) then the GMFESM is singular and so the GFESM goes to zero.[5] The former problem can be solved by excluding collinear variables or through other adjustments but the latter cannot.

Since the GFESM tends to zero when $KH > N$, the literature generally steers clear of this case. Although Clements and Hendry (1995) study the "small sample properties" of the GFESM, their analysis remains well away from the singular GMFESM. Other studies limit their use of the GFESM by focusing on either variables or horizons. For example, Clements and Hendry (1997) focus on cross-variable forecast errors from five different transformations of the same variable. Paap et al. (1997) use the cross-horizon forecast errors of consumption data to compute the GFESM. Other studies that take this approach include Kunst and Franses (1998), De Gooijer and Ray (2003), De Gooijer and Vidiella-i-Anguera (2004), Veloce (2004), Castle and Hendry (2008), Athanasopoulos et al. (2011) and Simionescu (2013). While the GMFESM is non-singular in most of these studies,

[5] Granger (1993) first pointed this out for the GFESM. Strictly speaking the former is considered singular while the latter is more precisely defined as ill-conditioned. Here they are referred to interchangeably.

Table 2.1: Overview of Previous Studies that use the GFESM

Study	K (max)	H (max)	KH	N (min)
Wallis (1993)	2	1	2	8
Clements and Hendry (1995)	2	20	40	50
Hoffman and Rasche (1996)	5	16	80	28
Clements and Hendry (1997)	5	8	40	22
Paap et al. (1997)	1	8	8	16
Kunst and Franses (1998)	3	8	24	28
Zeng and Swanson (1998)	5	5	25	1,400
Anderson et al. (2002)	6	16	96	36
Vahid and Issler (2002)	3	16	48	100
De Gooijer and Ray (2003)	2	5	10	10
Clements and Galvão (2004)	2	24	48	100
De Gooijer and Vidiella-i-Anguera (2004)	2	5	10	40
Veloce (2004)	1	8	8	13
Castle and Hendry (2008)	1	8	8	20
Bowsher and Meeks (2008)	6	1	6	72
Baillie and Morana (2009)	1	12	12	940
Athanasopoulos et al. (2011)	3	16	48	90
Simionescu (2013)	4	1	4	40

in some cases there are only just enough forecast-error observations. See Table 2.1 for an indication of how close previous studies have gotten to $KH > N$.

The GFESM is not limited by the number of observations in all applications, as some (particularly in finance) have relatively large samples of forecast-error observations. In these applications, collinear variables are more likely to be a concern. For example, see Zeng and Swanson (1998), Clements and Galvão (2004), Bowsher and Meeks (2008) and Baillie and Morana (2009).

Some studies attempt to use the GFESM even when there are relatively few observations. For example, Hoffman and Rasche (1996) apply the GFESM when $KH > N$. They impose restrictions on the GMFESM based on assumptions about its structure. The crucial assumption is that the forecast errors are uncorrelated across variables. While the "cross-variable implications of the system forecast ranking" are lost by imposing these restrictions, they justify this by claiming that the rankings do not change substantially at shorter horizons when comparisons between the restricted and unrestricted GMFESM are feasible. Anderson et al. (2002) follow a similar approach.

Several studies, including Kunst and Franses (1998) and De Gooijer and Ray (2003) come close to the boundary of the singular GMFESM, but do not analyze the performance of the estimates. While the GFESM is estimable in such cases, it is unclear how estimates and rankings are affected by assumptions about the error structure or as the GMFESM gets close to the singular case. The rest of this chapter addresses these questions.

3 Methodological Approach

While few approaches exist for estimating moment matrices with relatively few observations, there is an extended literature for covariance matrices. Several methods in the literature ensure estimates of covariance matrices remain well-conditioned (i.e. positive definite). However, they typically require imposing relatively stringent assumptions about the structure. Abadir et al. (2014) note that there are three main avenues in the literature to ensure positive definiteness, namely the imposition of restrictions, the use of factor models, and shrinkage methods.

Few of these approaches have been used with the GFESM. An exception is the method proposed by Hoffman and Rasche (1996). Their "constrained" method is calculated using the "standard" approach except that cross-variable covariances of the GMFESM are set to zero, which ensures the calculated GMFESM is positive definite as long as $H \leq N$. Thus, there are potentially large improvements in estimates of the GFESM by using more flexible methods when $KH > N$.

One method that was developed explicitly for time series applications and which could potentially be applicable to the GFESM was proposed by McMurry and Politis (2010) and extended by Jentsch and Politis (2015). This "tapered" approach can be considered a shrinkage method in that it down-weights elements of the covariance matrix as a function of distance from the main diagonal.[6] While the benefit of this method is clear in terms of standard autocovariance matrices, it is less clear for the GMFESM where covariances are less likely to taper off.

This chapter focuses instead on a more flexible approach to ensure that the GMFESM remains positive definite. A "design-free" method was proposed by Abadir et al. (2014) which does not impose any assumptions about the distribution of the sample or the parametric structure of the matrix. This gives the method considerable flexibility and Abadir et al. (2014) demonstrate that it performs well relative to alternative approaches in a variety of settings.

The crucial assumption imposed by this method is that the data are IID. However, this assump-

[6] Even so, this approach only generates estimates that are asymptotically positive definite. Thus, in finite samples the maximum between the estimated eigenvalues and some convergence factor is used to ensure positive definiteness.

tion is not satisfied, since multi-step forecast errors from any dynamic model are serially correlated. Abadir et al. (2014) propose filtering the data using an appropriate model. This chapter proposes a model-free transformation of the stacked forecast errors to eliminate serial correlation.

In the known-parameters case, if (2.3) is substituted into (2.5), Clements and Hendry (1998, p. 315) show that the stacked forecast errors can be decomposed as

$$\widetilde{\mathbf{W}}_{H,N|T} = \begin{pmatrix} \mathbf{\Gamma}_{p,0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Gamma}_{p,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{\Gamma}_{p,H-1} & \cdots & \mathbf{\Gamma}_{p,1} & \mathbf{\Gamma}_{p,0} \end{pmatrix} \begin{pmatrix} \mathbf{v}_{T+1} & \cdots & \cdots & \mathbf{v}_{T+N} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{v}_{T+H} & \cdots & \cdots & \mathbf{v}_{T+H+N-1} \end{pmatrix} = \mathbf{\Psi}_{p,H}\mathbf{Z}_{H,N}. \quad (3.1)$$

This result also follows from (2.7). $\mathbf{Z}_{H,N}$ no longer contains terms that are serially-correlated. When this decomposition is applied to the GMFESM, it becomes

$$\widehat{\mathbf{\Phi}}_{H,N} = \mathbf{\Psi}_{p,H}\left(\mathbf{\Sigma}_{H,N} + \mathbf{\Xi}\right)\mathbf{\Psi}'_{p,H}, \quad (3.2)$$

where $\mathbf{\Xi}$ is the squared bias, $\mathbf{\Sigma}_{H,N} = \frac{1}{N}\mathbf{Z}_{H,N}\mathbf{Q_N}\mathbf{Z'}_{H,N}$ and $\mathbf{Q}_N$ is the demeaning matrix. In the known-parameters case, $\mathbf{\Xi} = \mathbf{0}$ such that the covariance and second-moment matrices are identical. In general, though, the squared bias needs to be removed to calculate the covariance matrix and then added back to get the GMFESM.

When the forecast model is known, $\mathbf{\Psi}_{p,H}$ can be derived from the model parameters under the assumption that the model is well-specified. When the underlying forecast model is not known, then $\mathbf{\Psi}_{p,H}$ can be estimated from the forecast errors. This is done by taking the h-step-ahead forecast error at $T+n$ as in (2.3) and subtracting the $(h-1)$-step-ahead forecast error at $T+n+1$. Assuming that neither the model nor the DGP changes substantially across observations, this transformation gives

$$\Delta\widetilde{\widetilde{\mathbf{u}}}_{T+n+h|T+n} = \widetilde{\mathbf{u}}_{T+n+h|T+n} - \widetilde{\mathbf{u}}_{T+n+h|T+n+1} = \mathbf{\Gamma}_{p,h-1}\mathbf{v}_{T+n+1}. \quad (3.3)$$

Equation (3.3) illustrates that the quasi-differenced h-step-ahead forecast errors are a function of the parameters $\mathbf{\Gamma}_{p,h-1}$ and the one-step-ahead forecast errors. Thus, $\mathbf{\Gamma}_{p,h-1}$ can be estimated by regressing the 1-step-ahead forecast errors, $\left\{\mathbf{u}_{T+n+1|T+n}\right\}_{n=0}^{N-2}$, on the quasi-differenced h-step-ahead forecast errors, $\left\{\Delta\widetilde{\widetilde{\mathbf{u}}}_{T+n+h|T+n}\right\}_{n=0}^{N-2}$: $\Delta\widetilde{\widetilde{\mathbf{u}}}_{T+n+h|T+n} = \widehat{\mathbf{\Gamma}}_{p,h-1}\mathbf{u}_{T+n+1|T+n} + \boldsymbol{\epsilon}_{T+n}$. Note that the estimation sample is $N-1$, meaning that estimates of $\mathbf{\Gamma}_{p,h-1}$ become less reliable in small samples. Repeating this analysis for $h = \{2,...,H\}$ traces out $\mathbf{\Psi}_{p,H}$. When $\mathbf{\Omega}$ is non-diagonal, then estimates of $\mathbf{\Gamma}_{p,h-1}$, and so estimates of $\mathbf{\Psi}_{p,H}$, are augmented by the additional dependence structure in $\mathbf{\Omega}$.

The forecast errors can be transformed to eliminate serial correlation using estimates of $\mathbf{\Psi}_{p,H}$: $\widetilde{\mathbf{Z}}_{H,N} = \widehat{\mathbf{\Psi}}_{p,H}^{-1}\widetilde{\mathbf{W}}_{H,N|T}$. If estimates are consistent so that $\widehat{\mathbf{\Psi}}_{p,H}^{-1} \approx \mathbf{\Psi}_{p,H}^{-1}$ and the forecast model is reasonably well-specified, then this transformation will eliminate most serial correlation. The transformed covariance matrix is $\widehat{\mathbf{\Sigma}}_{H,N} = \frac{1}{N}\widetilde{\mathbf{Z}}_{H,N}\widehat{\mathbf{Q}}_N\widetilde{\mathbf{Z}}'_{H,N}$, which remains singular when $KH > N$.

Even with this transformation, $\widetilde{\mathbf{Z}}_{H,N}$ is only quasi-IID. While the transformation ensures independence, in practice there is no guarantee that the residuals are identically distributed over time. Hendry and Mizon (2014) argue that shifts in the distribution are a major source of forecast failure. One way to deal with shifts and fat-tails is to use indicator saturation methods when estimating $\mathbf{\Gamma}_{p,h-1}$; see Hendry et al. (2008), Johansen and Nielsen (2009) and Castle et al. (2012). Using these methods, it is possible to extract shifts or outliers and include them within the estimate of $\mathbf{\Xi}$.

Let $\widehat{\mathbf{\Sigma}}_{H,N}$ be the $KH \times KH$ positive semi-definite covariance matrix of the transformed stacked forecast errors. This generalizes Abadir et al. (2014), which is a special case where $H = 1$, $\mathbf{\Xi} = \mathbf{0}$, and $\mathbf{\Pi} = \mathbf{0}$ so that $\widetilde{\mathbf{W}}_{N|T} = \widetilde{\mathbf{Z}}_N$. The transformed covariance matrix can be decomposed as

$$\widehat{\mathbf{\Sigma}}_{H,N} = \frac{1}{N}\widetilde{\mathbf{Z}}_{H,N}\mathbf{Q}_N\widetilde{\mathbf{Z}}'_{H,N} = \widehat{\mathbf{P}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{P}}', \tag{3.4}$$

where $\widehat{\mathbf{P}}$ is an orthogonal matrix of eigenvectors with a typical column $\widehat{\mathbf{p}}_i$ ($i = 1,...,KH$) and $\widehat{\mathbf{\Lambda}}$ is the diagonal matrix of eigenvalues. Since $\widehat{\mathbf{P}}$ is orthogonal, it is well-conditioned for any value of N and KH. Therefore, the only source of ill-conditioning in $\widehat{\mathbf{\Sigma}}_H$ when $KH > N$ stems from the estimation

of the eigenvalues, $\widehat{\mathbf{\Lambda}}$, since the eigenvalues beyond N tend to zero. This illustrates why the GFESM is limited by the number of observations but the Trace is not: The determinant is the product of the eigenvalues, $\prod_{i=0}^{KH} \lambda_i$, while the trace is the sum of the eigenvalues, $\sum_{i=0}^{KH} \lambda_i$, see Abadir and Magnus (2005, Chapter 7). However, both measures suffer from information loss when $KH > N$.

The "design-free" method relies on estimates of the eigenvectors to generate alternative estimates of the eigenvalues. Using the fact that $\widehat{\mathbf{P}}'\widehat{\mathbf{P}} = \mathbf{I}_{KH}$, then (3.4) can be rearranged so that

$$\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{P}}'\widehat{\mathbf{\Sigma}}_{H,N}\widehat{\mathbf{P}} = diag(\widehat{var}(\widehat{\mathbf{p}}_1'\widetilde{Z}_{H,N}), \ldots, \widehat{var}(\widehat{\mathbf{p}}_{K\times H}'\widetilde{Z}_{H,N})), \tag{3.5}$$

where $\widehat{var}(\widehat{\mathbf{p}}_i'\widetilde{Z}_{H,N}) = \frac{1}{N}\widehat{\mathbf{p}}_i'\widetilde{\mathbf{Z}}_{H,N}\widehat{\mathbf{Q}}_N\widetilde{\mathbf{Z}}_{H,N}'\widehat{\mathbf{p}}_i$. The second equality follows since $\widehat{\mathbf{\Lambda}}$ is diagonal. The innovation of this method is to base estimates of the eigenvectors, $\widehat{\mathbf{P}}$, on a subsample of observations. The IID assumption for $\widetilde{\mathbf{Z}}_{H,N}$ implies that it is efficient to use $s < N$ observations. Splitting the transformed stacked forecast errors into two samples gives

$$\widetilde{\mathbf{Z}}_{H,N} = (\widetilde{\mathbf{Z}}_{H,s} : \widetilde{\mathbf{Z}}_{H,N-s}), \tag{3.6}$$

where $\widetilde{\mathbf{Z}}_{H,s}$ and $\widetilde{\mathbf{Z}}_{H,N-s}$ are $KH \times s$ and $KH \times (N-s)$, respectively. The first-step estimator of the eigenvectors, $\widehat{\mathbf{P}}_s$, is obtained by estimating $\widehat{\mathbf{\Sigma}}_{H,s}$ from s observations

$$\widehat{\mathbf{\Sigma}}_{H,s} = \widehat{\mathbf{P}}_s\widehat{\mathbf{\Lambda}}_s\widehat{\mathbf{P}}'_s. \tag{3.7}$$

The first-step estimator can then be used to orthogonalize the remaining $N-s$ observations in order to derive an estimate of the eigenvalues

$$\widetilde{\mathbf{\Lambda}}_s = diag(\widehat{\mathbf{P}}_s\widehat{\mathbf{\Sigma}}_{H,N-s}\widehat{\mathbf{P}}'_s). \tag{3.8}$$

The IID assumption for $\widetilde{\mathbf{Z}}_{H,N}$ implies that it is efficient to use $N-s$ observations for this second step.

Furthermore, Abadir et al. (2014) show that reusing any of the s observations worsens estimates of $\widetilde{\mathbf{\Lambda}}_s$, which replaces $\widehat{\mathbf{\Lambda}}$ in (3.4). Adding the squared bias, $\mathbf{\Xi}$, and then multiplying by the parameter matrix $\mathbf{\Psi}_H$ unravels the error transformation and gives a "design-free" estimate of the GMFESM:

$$\widetilde{\mathbf{\Phi}}_{H,s} = \widehat{\mathbf{\Psi}}'_{p,H} \left[\widehat{\mathbf{P}} \widetilde{\mathbf{\Lambda}}_s \widehat{\mathbf{P}}' + \widehat{\mathbf{\Xi}} \right] \widehat{\mathbf{\Psi}}_{p,H}, \tag{3.9}$$

which is positive definite and non-singular as long as $N > s$ even when $KH > N$.[7]

There are many possible ways to choose subsamples of size s from N. Given the IID assumption, the "design-free" method can be generalized by averaging over R subsample re-samples of s to "reduce the variability that comes with the choice of any one specific combination of s observations" (p. 167). As a result, the generalized "design-free" method is

$$\widetilde{\mathbf{\Phi}}_{H,s,R} = \widehat{\mathbf{\Psi}}'_{p,H} \left[\widehat{\mathbf{P}} \left(\frac{1}{R} \sum_{r=1}^{R} \widetilde{\mathbf{\Lambda}}_{s,r} \right) \widehat{\mathbf{P}}' + \widehat{\mathbf{\Xi}} \right] \widehat{\mathbf{\Psi}}_{p,H}. \tag{3.10}$$

The generalized "design-free" approach adds two additional parameters to the estimation of the GMFESM: the size of the first-step estimator, s, and the number of subsample re-samples, R. First, consider how to choose the optimal number of subsample re-samples. In theory, R can be as large as $\binom{N}{s}$. However, given the IID assumption, each of these $\binom{N}{s}$ possibilities are equivalent so that the choice of R is of minor importance.

Choosing the optimal size of the first-step estimator is more complicated. As Abadir et al. (2014, p. 169) note, there is a trade off between the estimation of $\widehat{\mathbf{P}}_s$, which improves with large s, and the estimation of $\widehat{\mathbf{\Sigma}}_{H,N-s}$, which improves with a small s. For example, assume that $\{\mathbf{v}_T\}$ has the same known variance for all K variables, and observations, $H+N$, but is independent across variables so that $\widehat{\mathbf{\Omega}}_{1+N} = ... = \widehat{\mathbf{\Omega}}_{H+N} = \mathbf{\Omega} = \sigma^2 \boldsymbol{I}_K$. Then, the covariance matrix of the transformed stacked forecast errors is a scalar matrix

$$\mathbf{\Sigma}_H = \sigma^2 \boldsymbol{I}_{KH}, \tag{3.11}$$

[7] Since $\widetilde{\mathbf{Z}}_{H,s}$ is de-meaned, then the GMFESM is non-singular as long as $N - 1 > s$.

where σ^2 is the same across variables and horizons. Abadir et al. (2014, p. 167) argue that in this case, the precision of the estimation of $\widetilde{\mathbf{\Phi}}_{H,s,R}$ is "invariant to [the eigenvectors], and the optimal choice of s is as small as possible to increase the precision of the eigenvalues estimated in the second-step". Relaxing the assumption on variance invalidates this result. However, the special case provides insight into the choice of the first-step estimator and also illustrates why alternative approaches such as zero restrictions or shrinkage are effective.

In practice, there is considerable variation in the optimal choice of s. Abadir et al. (2014) argue that "choosing small s delivers the largest improvements in the conditioning of the estimated variance matrix" (p. 174) but also suggest that choosing s to be around $0.5N$ provides the best estimate of the covariance matrix (p. 177). This supports the choice of a range of values of s for evaluating the stability of the performance of the "design-free" method.

By choosing the optimal first-step estimator to orthogonalize the unused observations, the approach proposed in Abadir et al. (2014, p. 166) reformulates a multivariate problem from one of "ill-conditioning and imprecision to a univariate problem" in each diagonal element in (3.5). While this approach generates well-defined eigenvalues, and therefore positive definite matrices, it is important to note that there is no additional information beyond N.

Once $KH > N$, then the eigenvalues beyond N tend to zero. The additional eigenvectors beyond N are then arbitrary and can be normalized to any value. Thus, while positive definite estimates of the GMFESM are generated, these estimates, and therefore the GFESM, may deviate substantially from the true underlying values. However, in principle as long as the normalization of the eigenvectors is applied uniformly across forecast models then it is possible that this approach would still generate invariant rankings of the forecasts. Furthermore, averaging across the subsample re-samples may help reduce substantial deviations.

4 Simulations

In this section, Monte Carlo experiments are run to study the small-sample properties of various estimators of the GFESM across different horizons, variables and DGPs. The simulation design is as follows: The DGP is a first-order vector-autoregressive process (VAR(1))

$$\mathbf{X}_t = \mathbf{\Pi}\mathbf{X}_{t-1} + \mathbf{v}_t, \text{ for } \mathbf{v}_t \sim IN_K[\mathbf{0}, \mathbf{\Omega}], \tag{4.1}$$

where $\mathbf{X}_t = (\mathbf{Y}_t - \mathbf{\Theta})$ is a vector of K variables that are demeaned by a vector of their long-run means. While the Gaussian case is the focus here, Abadir et al. (2014) illustrate that the "design-free" method also performs well with alternative fat-tailed distributions which are relevant for finance and turbulent macroeconomic environments. In the baseline case, the system is composed of two variables ($K = 2$) forecast across four horizons ($H = 4$) where $\mathbf{\Pi} = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}$ and $\mathbf{\Omega} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$. The known-parameters model is considered first so that $\mathbf{\Pi}$ is known, there is no bias and no parameter estimation uncertainty. This implies that all serial correlation is eliminated using $\mathbf{\Psi}_{1,H}^{-1}$; see (3.1).

The simulations explore the performance of estimates of the GMFESM across a range of forecast-error observations, including when the standard approach is singular. This analysis assesses the performance of a range of methods as they relate to the GFESM. Four approaches are considered: the "standard", "constrained", "tapered", and "design-free" methods. See Abadir et al. (2014) for comparisons of the "design-free" method against alternative approaches in another context.

The "standard" approach was originally proposed by Clements and Hendry (1993). Adapting (2.6) to illustrate the transformed stacked forecast errors from (3.1), then it is computed as

$$\left|\widehat{\mathbf{\Phi}}_{H,N}\right|^{1/H} = \left|\widehat{\mathbf{\Psi}}_{1,H}\left(\widehat{\mathbf{\Sigma}}_{H,N} + \widehat{\mathbf{\Xi}}\right)\widehat{\mathbf{\Psi}}'_{1,H}\right|^{1/H} = \left|\widehat{\mathbf{\Sigma}}_{H,N} + \widehat{\mathbf{\Xi}}\right|^{1/H}, \tag{4.2}$$

where $\widetilde{\mathbf{W}}_{H,N|T} = \widehat{\mathbf{\Psi}}_{1,H}\widetilde{\mathbf{Z}}_{H,N}$ is calculated from (2.5) and the forecast errors, $\widetilde{\mathbf{u}}_{T+h+n|T+n}$ are generated from the known-parameters model, see (2.3, when $p = 1$ and $\widehat{\mathbf{\Pi}} = \mathbf{\Pi}$). The second equality

follows from the fact that the determinant of the coefficient matrix is unity, so that the "standard" approach is invariant to the use of the transformed or non-transformed stacked forecast errors. Since the system consists of two variables and four forecast horizons ($KH = 8$), the minimum number of forecast-error observations, N, for the "standard" approach to be well-conditioned is 8.

Next, the "design-free" method is considered. This approach computes the GFESM as

$$\left|\widetilde{\mathbf{\Phi}}_{H,s,R}\right|^{1/H} = \left|\widehat{\mathbf{P}}\left(\frac{1}{R}\sum_{r=1}^{R}\widetilde{\mathbf{\Lambda}}_{s,r}\right)\widehat{\mathbf{P}}' + \widehat{\mathbf{\Xi}}\right|^{1/H}, \tag{4.3}$$

where the "standard" approach, $\widehat{\mathbf{\Phi}}_H$, is replaced by the general "design-free" method from (3.10). Unlike (4.2), the "design-free" method is not invariant to transformations of the forecast errors. The "design-free" method requires additional choices in terms of R and s. Setting $R = 20$ is sufficient to reap most benefits.[8] Abadir et al. (2014, p. 173) find "robust performance as s varies around its optimal value, more specifically around approximately $s \in [0.2N, 0.8N]$". Thus, a range is used $s = \{s_1, s_2, s_3\} = \{0.2N, 0.5N, 0.8N\}$ where the resulting number is then rounded to the nearest integer less than $N - 1$. For example, when $N = 20$ then $s_1 = 4$, $s_2 = 10$, $s_3 = 16$. For simplicity, s_1 is referred to as "min", s_2 is referred to as the "mid", and s_3 is referred to as "max".[9] These three values give the range for how (4.3) varies with s.

The "constrained" and "tapered" methods are considered for comparison. They are calculated using the transformed forecast errors following Hoffman and Rasche (1996) and Jentsch and Politis (2015, pp. 7-9) respectively.[10] The "constrained" method is invariant to transformations of the stacked forecast errors when its restrictions are valid. However, the "tapered" method is not invariant to transformations of the forecast errors.

To ascertain how close the methods get to the actual GFESM, it is necessary to have a baseline

[8] See Abadir et al. (2014, p. 169). Simulations show that there is a trade-off in that there are marginal benefits to increasing R when the estimator is singular, but that with higher R performance declines as N increases.

[9] Abadir et al. (2014) also propose a 'grand average' method which averages over the different subsamples. However, it is excluded here since its performance generally lies between the subsamples considered.

[10] For the "tapered" approach, elements more than two positions away from the main diagonal are set equal to zero and the convergence factor is specified as $\frac{1}{N}$. This implies that the user defined constants are all set equal to 1.

for comparison. In a Monte Carlo setting, the DGP is known, so it is possible to estimate the "true" GFESM. From (2.7) when $p = 1$, the "true" GFESM for the known-parameters forecast model is

$$|\mathbf{\Phi}_H|^{1/H} = \left| \begin{array}{ccc} \widehat{\mathbf{\Omega}}_{1+N} & \cdots & \widehat{\mathbf{\Omega}}_{1+N}\mathbf{\Pi}'^{H-1} \\ \vdots & \ddots & \vdots \\ \mathbf{\Pi}^{H-1}\widehat{\mathbf{\Omega}}_{1+N} & \cdots & \sum_{i=0}^{H-1}\mathbf{\Pi}^{i}\widehat{\mathbf{\Omega}}_{H+N-1-i}\mathbf{\Pi}'^{i} \end{array} \right|^{1/H}, \tag{4.4}$$

where $\widehat{\mathbf{\Omega}}$ is estimated from the correctly specified model such that

$$\widehat{\mathbf{\Omega}}_{N+h} = \frac{1}{T+N}\sum_{i=h}^{T+N+h-1}\mathbf{v}_i\mathbf{v}_i'. \tag{4.5}$$

The "true" GFESM is used as the baseline against which the various methods are compared. The analysis first examines how different methods estimate eigenvalues of the GMFESM and then the GFESM itself. Simulations are run with 1,000 replications.[11]

4.1 Eigenvalues

The relative performance of the methods is given by examining how well the eigenvalues of the GMFESM are estimated. These estimates are important because the GFESM is the product of the eigenvalues. Table 4.1 illustrates the "standard" approach overstates the largest eigenvalues and substantially under-reports the smallest eigenvalues when $KH > N$. The "standard" approach goes to zero because it severely underestimates any eigenvalues greater than the number of observations.

Each of the methods attempts to smooth this dynamic by reducing the over- (under-)reporting of the largest (smallest) eigenvalues. The smoothing done by the "constrained" method is efficient only when its restrictions are valid. The "tapered" and "design-free" methods do not impose constraints, meaning that the resulting changes are less efficient but are valid when there is additional correlation across variables. These methods require larger reductions in the largest eigenvalues in order

[11] All numerical results were obtained using OxMetrics Version 7.10; see Doornik (2013); Doornik and Hendry (2013a).

Table 4.1: Largest to Smallest Eigenvalues by Method, when $KH = 8$ and $N = 6$

Eigenvalue	"true"	"standard"	"const."	"tapered"	"design-free"		
					"min"	"mid"	"max"
1	3.99	5.49	5.11	6.38	5.94	5.38	4.84
2	1.67	2.23	1.88	2.01	1.90	1.60	1.30
3	1.29	1.29	1.21	1.24	1.14	0.93	0.74
4	1.02	0.75	0.83	0.88	0.84	0.70	0.52
5	0.82	0.37	0.57	0.64	0.60	0.53	0.38
6	0.66	0.13	0.37	0.38	0.45	0.40	0.27
7	0.53	0.00	0.22	0.21	0.33	0.30	0.19
8	0.38	0.00	0.10	0.11	0.22	0.20	0.11
GFESM	1.20	0.00	0.66	1.21	3.50	1.11	0.09
Trace	6.38	5.13	5.81	7.13	9.21	6.13	4.26

Notes: Estimates are averages from 1,000 replications.

to obtain similar increases in the smallest eigenvalues. This is especially clear for the "design-free" method where all but the largest eigenvalues are under-reported with "mid" under-reporting the smallest eigenvalues the least. This has implications for the GFESM. While the GFESM is greater than zero for each of the methods, "tapered" and "mid" are closest to the "true" GFESM. The "standard" Trace suffers from information loss such that it too benefits from the "design-free" methods.

4.2 The GFESM

Now consider the impacts of the different methods on estimates of the GFESM, where bias and the relative mean square error (MSE) are used to judge the performances. Bias is calculated as the percentage difference between the estimated and the "true" GFESM. The relative MSE is calculated as the ratio of the mean square error of the estimated GFESM (relative to "true") to that of the singular GFESM (relative to "true"). This standardizes the MSE across methods. The results are presented in percentages where a value of 100 for bias (−) or MSE (+) indicates that the estimate is no better than the singular case. Values less than $abs(\pm 100)$ represent an improvement relative to the singular case.

For the baseline case in Figure 4.1 where $K = 2$ and $H = 4$, the "standard" GMFESM is initially non-singular until $N = 8$ when the MSE and bias start to decrease. Put slightly differently, the "standard" approach deteriorates as N decreases relative to KH. Furthermore, while the "standard" approach is consistent for large values of N and fixed KH, it is biased in finite samples.

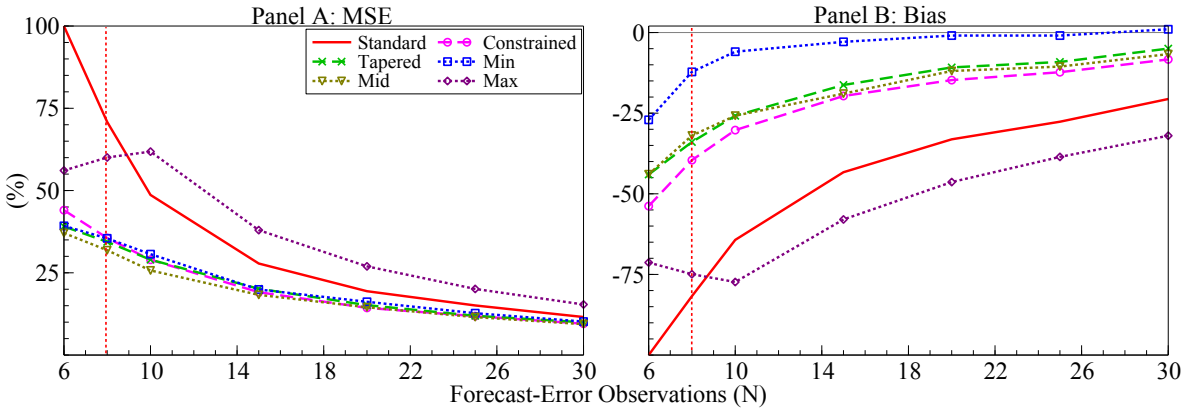


Figure 4.1: Comparing Performance Across Methods

These results are of considerable interest given the relevance to previous applications of the GFESM. While most studies shy away from the GFESM when $KH > N$, several studies including Kunst and Franses (1998) and De Gooijer and Ray (2003) come right up to this boundary using the "standard" approach. If these results can be generalized, then estimates of the GFESM using the "standard" approach at the borderline case are downward-biased.

Figure 4.1 shows that the alternative methods all remain non-singular when $KH > N$. In fact, with the exception of "max", each of the other methods has lower biases and MSEs than the "standard" approach across all values of N. The largest gains are when $KH > N$ where the MSE and bias are reduced by $25-70$ percent. While the methods cluster together, "mid" performs best in terms of MSE and "min" does best in terms of bias. Panel B in Figure 4.1 illustrates that each of the methods underestimates the GFESM but that this bias decreases as N increases. These results remain broadly similar for alternative values of $\mathbf{\Pi}$ and $\mathbf{\Omega}$ since the transformation of the forecast errors using $\widehat{\mathbf{\Psi}}_{1,H}^{-1}$ removes most of the correlation in the known-parameters case.

Figure 4.2 illustrates that the impact of an increase in K is not equivalent to that of an increase in H. In both panels, the minimum required number of forecast-error observations is 16. However, comparing Panel A of Figure 4.2 with Panel A of Figure 4.1, it is clear that the "design-free" and "tapered" methods perform well in spite of the doubling of H and a decline in the relative size of N. On the other hand, the "standard" and "constrained" approaches perform worse as the forecast

horizon increases. The "standard" approach now requires 16 forecast-error observations to be well-conditioned (compared to 8 previously), while the "constrained" approach requires eight forecast-error observations (compared to four).

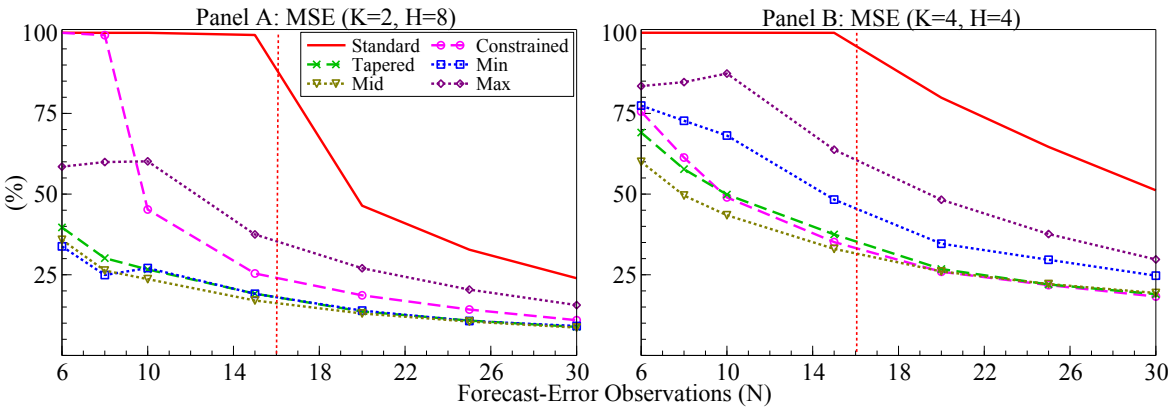


Figure 4.2: Comparing Performance Across Methods for Different H and K

Comparing Panel B of Figure 4.2 with Panel A of Figure 4.1 illustrates that a doubling of K exacerbates the underlying differences. While the MSE increases by around 20 percent relative to the baseline for all methods (Figure 4.1), there is now greater differentiation between methods, with "mid" retaining the lowest MSE, followed by the "tapered" and "constrained" methods.

5 Comparing Performance Across Forecast Models

The previous section demonstrated the relative performances of the different methods for the known-parameters model. However, it is not clear if these results hold for other forecast models. Furthermore, since the previous section only focused on one model, nothing could be said about how the performance of the methods relates to their ability to capture the actual rankings. This section extends the simulation results to the unknown parameters (VAR(1)) and random walk (RW) models. While the VAR(1) performs well when there are no structural breaks, the RW is often used as a baseline and is more robust in the presence of structural breaks.

To compare estimates of the GFESM across models, the target values, which differ for each model, need to be computed. These "true" GFESM's are described before presenting the results.

This section focuses on models from two different cases: unknown, and fixed parameters.

Unknown Parameters

In the VAR(1) model, the true parameters are unknown. Thus, the parameters (including a constant term) are re-estimated following each of the $T+n$ observations. Building on the work of Schmidt (1974), Clements and Hendry (1998, p. 75) show that when the true parameters are unknown, the forecast error variance for the h-step-ahead forecast becomes

$$\mathbb{E}_T\left[\widehat{\mathbf{u}}_{T+h|T}\widehat{\mathbf{u}}'_{T+h|T}\right] = \sum_{i=0}^{h-1}\mathbf{\Pi}^i\mathbf{\Omega}\mathbf{\Pi}^{i'} + (\boldsymbol{I}_N \otimes \mathbf{X}'_T)\mathbf{Q}(h)'\mathbb{V}[\widehat{\mathbf{\Pi}}^v]\mathbf{Q}(h)(\boldsymbol{I}_N \otimes \mathbf{X}_T), \tag{5.1}$$

where $\mathbf{Q}(h)$ is the derivative of $(\mathbf{\Pi}^h)^v$ with respect to $\mathbf{\Pi}^v$ when $(.)^v$ denotes vectoring. When the system is mapped to an I(0) representation and there are no unrestricted variables, Doornik and Hendry (2013b) argue that the parameter estimation uncertainty can be expressed as

$$\left(\sum_{i=0}^{h-1}\widehat{\mathbf{\Pi}}_1^i \otimes \widehat{\mathbf{b}}'_{T+h-i}\right)\mathbb{V}\left[\widehat{\mathbf{\Pi}}'^v\right]\left(\sum_{i=0}^{h-1}\widehat{\mathbf{\Pi}}_1^i \otimes \widehat{\mathbf{b}}'_{T+h-i}\right)', \tag{5.2}$$

if $\mathbf{b}'_{T+h-i} = \left(\mathbf{X}'_{T+h-i,h-1-i} : \mathbf{Z}'_{T+h-i}\right)$, where for $\mathbf{X}'_{a,b}$ a denotes the end and b denotes the start of the series, $\mathbf{Z}$ denotes the vector of additional variables included in the model and $\widehat{\mathbf{\Pi}}_1$ denotes the estimated parameters for $\mathbf{X}$. Equation (5.2) is added to each term in the baseline GMFESM with modifications to account for the covariances. Given that parameters are re-estimated after each additional observation, if $\mathbf{A}_{h,n} = \left(\sum_{i=0}^{h-1}\widehat{\mathbf{\Pi}}_1^i \otimes \widehat{\mathbf{b}}'_{T+n+h-i}\right)$ and $\widetilde{\mathbf{V}}_N = \mathbb{V}\left[\widehat{\mathbf{\Pi}}'^v\right]$, the "true" GMFESM augments (4.4) by

$$\begin{pmatrix} \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{A}_{1,n}\widetilde{\mathbf{V}}_n\mathbf{A}_{1,n} & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{A}_{1,n}\widetilde{\mathbf{V}}_n\mathbf{A}_{H,n} \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{A}_{H,n}\widetilde{\mathbf{V}}_n\mathbf{A}_{1,n} & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{A}_{H,n}\widetilde{\mathbf{V}}_n\mathbf{A}_{H,n} \end{pmatrix}. \tag{5.3}$$

Fixed Parameters

The RW model fixes the parameters. Thus, while there is no parameter estimation uncertainty, the estimate of the "true" GFESM needs to account for the difference between the actual and fixed parameters. The RW model does not change the VAR(1) other than to fix the parameter values, denoted as $\bar{\mathbf{\Pi}}$. When $\bar{\mathbf{\Pi}} = \boldsymbol{I}_K$, then the forecast error variance is

$$\mathbb{E}_T\left[\widehat{\mathbf{u}}_{T+h|T}\widehat{\mathbf{u}}'_{T+h|T}|\mathbf{X}_T\right] = \left(\mathbf{\Pi}^h - \boldsymbol{I}_K\right)\mathbf{X}_T\mathbf{X}'_T\left(\mathbf{\Pi}^h - \boldsymbol{I}_K\right)' + \sum_{i=0}^{h-1}\mathbf{\Pi}^i\mathbf{\Omega}\mathbf{\Pi}^{i'}. \tag{5.4}$$

Rather than taking into account the parameter uncertainty, this modified equation takes into account the difference between the values of the true and fixed parameters. As the persistence increases towards a unit root ($\mathbf{\Pi} \rightarrow \boldsymbol{I}_K$), the first term goes to zero and the forecast error decreases towards the known-parameters model. Thus, if $\mathbf{J}_n = \mathbf{X}_{T+n}\mathbf{X}'_{T+n}$ and $\mathbf{F}_h = \left(\mathbf{\Pi}^h - \boldsymbol{I}_K\right)$, the "true" GMFESM for the RW model augments (4.4) with

$$\begin{pmatrix} \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{F}_1\mathbf{J}_n\mathbf{F}_1 & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{F}_1\mathbf{J}_n\mathbf{F}_H \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{F}_H\mathbf{J}_n\mathbf{F}_1 & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{F}_H\mathbf{J}_n\mathbf{F}_H \end{pmatrix}. \tag{5.5}$$

Equations (5.3) and (5.5) augment (4.4) in different ways to give the "true" GMFESM for the respective forecast models, the determinants of which are the "true" GFESMs that are used as baselines to compare the various estimators. As in the known-parameters case, the error variance in each of these cases is estimated from the correctly specified model in (4.5).

The GFESM is calculated using the forecast errors from (2.3) when $p = 1$, for the known parameters, the VAR(1), and the RW models with modifications for how $\widehat{\mathbf{\Pi}}$ is treated and whether a constant is estimated. The forecast errors for each model are stacked into matrices following (2.5), transformed by the estimated parameter matrix in (3.1) and then used to calculate the GFESM following the methods described in section 4.

Equations (5.3) and (5.5) illustrate that when forecast models are misspecified, it is no longer straightforward to decompose the GMFESM as in (3.2). As a result, transforming the forecast errors as described in section 3 may not remove all serial correlation from the forecast errors of misspecified models. An IID test on the transformed forecast errors could be used. With the correct transformation, the null hypothesis of IID forecast errors will not be rejected. When the residuals are normally distributed then testing for normality and no autocorrelation is akin to testing the IID assumption. In the following simulations, the null hypotheses of normality and no autocorrelation are not rejected across all models and for all values of N. They are rejected in the presence of fat-tails and shifts in the distribution. Therefore, it is useful to extract shifts and outliers as discussed in section 3.

5.1 Simulation Results

Figure 5.1 presents the simulation results for the alternative forecast models. Panel A presents the results for the VAR(1) while panel B presents those for the RW. They illustrate that the various methods perform differently across models. The results for the VAR(1) in Panel A of Figure 5.1 follow the known-parameters model in Panel A of Figure 4.1 very closely. This stems from the fact that the VAR(1) is well-specified. It also suggests that the parameter estimation uncertainty term from (5.3) does not impact estimates of the GFESM substantially. Note that differences are larger when the estimation period (T) is limited.

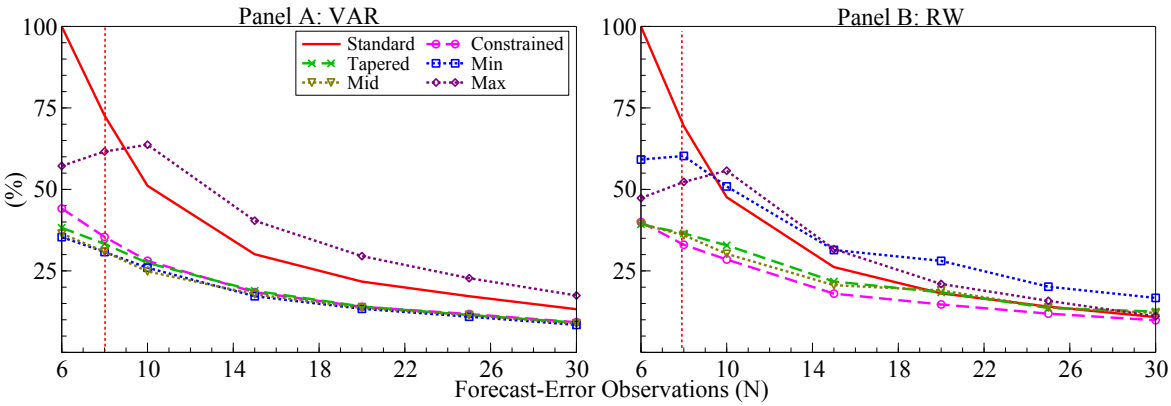


Figure 5.1: Comparing MSE Performance Across Methods and Models

The results for the RW model in Panel B of Figure 5.1 illustrates that the previous results do

not extend completely to misspecified models. In this case, the forecasts are misspecified such that the forecast error increases with the forecast horizon. Thus, heteroskedasticity increases along the diagonal of the GMFESM; see (5.5). As a result, a homogeneous estimate of the eigenvectors (i.e. smaller s) does poorly, meaning that increasing the size of the first-step estimator (i.e. higher s) improves the estimates, which is why "min" deteriorates. "Mid" and the "tapered" method do well when $KH > N$, while "constrained" performs best for $KH \leq N$. This is due to the fact that differences on the diagonal outweigh additional covariance terms, particularly with more observations. Increasing the cross-variable correlations and decreasing the differences along the diagonal of the forecast errors eliminates some of these advantages. Simulations for alternatively misspecified models suggests that these results continue to hold.

5.2 Ranking the Forecast Models

The analysis has focused on how well the GFESM can be estimated. However, it is important that estimates of the GFESM provide a valid ranking of forecast models. In fact, poor estimates can still provide informative rankings of the forecasts. The remaining analysis in this section focuses on how well the methods rank the forecasts. The DGP evaluated here contains two variables and four forecast horizons. Unlike the previous results, this analysis focuses on the share of simulations for which the method ranks the forecast models correctly. Table 5.1 presents the rankings across different values of N. For the sake of simplicity three cases are considered: the singular case ($N = 6$), the borderline case ($N = 8$), and the non-singular case ($N = 30$). Rows present the methods and the minor columns represent different forecast models. The "true" GFESM ranking across all three cases from best to worst is: known-parameters, VAR(1) and RW.

Consider how well the different methods rank the forecasts. In the first major column of Table 5.1, the "standard" approach is unable to distinguish between models, since $KH > N$. However, each of the other methods provides a ranking. Similar to the results in sections 3 and 4, while all methods perform closely, "min" and "tapered" correctly specify the complete ranking of the models most often

Table 5.1: Share of Simulations that Correctly Rank Forecast Models, by Method

Method	$KH > N: N = 6$				$KH = N: N = 8$				$KH < N: N = 30$			
	Known	VAR(1)	RW	Total	Known	VAR(1)	RW	Total	Known	VAR(1)	RW	Total
Rank:	1	2	3		1	2	3		1	2	3	
"standard"	-	-	-	-	39.6	42.5	58.5	31.5	72.9	71.8	97.1	71.3
"min"	57.8	49.6	82.3	48.0	59.6	53.6	86.3	52.6	73.2	71.6	97.3	71.3
"mid"	59.1	50.5	82.2	49.1	60.1	53.5	85.5	52.3	72.6	70.7	97.2	70.5
"max"	56.2	48.3	79.8	46.2	55.9	50.6	79.4	47.8	71.2	68.7	95.1	68.2
"constrained"	51.8	48.5	73.2	42.7	52.8	48.0	78.0	45.4	72.2	70.5	96.4	70.1
"tapered"	57.5	50.8	80.6	48.9	57.7	52.3	84.4	50.9	73.4	70.9	95.8	70.5
Trace	55.1	49.4	75.5	46.2	57.4	51.8	79.3	48.9	74.0	72.3	97.3	72.0

Notes: Shares are based on 1,000 replications.

at 49 percent, whereas the constrained method performs worst with 43 percent. Note that the Trace does well at 46 percent but does not outperform the "design-free" method when $KH > N$. Overall the shares are low since all methods have difficulty distinguishing between the known-parameters and VAR(1) models with so few observations. However, the "design-free" method is best at distinguishing between misspecified models (RW).

The second major column of Table 5.1 presents the borderline case where $KH = N$. While the "standard" approach provides rankings across all models, it is poor at distinguishing between them in that it only produces the correct ranking 32 percent of the time. This is worse than the other approaches even if they have fewer observations. Now "min" and "mid" perform best by correctly ranking the models 53 percent of the time, whereas the Trace and the "constrained" approach correctly rank the forecasts 49 and 45 percent of the time respectively.

The non-singular case where $KH < N$ is presented in the third major column of Table 5.1. All of the methods perform approximately similarly: the Trace, "standard", "min", "mid" and "tapered" methods" specify the complete ranking around 72 percent of the time, while the "standard" and "max" approaches perform slightly worse with 70 and 68 percent respectively. The Trace ranks the forecast models relatively well in these simulations, but previous studies have found that rankings can differ substantially between the GFESM and the Trace; see Clements and Hendry (1997).

These results illustrate that the "design-free" method is best able to distinguish between different models when N is relatively small. Furthermore, when N is relatively small, an important concern

is the ability to distinguish between well-specified models. This is especially true when there is a structural break at the end of the estimation sample. We consider this case in the next section.

6 Allowing for Structural Breaks

In this section we modify the DGP to allow for breaks so that it becomes

$$\mathbf{X}_t = \mathbf{\Pi}\mathbf{X}_{t-1} + \boldsymbol{\delta}_{(t \geq t^*)} + \mathbf{v}_t, \tag{6.1}$$

where $\boldsymbol{\delta}$ is a $(K \times 1)$ vector that shifts the mean when $t \geq t^*$ where t^* is the break point. If there is no break then $\boldsymbol{\delta} = \mathbf{0}$. If there is a break, then $\boldsymbol{\delta}$ is non-zero for all variables K when $t \geq t^*$.

In addition to modifying the DGP we also consider two other models, the intercept corrected (IC) model and the robust random walk (RRW) model. The IC model differs from the unknown parameters VAR model in that it estimates an additional constant term starting at $t*$ to coincide with the break date. As a result, the forecast-error variance in the no-break case is identical to the VAR case except that there is an additional vector of variables that add to the parameter estimation uncertainty. The RRW model adds another lagged dependent variable in (6.1) so the h-step-ahead forecast is

$$\widehat{\mathbf{X}}_{T+h|T} = \widehat{\mathbf{X}}_{T+h-1|T} + \Delta\widehat{\mathbf{X}}_{T+h-1|T} = \mathbf{X}_T + \boldsymbol{h}\Delta\mathbf{X}_T, \tag{6.2}$$

where the fact that $\bar{\mathbf{\Pi}} = \boldsymbol{I}_K$ is used throughout. Then the h-step-ahead forecast error becomes

$$\begin{aligned} \widehat{\mathbf{u}}_{T+h|T} &= \left(\mathbf{\Pi}^h - \boldsymbol{I}_K\right)\mathbf{X}_T - \boldsymbol{h}\Delta\mathbf{X}_T + \sum_{i=0}^{h-1}\mathbf{\Pi}^i\left(\mathbf{v}_{T+h-i} + \boldsymbol{\delta}\right) \\ &= \left(\left(\mathbf{\Pi}^h - \boldsymbol{I}_K\right)\mathbf{\Pi} + \boldsymbol{h}(\boldsymbol{I}_K - \mathbf{\Pi})\right)\mathbf{X}_{T-1} + \sum_{i=0}^{h}\mathbf{\Pi}^i\left(\mathbf{v}_{T+h-i} + \boldsymbol{\delta}\right) - (\boldsymbol{h}+\mathbf{1})(\mathbf{v}_T + \boldsymbol{\delta}). \end{aligned} \tag{6.3}$$

The first line of (6.3) is similar to the forecast error for the RW model with an additional bias term. However, unlike the RW model, an additional error term remains even as $\mathbf{\Pi} \rightarrow \boldsymbol{I}_K$. If

$\mathbf{D}_h = \left(\mathbf{\Pi}^{h+1} - \mathbf{\Pi} + \boldsymbol{h}(\boldsymbol{I}_K - \mathbf{\Pi})\right)$, $\mathbf{J}_N = \mathbf{X}_{T+N}\mathbf{X}'_{T+N}$ and $\mathbf{B}_h = (\boldsymbol{h}+\mathbf{1})\boldsymbol{I}_K$, then the forecast error variance for an h-step-ahead RRW forecast is

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathbf{u}}_{T+h|T}\widehat{\mathbf{u}}'_{T+h|T}|\mathbf{X}_T\right] = \quad & \sum_{i=0}^{h-1}\mathbf{\Pi}^i\mathbf{\Omega}\mathbf{\Pi}^{i'} + \mathbf{B}_h\mathbf{\Omega}\mathbf{B}_h' + \mathbf{D}_h\mathbf{J}_{-1}\mathbf{D}_h' + \left(\mathbf{B}_h - \sum_{i=0}^{h}\mathbf{\Pi}^i\right)\boldsymbol{\delta}\boldsymbol{\delta}'\left(\mathbf{B}_h - \sum_{i=0}^{h}\mathbf{\Pi}^i\right)' \\ & - \left(\mathbf{B}_h - \sum_{i=0}^{h}\mathbf{\Pi}^i\right)\boldsymbol{\delta}\mathbf{X}'_{T-1}\mathbf{D}_h' - \mathbf{D}_h\mathbf{X}_{T-1}\boldsymbol{\delta}'\left(\mathbf{B}_h - \sum_{i=0}^{h}\mathbf{\Pi}^i\right)'. \end{aligned} \tag{6.4}$$

This illustrates that since the parameters are fixed, a structural break only impacts the true GFESM through the squared bias terms. In the absence of a structural break (i.e. $\boldsymbol{\delta} = \mathbf{0}$), the true GMFESM for the RRW model augments (4.4) with

$$\begin{pmatrix} \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{D}_1\mathbf{J}_{n-1}\mathbf{D}_1' & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{D}_1\mathbf{J}_{n-1}\mathbf{D}_H' \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{D}_H\mathbf{J}_{n-1}\mathbf{D}_1' & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{D}_H\mathbf{J}_{n-1}\mathbf{D}_H' \end{pmatrix} + \begin{pmatrix} \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{B}_1\widehat{\mathbf{\Omega}}_n\mathbf{B}_1' & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{B}_1\widehat{\mathbf{\Omega}}_n\mathbf{B}_H' \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{B}_H\widehat{\mathbf{\Omega}}_n\mathbf{B}_1' & \cdots & \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{B}_H\widehat{\mathbf{\Omega}}_n\mathbf{B}_H' \end{pmatrix}. \tag{6.5}$$

Consider a 5 standard deviation shift in the mean ($\boldsymbol{\delta} = \mathbf{5}$) two periods before the first forecast ($t^* = T-2$). While the known parameters model is unaffected, all other models experience some degree of forecast failure. Figure 6.1 provides evidence that the "design-free" estimator is not robust to misspecified forecast models when s is small. The performance of the estimators does not substantially change for the robust forecast models, except that the RRW is hard to estimate given how poorly it does, as seen from Panel D in 6.1. Performance does change for the VAR such that "min" initially performs worse than the "standard" estimator. These results suggest that the non-constancy in the DGP induces an increase in the heterogeneity across the diagonal of the GMFESM in non-robust models such that it shifts the trade-off between estimating the eigenvalues and eigenvectors in favor of estimating the eigenvectors. Generally speaking choosing a mid-ranged s for the "design-free" estimator is more robust to the presence of structural breaks and misspecification.

Overall these results illustrate the benefit of using the "design-free" estimator across various mod-

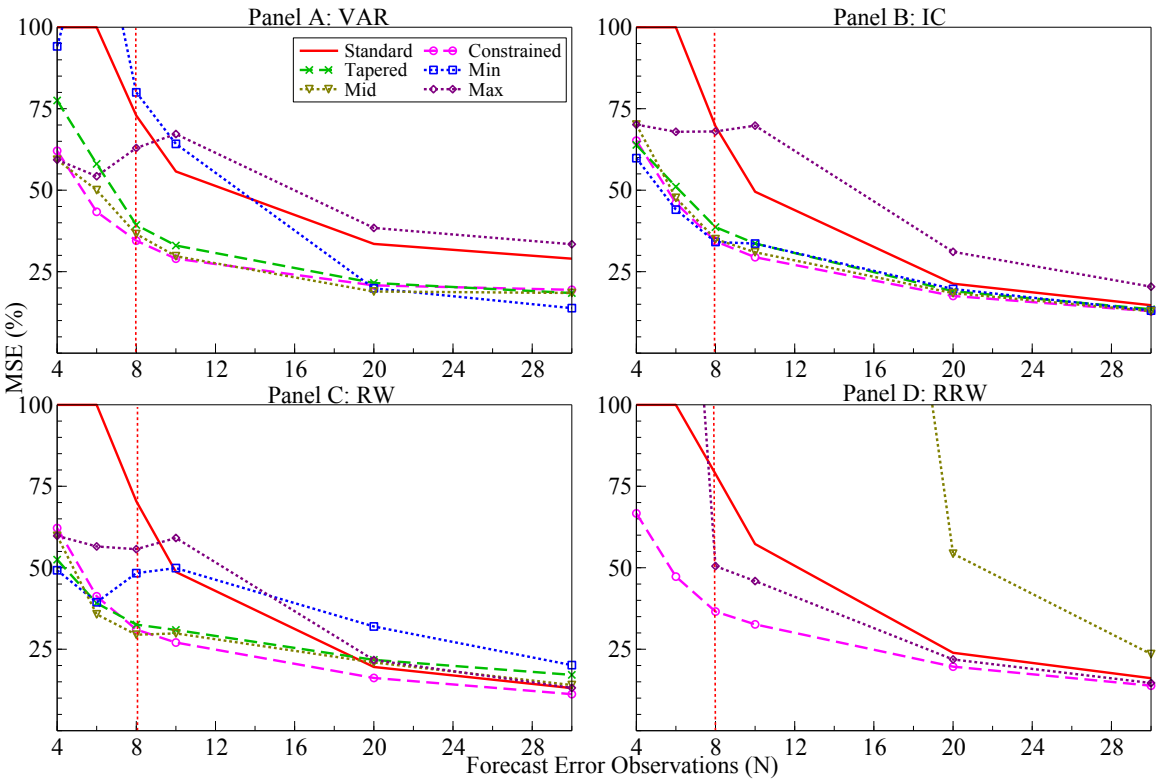


Figure 6.1: Comparing MSE Performance Across Methods and Models with a Structural Break

els. These benefits are particularly strong for the borderline case where the "standard" estimator does poorly. The "design-free" estimator generally outperforms the "standard" estimator where "min" typically dominates in terms of MSE and bias for well-specified models while "mid" and "max" perform better for misspecified forecast models. The "design-free" estimator also performs well against the alternative estimators considered here. In well-specified models, all of the estimators generally fall in the same range, while for misspecified models alternative approaches only improve upon the "design-free" estimator when their imposed restrictions are valid.

Now consider what a break does to the estimated rankings. Table 6.1 presents the estimated rankings for this case. The "true" rankings illustrate a more complicated picture in that the rankings change with the observations. When $(K \times H) > N$, the "true" ranking from best to worst is: known parameters, RW, IC, VAR and RRW. When $(K \times H) \leq N$, this changes such that the IC model does better than the RW model. Thus, the GFESM punishes the IC model after the break due to the decline in error covariance across horizons. As observations increase after the break, the IC model's error covariance increases and so its GFESM declines.

Table 6.1: Share of Simulations that Correctly Rank Models with a Structural Break, by Method

Method	$KH > N: N = 6$						$KH = N: N = 8$						$KH < N: N = 30$					
	Known	VAR	RW	IC	RRW	Total	Known	VAR	RW	IC	RRW	Total	Known	VAR	RW	IC	RRW	Total
Rank:	1	4	3	2	5	-	1	4	2	3	5	-	1	42	2	3	5	-
"standard"	-	-	-	-	-	-	54.8	36.1	27.9	34.9	65.5	12.2	96.6	95.7	78.4	78.5	100	75.3
"min"	93.6	86.4	53.7	53.7	98.3	48.6	95.6	93.4	57.4	57.0	99.8	53.8	97.3	95.6	78.2	79.0	100	76.0
"mid"	94.7	91.1	55.0	53.8	99.5	50.7	94.9	93.2	57.2	55.6	99.8	53.0	97.5	95.5	78.7	79.4	100	76.5
"max"	92.4	88.5	53.8	52.1	98.6	47.0	89.0	85.3	53.3	54.1	98.7	47.7	96.0	92.3	75.1	77.2	100	71.9
"const."	81.7	69.3	47.0	49.8	90.6	35.8	87.0	81.4	53.9	54.2	97.5	46.1	97.4	94.3	78.4	80.4	100	76.1
"tapered"	91.3	84.3	52.6	54.0	98.5	46.6	91.2	86.8	55.7	56.3	99.9	50.5	97.5	93.1	74.9	78.4	100	73.0
Trace	91.5	61.4	47.5	53.5	73.7	28.8	93.0	74.1	56.4	57.7	82.1	39.7	98.2	98.8	82.2	80.6	99.9	80.3

Notes: Shares are based on 1,000 replications.

None of the estimators perfectly captures the ranking of the forecasts in the singular case. While the estimators do well capturing both the best and the worst forecast models such that they almost always correctly rank the known parameters, the RRW, and the VAR models, they have a hard time correctly ranking the IC and RW models. This relates back to the inability to distinguish between well-specified forecast models in small samples. Overall, in the singular case the "mid" estimator produces completely correct rankings most often (just over 50% of the time). In the borderline case the estimators continue to struggle distinguishing between the RW and IC models with the "min" producing completely correct rankings almost 54% of the time. Once there are a large number of observations then most methods are able to generate correct rankings most of the time.

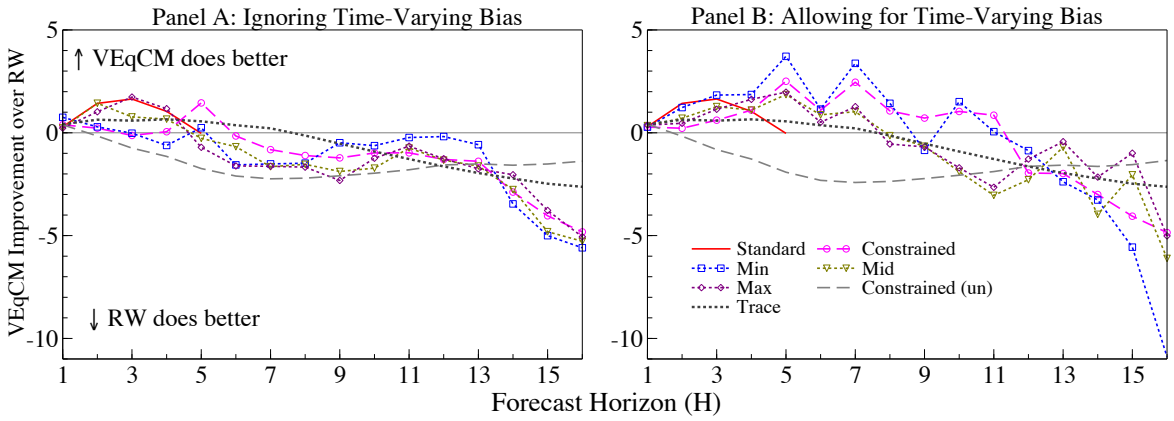
7 Evaluating a Vector of Forecasts for the US Economy

This section recreates the Anderson et al. (2002) analysis of a vector of forecasts of the US economy, and compares the recreated results with the "design-free" method. Following Anderson et al. (2002), forecasts are generated by a vector equilibrium correction model (VEqCM) and a RW model for six variables up to 16 quarters ahead from 1989 until 1997. Thus, 96 forecast-error observations are required to evaluate the forecasts using the "standard" approach, but a maximum of 36 forecast-error observations are available. Anderson et al. (2002) use the "constrained" method to compute reductions in the VEqCM's GFESM relative to the RW.

While this analysis uses the original data to estimate a VEqCM, generate forecasts that exhibit the same general features and the same forecast rankings for individual variables and horizons, the

original GFESM results could not be replicated exactly; see Anderson et al. (2002, Table 6-8). See Appendix A for details on how the VEqCM is estimated. Using the methods described above, the improvements of the VEqCM over the RW in terms of the log GFESM and the Trace (which is calculated here as the Trace of the log GMFESM) for 1 – 16 quarters ahead forecasts were computed.

Figure 7.1 presents the results from this analysis. Panel A presents the case where outliers and time-varying bias are ignored. Panel B presents the results for when outliers and time-varying bias are taken into account using indicator saturation methods. The null hypotheses of normality and no autocorrelation are not rejected across all horizons once the forecast errors are transformed and any time-varying bias is removed. Although the "standard" approach is only available through 5 horizons, all of the other methods can be calculated through 16 horizons.[12]



Note: The improvement in the GFESM methods is calculated as the difference between the log GFESMs. The improvement in the Trace is calculated as the difference between the traces of the log GMFESM. Constrained (un) refers to the non-transformed version of the "constrained" approach.

Figure 7.1: VEqCM Forecast Performance for the U.S. Economy, 1989-97

The results illustrate several important points. First, outliers and time-varying bias can have a substantial impact on the results. When potential outliers and time-varying bias are ignored then most methods indicate that the VEqCM only outperforms the RW for up to one-year-ahead. However when taking potential outliers into account all the methods indicate that the VEqCM outperforms the RW for up to two years ahead. This suggests that it is important to account for potential outliers and time-varying biases when using these methods.

[12]The "tapered" method is excluded since it produces very volatile results.

The second important result is that transforming the forecast errors has a substantial impact. While the "constrained" approach using the transformed forecast errors follows the "design-free" method closely in both panels, note that the "constrained" approach without transforming the forecast errors is considerably different from the other methods (especially when allowing for outliers and shifts) in that it indicates that the VEqCM only outperforms the RW up to 1-quarter-ahead. This strongly suggests that the assumptions imposed by the "constrained" approach are not valid in this application. Note that the Trace gives similar rankings across horizons but underestimates the magnitude of the differences between the forecast models.

8 Conclusions

This chapter develops a new approach for evaluating multi-step system forecasts when there are relatively few forecast-error observations. In the process, it proposes a solution to one of the long-standing limitations of the GFESM. Section 3 shows how the approach developed by Abadir et al. (2014) can be generalized to estimate the GMFESM by transforming the stacked forecast errors. Combining this approach with Clements and Hendry (1993) allows for precise estimation of the GFESM even when $KH > N$.

The simulations in sections 4-7 show that the "standard" approach is downwards biased and provides poor rankings of the forecast models when there are relatively few observations. This has large implications for the previous analyses which applied the GFESM with relatively few forecast-error observations. The simulations also illustrate that the "design-free" method substantially reduces the bias in both the singular and borderline cases and is the most likely approach to provide correct rankings of the forecast models.

While there is limited differentiation between the alternative methods for well-specified models, this is not the case for misspecified models or for larger systems. Overall, both the simulations and the empirical results suggest that the "mid" approach is best in the largest number of situations. While the "constrained" approach does well when the forecast errors are transformed, the empirical

application illustrates that its performance deteriorates substantially when the forecast errors are not transformed. The empirical example also illustrates the importance of accounting for outliers and time-varying bias when applying these methods.

Finally, while the trace MSFE performed relatively well in both the simulations and the empirical application, it is important to note that it too suffers from information loss when there are relatively few forecast error observations. Furthermore, a wide range of situations exist in the literature in which the rankings of the trace MSFE differ from the GFESM. Overall, this illustrates the benefits of being able to extend estimates of the GFESM to cases when there are more variables times horizons than forecast-error observations.

A Appendix: Estimating the VEqCM

This appendix re-estimates the forecasting model and forecasts as specified in Anderson et al. (2002). It starts by describing the data used and then recreates and tests the vector equilibrium correction model (VEqCM) which is then used to generate the forecasts. An initial comparison of the forecasts from the VEqCM and random walk (RW) models is done using root mean square errors.

A.1 Data Sources and Description

This analysis follows Anderson et al. (2002) with the same six variables that were included in the original analysis. These variables are the log of real GDP (gdp); the log of real M1 money balances (m1p); two measures of inflation: the first differences of the log of the GDP deflator (infdef) and of the CPI (infcpi) both expressed as percent annual rates; the constant-maturity yield on 10-year Treasury securities (lrate) and the federal funds rate (funds).[13] While the original analysis does not specify whether the variables used are seasonally adjusted or not, here seasonally adjusted are used when they are available. The data are quarterly and span 1955 Q1-1998 Q2.

The data originally come from the Federal Reserve Bank of St. Louis' database (FRED) and were as of December 1998. Here vintage data from the Federal Reserve Bank of St. Louis' archival database (ALFRED) as of the last revision prior to 31 December 1998 are used. ALFRED and FRED only contain M1 data starting in 1958. As a result, the data for M1 from 1955 to 1958 are based on calculations in Rasche (1987). Additionally, M1 is augmented by estimates of the amount of transaction deposits swept by banks into money market deposit accounts, beginning in 1994.[14]

The six variables are all plotted in Figure 6.1. The figures illustrate some important properties about the data. Most noticeably are the peaks in the various interest rate measures and inflation measures around the mid-1970s and the early 1980s with inflation annual percent rate of inflation

[13] CPI is available monthly. There are several ways to compute the first differences of the logs. One is to use the average log CPI over the months in a given quarter to compute the first differences of the logs. Alternatively one can take the difference between log of the last month of the previous quarter and the last month of the current quarter. While the latter suffers from fewer methodological concerns, the former is used since the results are closer to Anderson et al. (2002).

[14] The early data and the data on sweeps can be found at http://research.stlouisfed.org/aggreg/. The sweeps data was discontinued in 2002 and there there are no historical vintages of these estimates available.

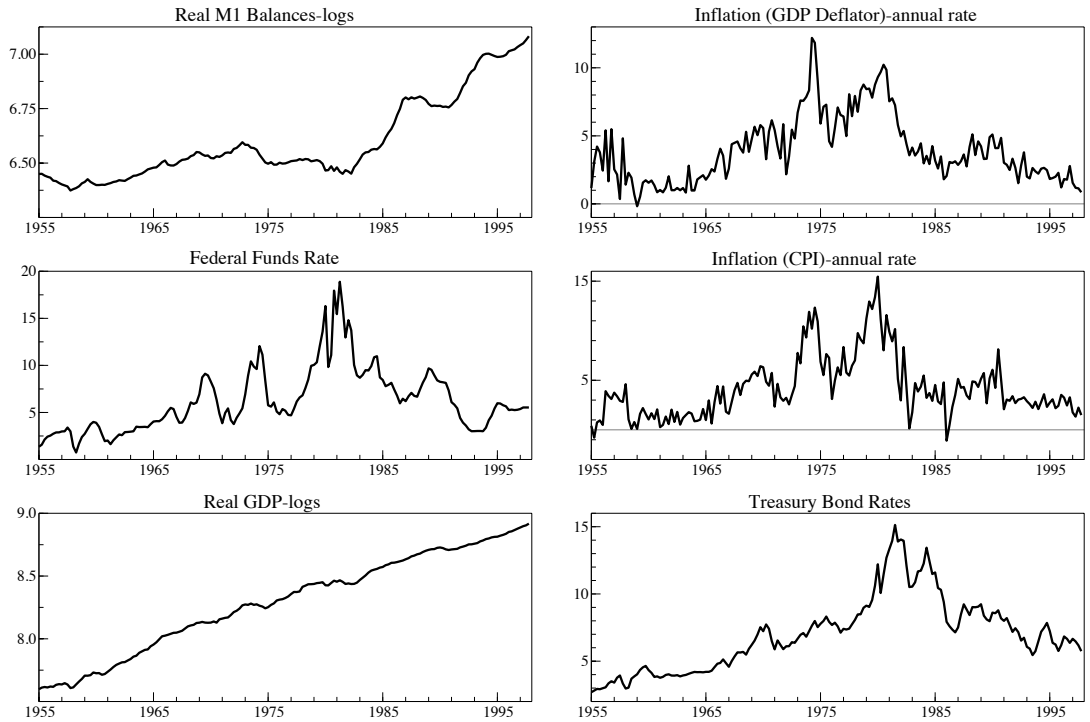


Figure A.1: Quarterly Macroeconomic Time Series for the U.S., 1955-1997

reaching as high as 15 percent in both cases depending on which measure of inflation was used. Interestingly, the long-term interest rate embodied in the Treasury Bond rates only shows the second peak in the early 1980s which suggests that long-term expectations about inflation were relatively low during the mid-1970s. Another interesting pattern to note is that, starting in the mid-1980s, a decline in interest rates and inflation appears to be associated with an increase in real money balances.

In addition to the six variables, several dummy variables are included. A step-shift variable ("D79") whose value is 1 starting in 1979 Q4 was included in order to capture the beginning of the Federal Reserve's "New Operating Procedure". Furthermore, the period 1979 Q4 to 1981 Q4 is "omitted" from the analysis using dummy variables. Anderson et al. (2002, p. 575) argues that the exclusion of these observations is necessary "because the semi-log functional form used in this model does not have adequate curvature to approximate M1 velocity in a range of nominal interest rates about 10%".[15] Including a dummy variable for each quarter between 1979 Q4 and 1981 Q4 effectively omits these observations while also allowing for the ability to test these restrictions. Testing

[15] See Baba et al. (1992) for an alternative approach to dealing with this issue.

first whether the step-shift variable actually captures a regime change for the whole system gives an F-statistic of 4.74 which is significant at the 1 percent level. It is possible to test whether all of the omitted observations are jointly significant for the whole system. This restriction is also significant with an F-statistic of 6.20. Each of the individual dummies can also be tested to see whether they are significant. The results indicate that the dummies for 1979 Q4 and 1980 Q1 are individually and jointly insignificant at the 1 percent level. However, even when including a step-shift dummy for the New Operation Procedure and using dummies for the observations 1979 Q4 to 1981 Q4, the individual series and the vector as a whole still fail diagnostic tests for residual normality and no autocorrelation. While Impulse Indicator Saturation (IIS) would provide a more general approach to correctly specify the model, it is not applied in the interest of maintaining comparability with the original analysis. See Hendry and Mizon (2011) for an application of IIS in this context.

A.2 A Forecast Model of the US Economy

The model is a reduced rank VAR following Johansen and Juselius (1990) and Johansen (1991):

$$\mathbf{X}_t = \sum_{i=1}^{l} \mathbf{\Pi}_i \mathbf{X}_{t-i} + \Upsilon \mathbf{D}_t + \mathbf{v}_t, \tag{A.1}$$

where D_t captures any deterministic terms (i.e. dummies and constant) and l denotes the maximum lag length. The next step is to determine l. In general, with quarterly data it is good to start with at least 4 to 8 lags to capture the dynamics. Anderson et al. (2002) specifies a model with 6 lags. A lag structure analysis suggests that this is the correct specification.

Once the correct lag structure is chosen, it is important to determine whether the system is cointegrated or not. (A.1) can be rewritten in vector equilibrium correction form by subtracting X_{t-1} from both sides and rearranging to give

$$\Delta \mathbf{X}_t = \mathbf{\Pi} \mathbf{X}_{t-1} + \sum_{i=1}^{l-1} \mathbf{\Gamma}_i \Delta \mathbf{X}_{t-i} + \Upsilon \mathbf{D}_t + \mathbf{v}_t, \tag{A.2}$$

where $\mathbf{\Gamma}_i = -\sum_{j=i+1}^{l} \mathbf{\Pi}_j$ and $\mathbf{\Pi} = \left(\sum_{i=1}^{l} \mathbf{\Pi}_i - \mathbf{I}_K\right)$. Testing for cointegration implies testing the rank of $\mathbf{\Pi}$. As illustrated in Johansen and Juselius (1990) $rank\,(\mathbf{\Pi}) = K$ means that the parameter matrix has full rank which indicates that the whole process is stationary. On the other hand, $rank\,(\mathbf{\Pi}) = 0$ indicates that all of the variables are I(1) processes and the correct specification is the differenced model. Finally, $0 < rank\,(\mathbf{\Pi}) = r < K$ indicates that the variables in the system are cointegrated and can be further decomposed into $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ represents how quickly the short-run model adjusts back to the long-run relationship (i.e. the equilibrium correction term), while $\boldsymbol{\beta}$ gives the long-run relationship (i.e. the cointegrating space) between the different variables.

Anderson et al. (2002) use economic theory to hypothesize that there are four long-run (cointegrating) relationships between the variables (i.e. $r = 4$). It is possible to test this hypothesis using a reduced rank restriction or trace test. When there are $r = 4$ cointegrating relations, $r - 1 = 3$ restrictions satisfy the conditions for identification for each of the cointegrating relations. Tests support the hypothesis of cointegration between the six variables over 1957 Q1-1997 Q4 in that the trace test statistic has a value of 24.5 for $r \leq 3$. This is somewhat lower than the trace test statistic of 26.1 in the original analysis. However, the null hypothesis is not rejected at the 5 percent level.

Additional restrictions are imposed on the cointegrating space using economic theory. The four hypothesized long-run relationships are: the money demand equation, two different Fisher relations and the term structure spread. To illustrate, the vector of variables can be ordered as

$$\boldsymbol{X}_t' = \{m1p_t, infdef_t, funds_t, infcpi_t, gdp_t, lrate_t\}.$$

The cointegration space that reflects the four hypothesized long-run relations is

$$\boldsymbol{\beta}' = \begin{bmatrix} \underline{\underline{1}} & \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{0}} & \underline{\text{-1}} & \beta_1 \\ \underline{\underline{0}} & \underline{\underline{1}} & \underline{\underline{0}} & \underline{\underline{0}} & \underline{0} & \underline{\text{-1}} \\ \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{1}} & \underline{\underline{0}} & \underline{0} & \underline{\text{-1}} \\ \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{1}} & \underline{0} & \underline{\text{-1}} \end{bmatrix} \tag{A.3}$$

where identification restrictions appear with a double underscore and the 'ones' down the main diagonal are normalizations. The first row in (A.3) represents the long-run money demand relation ($mdciv$) which assumes a 1-to-1 relationship between real output and real money balances that is moderated by the long-term interest rate. The second and fourth rows in (A.3) represent two different Fisher relationships using the two measures of inflation in the system ($defrciv$ and $cpirciv$ respectively). The third row captures the term-structure spread linking the short-term Federal funds rate and the long-term government bond rate ($termciv$). Anderson et al. (2002, p. 573) illustrate that this representation of the cointegration space is not unique and alternative normalizations "yield long-run relations that embody [the] system of cointegrating relations".

Table A.1: Recursive Estimates of the Semi-Elasticity of Money Demand

Sample ending	Estimated semi-elasticity		Imposing the semi-elasticity
	coeff and se	LR test on restrictions	LR test on restriction
88:4	0.087 (0.007)	13.36	0.25
89:4	0.086 (0.007)	15.25*	0.35
90:4	0.084 (0.007)	12.27	0.01
91:4	0.084 (0.007)	17.00*	0.01
92:4	0.086 (0.007)	18.78**	0.02
93:4	0.085 (0.006)	18.92**	0.06
94:4	0.085 (0.006)	17.96*	0.02
95:4	0.083 (0.006)	22.34**	0.00
96:4	0.085 (0.006)	13.21	0.01
97:4	0.085 (0.006)	14.17*	0.01

Notes: The last column presents the test statistics for 1 additional restriction.
* denotes rejection at the 5% level while ** denotes rejection at the 1% level.

Rather than testing each of the restrictions, we can use economic theory to specify the coefficients in the cointegrating space. As (A.3) shows, 16 restrictions are used to identify the 4 cointegrating vectors (double underscore). Additionally, 7 over-identifying restrictions are imposed on the cointe-

grating space, which appear with a single underscore in (A.3). These give the long-run money demand relation, the two Fisher equations, and the term structure. One coefficient in the cointegrating space is left unrestricted in order to provide an estimated interest rate semi-elasticity. In their analysis, Anderson et al. (2002) find that the likelihood ratio test statistic of these restrictions to be 14.12 while the current analysis finds a likelihood ratio test statistic of 14.17. In both cases, the imposed restrictions are weakly rejected at a critical value of 5 percent. The original estimated interest rate semi-elasticity is $\hat{\beta}_1 = 0.085$ with a standard error of 0.005. While we find the same semi-elasticity with a slightly higher standard error of 0.006 (See the last row of Table A.1).

Anderson et al. (2002) evaluate the stability of the system by examining the constancy of the estimated interest rate semi-elasticity. This analysis is replicated in Table A.1 from 1988 Q4 through 1997 Q4. The estimated coefficients of the semi-elasticity are generally in line with the original results. In fact, a slightly narrower range on the estimated semi-elasticities is found (0.083-0.087 vs. 0.082-0.091). We can also evaluate the constancy of the imposed restrictions on subsamples for each re-estimate of the model. While repeated testing is a concern, a 1 percent level requires 100 tests in order to reject one test by chance. The results show that for samples ending in 1992 Q4 through 1995 Q4 the imposed restrictions are strongly rejected. Thus, either the imposed restrictions are not valid for those subsamples or there may be omitted variables that could capture the instability during this period. Finally, it is also possible to test whether our estimate significantly different from Anderson et al. (2002) by imposing one additional restriction on the cointegrating space and then using a likelihood ratio test. The test statistics in the last column of Table A.1 illustrate that there is no statistical difference between estimating and imposing the original semi-elasticity.

Once the cointegrating relation is imposed and the VEqCM is formulated, it is possible to test whether any of the lagged variables can be excluded from any of the equations in the system. Table A.2 presents the F-statistics for excluding all lags of a given variable from a specific equation. Here the results depart from the original findings in Anderson et al. (2002, Table 3). The removal of the lags of real money balances, the federal funds rate and the long-term interest rate from the real money

Table A.2: F-tests for Excluding Lagged Variables

	Dependent variable (equation)					
	$\Delta m1p$	$\Delta infdef$	$\Delta funds$	$\Delta infcpi$	Δgdp	$\Delta lrate$
$\Delta m1p$	7.48**	0.94	2.20	2.95*	0.19	0.59
$\Delta infdef$	0.53	0.65	0.75	0.64	0.69	1.31
$\Delta funds$	s2.33	1.58	10.54**	3.61**	0.65	3.78**
$\Delta infcpi$	3.43**	0.58	1.83	0.31	0.84	0.49
Δgdp	1.55	0.15	5.96**	2.11	0.99	2.01
$\Delta lrate$	4.57**	0.48	0.50	2.20	0.45	3.52**

Notes: Equations are shown in columns. * denotes rejection at the 5% level while ** denotes rejection at the 1% level.

balances equation are all rejected at statistically significant levels. Unlike the previous analysis, the lags of CPI inflation are also significant. There are also differences in the inflation equations where previously the lags on the federal funds was significant in both inflation equations while the lagged long-term interest rate was significant in the CPI inflation equation. Now none of the lags are significant for the GDP deflator inflation equation and no significance is found for the lagged long-term interest rate in the CPI inflation equation. Regardless of these differences, the full system with all of the lags is used to generate the forecasts from the VEqCM.

Table A.3: Estimated Matrix of Equilibrium-Correction Coefficients

Equation	Error correction terms			
	$mdciv_{-1}$	$defrciv_{-1}$	$termciv_{-1}$	$cpirciv_{-1}$
$\Delta m1p$	**-0.034 (-3.44)**	-0.002 (-0.92)	0.000 (-0.51)	-0.001 (-0.90)
$\Delta infdef$	**-3.318 (-2.30)**	**-0.561 (-2.27)**	0.116 (1.12)	0.270 (1.15)
$\Delta funds$	0.861 (1.11)	0.208 (1.56)	**-0.144 (-2.58)**	-0.040 (-0.32)
$\Delta infcpi$	-0.755 (-0.47)	**0.997 (3.63)**	0.145 (1.26)	**-0.951 (-3.63)**
Δgdp	-0.021 (-1.77)	0.001 (0.63)	-0.002 (-1.80)	-0.003 (-1.77)
$\Delta lrate$	0.502 (0.91)	**0.225 (2.38)**	0.028 (0.71)	-0.075 (-0.84)

Notes: t-ratios are in parentheses. Bolded values are significant at 95% confidence level.

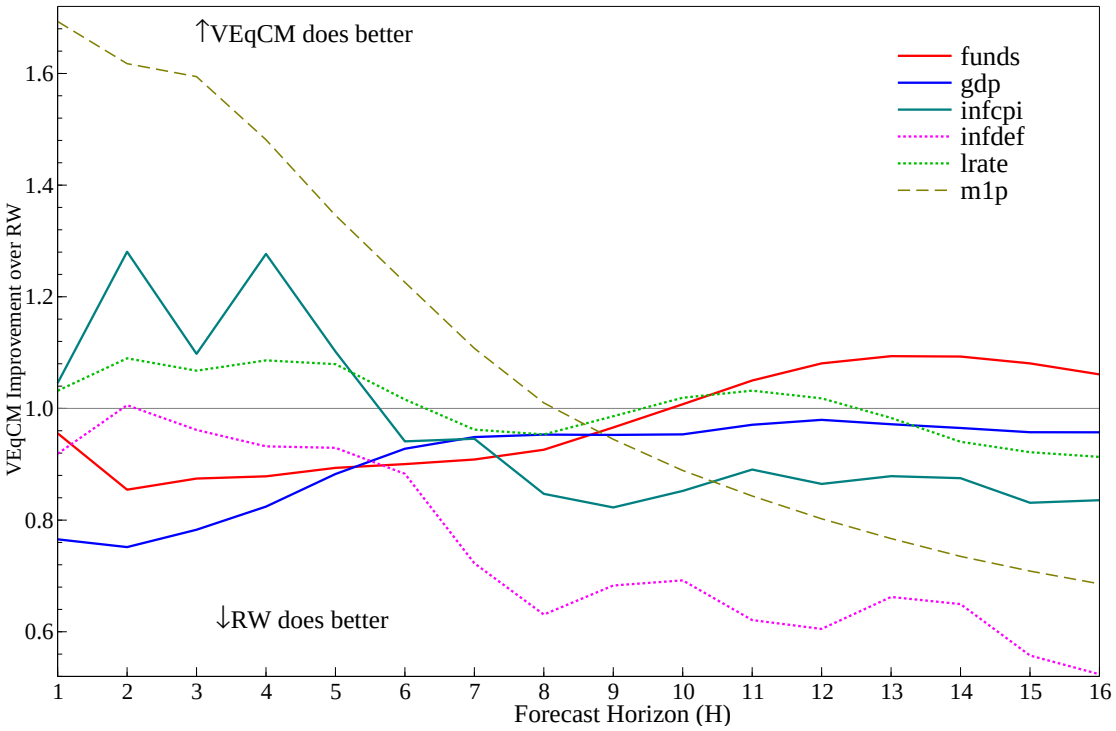
Differences in the significance of the lags in the system is not the only change in the current analysis. Anderson et al. (2002) finds that none of the variables are "weakly exogenous", in that for each of the six variables in the system, at least one of the estimated equilibrium correction coefficients (α) is significantly different from zero. That is, the hypothesized long-run relationship has a significant direct impact on each of the variables. Table A.3 shows that this no longer holds. None of the estimated equilibrium correction coefficients are significantly different from zero for the GDP equation, although three of them are close. Furthermore, while the other equilibrium correction co-

efficients are similar to the original results, their significance differs substantially. Changes in real balances only respond significantly to the long-run money demand equation. Changes in inflation, as measured by the GDP deflator, respond significantly to the long-run money demand equation and one of the Fisher relationships (defrciv). Changes in the federal funds rate only respond significantly to the term-structure spread (termciv), while changes in inflation, as measured by CPI, respond significantly to both of the Fisher relationships. Finally, changes in the long-term interest rate only respond significantly to one of the Fisher equations. Despite these differences, Anderson et al. (2002, p. 579) do not impose further restrictions since economic theory is unable to guide the restrictions and the empirical relationships can change over time.

A.3 An Initial Assessment of Forecast Performance

Once model estimation is finalized, it can be used to generate forecasts. The model is estimated through 1988 Q4 to produce the initial round of forecasts. It is then re-estimated each quarter through 1997 Q3 to produce forecasts through 1997 Q4. In total this procedure generates 36 one-quarter-ahead and 21 16-quarter-ahead forecast error observations. The VEqCM forecasts are then compared against a RW model for each variable. Specifically, a pure RW model is used for the interest rate variables and a RW with drift is used for all other variables.

Figure A.2 plots the relative root mean square forecast errors for the VEqCM and the RW for each variables in the system. In line with Anderson et al. (2002, Tables 7 and 8), the results suggest that there is no strong case for the VEqCM forecasts on a variable by variable basis. While the VEqCM does perform relatively well for some variables at shorter horizons, at longer horizons the RW forecasts appear to dominate for most variables, with the exception of the federal funds rate. However, the RMSE rankings do not capture the error-covariances across variables and horizons which are exploited in large forecasting systems. Thus there is a case for using the GFESM to analyze these forecasts.



Notes: Relative RMSE is the ratio of the random walk RMSE over the VEqCM RMSE for each variable at each horizon.

Figure A.2: Relative RMSEs by Variable and Horizon

Chapter 3

Testing for Differences in Path Forecast Accuracy

The trajectory and path of future outcomes play crucial roles in policy decisions. However, analyses of forecast accuracy typically focus on forecasts at individual horizons. Examining the path across multiple horizons provides additional insight into the forecast dynamics. I propose a test for differences in path forecast accuracy using the link between path forecast evaluation metrics and the joint predictive density. This path test is linked with and extends existing joint multi-horizon accuracy tests. Simulations highlight that there are trade-offs when using path forecast accuracy tests to detect a broad range of differences in forecasts. I use the tests to compare the Federal Reserve's Greenbook path forecasts against four model forecasts, and to evaluate the forecast accuracy of Hurricane Irma's path in 2017. The results show that differences in forecast dynamics play an important role in the assessment of path forecast accuracy.

"Any quibbles I might have with the Greenbook over point estimates for growth are minor because the baseline forecast path in the Greenbook is consistent with our own Atlanta model forecast." Jack Guynn (President of the Federal Reserve Bank of Atlanta) FOMC Meeting Transcript, December 9, 2003.

1 Introduction

Forecast paths play important roles in policy decisions. For example, central banks care about the future trajectory of inflation, output, and unemployment given that monetary policy is required to react to the expected future path of the economy due to lags in policy implementation and the monetary transmission mechanism. The Federal Reserve incorporates future economic paths into its decision making process by using optimal control monetary policy rules (see Yellen, 2012, and Brayton et al., 2014), while several central banks, including Sveriges Riksbank and Norges Bank, publish their expected future paths of policy rates.[1]

Path forecasts are not just useful for monetary policy. Fiscal policy decisions are judged in part by their expected impact on the trajectory of the debt and deficit; see Martinez (2015). Similarly, in natural disasters individuals make contingency plans and choose to evacuate based on the expected future path and intensity of the event; see Chapter 4. Thus, the importance and usefulness of having good path forecasts is relevant for a variety of disciplines.

This chapter constructs a test for differences in path forecast accuracy. The forecast path contains both individual point forecasts at each horizon as well as their relationship with the point forecasts at all other horizons. Testing for differences in the path allows us to capture the expected dynamics of individual variables across horizons as well as the dynamic relationships between variables in a system. Despite their prominence, tests of point forecasts at individual horizons do not capture these important aspects of the future trajectory. This has implications for key questions in the literature. For example, Faust and Wright (2013) and Schorfheide and Song (2015) find that improvements

[1] E.g. see Riksbank Monetary Policy Report April 2018 and Norges Bank Monetary Policy Report March 2018.

in nowcasts result in large improvements in longer-horizon forecasts. This implies that there is a high degree of persistence in the forecasts. However, to answer this question directly, we need to understand how changes in the nowcasts persist across the forecast path.

We start by demonstrating the link between general forecast error metrics and joint densities. Building on this insight, we draw on methods from both the point and density forecast accuracy literature to construct a general test of differences in path forecast accuracy. We show that this general path forecast accuracy test is related to existing joint multi-horizon tests while also being able to capture differences in the forecast dynamics across models. We also derive the asymptotic distribution of this test for the special case where both forecasts are correctly specified.

Monte Carlo simulations illustrate the trade-offs associated with this test. Path forecast accuracy tests are less able to detect differences in biases. The is due to the fact that bias tends to be highly correlated across horizons and so is offset by higher forecast-error covariances. However, path forecast accuracy tests have a much greater ability to capture differences in variances, covariances, and dynamics across forecast models. This is particularly true for differences in forecast dynamics across horizons, which other tests are unable to capture. Thus, our path forecast accuracy test captures a new aspect of differences in forecast accuracy. Furthermore, we also demonstrate how path tests can be extended for use in high dimensional settings.

Finally, we use these methods in two different applications. In the first we compare the Federal Reserve Board's Greenbook point and path forecasts with forecasts from four Dynamic Stochastic General Equilibrium (DSGE) models. The Greenbook's point and path forecasts of inflation dominate the other forecasts, but the DSGE models' ability to track the long-run mean performs well for the GDP growth forecasts over the sample period. We decompose the differences in the path forecasts to show that they are primarily driven by the differences in forecast dynamics across horizons and that these differences are often common across different DSGE models. In the second application we evaluate the path forecast accuracy of models for Hurricane Irma in 2017. We narrow down the more than 100 methods into a handful of the best performing models. The path tests show that while the

official forecasts tend to dominate all methods, other models can match its accuracy depending on the specific weights assigned to the path.

We contribute to the literature on the analyses of path forecasts. Path forecast evaluation using uncertainty bands was first discussed by Jordà and Marcellino (2010), which spurred a growing literature on simultaneous confidence bands; see Staszewska-Bystrova (2011), Jordà et al. (2013), Wolf and Wunderli (2015), Montiel Olea and Plagborg Møller (2017) and Knüppel (2018) among others. Our innovation is to focus on the path forecast's accuracy rather than the uncertainty around the path. Moreover, we test for differences in accuracy between alternative path forecasts. This allows us to better understand differences in the forecast dynamics between alternative models.

We also contribute to the literature on forecast accuracy metrics. Assessments of path forecast accuracy cannot rely on traditional point accuracy metrics. Instead, we show that the general forecast-error second-moment (hereafter GFESM), proposed by Clements and Hendry (1993), can be reinterpreted as a metric of path forecast accuracy. This metric is effectively a multi-horizon generalization of the log determinant measure popularized by Doan et al. (1984). The log determinant is widely used to evaluate forecasts of vector autoregressive (VAR) and DSGE models.[2] While the log determinant captures covariances between variables, the GFESM also captures covariances across horizons and between variables across horizons.

Finally, we build on and extend the literature of tests for path forecast accuracy. Despite an extensive literature on testing for differences in point forecast accuracy (see Clark and McCracken 2013b), there are no tests for differences in path forecast accuracy. We extend the framework of Giacomini and White (2006) and Amisano and Giacomini (2007) to tests of path forecast accuracy. Our test is related to other joint multi-horizon tests by Capistrán (2006) and Quaedvlieg (2019). However, our test captures differences in the forecast-error dynamics, which existing approaches cannot do. Our test is also related to the multi-horizon forecast encompassing test by Hungnes (2018) in that better path forecast accuracy is necessary but not sufficient for one path forecast to forecast-

[2]For example, see Adolfson et al. (2007); Del Negro et al. (2007); Schorfheide and Song (2015); Berg (2016).

encompass another path forecast; see Ericsson (1992).

The rest of the chapter is structured as follows. The next section introduces measures of path forecast accuracy. Section 3 formulates a general path forecast accuracy test and compares it against existing approaches. Section 4 examines the properties of a special case of the path forecast accuracy test. Section 5 compares the test against alternative loss functions in a variety of simulations. Section 6 evaluates the Federal Reserve Board's Greenbook path forecasts against the path forecasts from four DSGE models. Section 7 evaluates of the path forecast accuracy of alternative models for Hurricane Irma in 2017. Section 8 concludes.

2 Measuring Path Forecast Accuracy

We begin by illustrating how the general matrix of the forecast-error second-moment and its determinant, the GFESM, can be interpreted as measures of path forecast accuracy. We also show how these measures relate to the mean square forecast error (MSE), the trace statistic, and the log determinant.

Let $\mathbf{y}_t$ be a K-dimensional random vector. We can denote point forecasts of this vector as $\widehat{\mathbf{y}}_t^m(h) = \widehat{\mathbb{E}}_t\left(\mathbf{y}_{t+h} \mid \mathbf{y}_t, \mathbf{y}_{t-1}, \ldots\right)$ where $\widehat{\mathbb{E}}_t\left(\cdot|\cdot\right)$ is the empirical conditional expectations operator, the sample size used to estimate the parameters required to generate $\widehat{\mathbf{y}}_t^m(h)$ is m and $\{t : m \leq t \leq T\}$. Then the $(K \times 1)$ vector of point forecast errors at each horizon is

$$\widetilde{\mathbf{u}}_t^m(h) = \mathbf{y}_{t+h} - \widehat{\mathbf{y}}_t^m(h). \tag{2.1}$$

Point forecasts can be evaluated using any number of loss functions, $L(\mathbf{y}_{t+h}, \widehat{\mathbf{y}}_t^m(h))$. However, the most common is the quadratic MSE loss function. In multivariate systems, this becomes

$$\widehat{\mathbf{\Sigma}}_{N,h}^m = \frac{1}{N}\sum_{t=m}^{T} \widetilde{\mathbf{u}}_t^m(h)\widetilde{\mathbf{u}}_t^m(h)', \tag{2.2}$$

where $N = T - m + 1$ is the sample of forecast-error observations. A common way to summarize

the information within the MSE matrix is to calculate the trace of $\widehat{\mathbf{\Sigma}}^m_{N,h}$, denoted by $Tr\left(\widehat{\mathbf{\Sigma}}^m_{N,h}\right) = \frac{1}{N}\sum_{t=m}^{T}\widetilde{\mathbf{u}}^m_t(h)'\widetilde{\mathbf{u}}^m_t(h)$, which ignores any covariances between errors across variables. Alternatively, Doan et al. (1984) compute the (log) determinant: $\left|\widehat{\mathbf{\Sigma}}^m_{h,N}\right|$ which captures the covariances between variables. See Komunjer and Owyang (2012) for asymmetric multivariate loss functions.

If $\widehat{\mathbf{Y}}^m_t(H)$ and $\mathbf{Y}_{t,H}$ denote the 1 to H-step predicted and observed paths, the path errors are

$$\widetilde{\mathbf{U}}^m_{t,H} = \mathbf{Y}_{t,H} - \widehat{\mathbf{Y}}^m_t(H) = \begin{bmatrix} \mathbf{y}_{t+1} \\ \vdots \\ \mathbf{y}_{t+H} \end{bmatrix} - \begin{bmatrix} \widehat{\mathbf{y}}^m_t(1) \\ \vdots \\ \widehat{\mathbf{y}}^m_t(H) \end{bmatrix}. \tag{2.3}$$

We can appeal to the general loss function proposed by Clements and Hendry (1993) to evaluate the path forecast errors. Their general matrix of the forecast-error second-moment (GMFESM) extends the MSE matrix in (2.2) to multiple horizons as the average of the squared path forecast errors

$$\widehat{\mathbf{\Phi}}^m_{N,H} = \frac{1}{N}\sum_{t=m}^{T}\widetilde{\mathbf{U}}^m_{t,H}\widetilde{\mathbf{U}}^{m\prime}_{t,H}, \tag{2.4}$$

where each K-dimensional block along the main diagonal of (2.4) represents $\widehat{\mathbf{\Sigma}}^m_{N,h}$ for $h = 1,...,H$, the off diagonals are co-movements between horizons and variables, and the determinant $\left|\widehat{\mathbf{\Phi}}^m_{N,H}\right|$, also known as the GFESM, summarizes this information across variables and horizons. Clements and Hendry (1993, 1995, 1997, 1998) demonstrate that the GFESM captures changes in forecast dynamics, while traditional MSE metrics do not. See Christoffersen and Diebold (1998) for an alternative metric that also captures cointegrating relationships and forecast dynamics.

2.1 The Joint Density as a General Loss Function

Granger (1999) notes that it is possible to use predictive distributions to consider all variations of general forecast error loss functions. However, the standard marginal predictive distribution is unable to handle loss functions with multiple horizons (or transformations thereof). Instead, Granger (1999,

p. 171) argues that it is necessary to consider the joint predictive distribution to handle multi-horizon loss functions. In this context, it is possible to show that under specific assumptions we can formulate a joint predictive distribution that links directly with the GFESM. This illustrates that assumptions about the joint distribution imply a specific loss function and vice versa.

Following Jordà et al. (2013), if we assume that the path forecast errors are iid and have an elliptically contoured distribution, then we can write the joint density as:

$$f_{t,U}\left(\widetilde{\mathbf{U}}_{t,H}^{m}\right) = C\left|\mathbf{\Phi}_{H}^{m}\right|^{-1/2} exp\left\{-g_t\left(\widetilde{\mathbf{U}}_{t,H}^{m\prime}\left\{\mathbf{\Phi}_{H}^{m}\right\}^{-1}\widetilde{\mathbf{U}}_{t,H}^{m}\right)\right\}, \tag{2.5}$$

where C is a normalizing constant, $g(\cdot)$ is a measurable density function, and $\mathbf{\Phi}_{H}^{m}$ is positive definite. Although the assumption of an elliptical distribution is relatively strong, it encompasses a broad class of relevant distributions including the multivariate normal and t distributions. The elliptical distribution also imposes that the loss function is symmetric but allows for a general correlation structure across variables and horizons while accommodating fat-tails.

Given this distributional assumption, we can construct a predictive likelihood following Clements and Hendry (1998).[3] We construct a quasi-profile predictive likelihood by taking the forecasts as being maximized conditional on current and past information and then concentrating the likelihood further, taking logs and ignoring the constant so that

$$\ell_p\left(\widehat{\mathbf{\Phi}}_{N,H}^{m}|\widehat{\mathbf{Y}}_{t}^{m}(H)\right) = -\frac{N}{2}ln\left(\left|\widehat{\mathbf{\Phi}}_{N,H}^{m}\right|\right) - \sum_{t=m}^{T} g_t\left(\widetilde{\mathbf{U}}_{t,H}^{m\prime}\left\{\widehat{\mathbf{\Phi}}_{N,H}^{m}\right\}^{-1}\widetilde{\mathbf{U}}_{t,H}^{m}\right), \tag{2.6}$$

which depends in part on the scaled log GFESM. The other term drops out for the multivariate normal density when plugging in (2.4) as a consistent estimator for $\mathbf{\Phi}_{H}^{m}$. Thus, assumptions about the joint distribution of the path forecast errors can be reinterpreted as choices about the joint loss function.

Interpreting path loss functions more generally using the joint density allows us to appeal to

[3] Bjørnstad (1990) notes that the predictive likelihood was simultaneously proposed by Hinkley (1979) and Mathiasen (1979) to extend the likelihood framework to forecasting.

scoring rules to evaluate path forecast accuracy. Scoring rules are frequently used to evaluate density forecasts. A common class of scoring rules approximate the normal density and only depend on the first and second moments; see Gneiting and Raftery (2007). They are also related to a wider class of loss functions; see Cichocki et al. (2015). The next section constructs a general test of differences in path forecast accuracy using the joint density as a loss function.

3 A General Path Forecast Accuracy Test

This section shows how to construct a path forecast accuracy test when appealing to the joint forecast-error density as a general loss function. We start by describing the notation and the environment. Next we describe the test and the necessary assumptions. Finally, we interpret this test in the context of specific distributional assumptions and against alternative joint multi-horizon tests.

Consider a stochastic process $\mathbf{Z} \equiv \left\{\mathbf{z}_t : \Omega \to \mathbb{R}^S,\ S \in \mathbb{N}.\ t = 1,2,...\right\}$ defined on a complete probability space (Ω, F, P). Partition the observed vector $\mathbf{z}_t$ as $\mathbf{z}_t \equiv \left(\mathbf{y}_t, \mathbf{x}_t'\right)'$, where $\mathbf{y}_t : \Omega \to \mathbb{R}^K$ is the vector of variables of interest and $\mathbf{x}_t : \Omega \to \mathbb{R}^{S-K}$ is a vector of predictors. Define the information set at time t as $\mathbf{F}_t = \sigma\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \mathbf{\Pi}_{m,t}\right)$ where $\mathbf{\Pi}$ is the matrix of population parameters for the z_t process and suppose that methods are used to produce a system of path forecasts for the stacked HK vector of variables of interest, $\mathbf{Y}_{t,H}$, using the information in $\mathbf{F}_t$. Denote the path forecasts for two different sets of methods $i \in \{1,2\}$ by $\widehat{\mathbf{Y}}_{t,i}^{m}(H) \equiv l_i\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \hat{\mathbf{\Pi}}_{i,m,t}\right)$ where $l_i(\cdot)$ represents a measurable function that is allowed to vary across variables and horizons. Subscripts indicate that the time t forecasts are measurable functions of the m most recent observations where m is finite.[4]

The $S(t-m+1) \times HK$ matrix $\hat{\mathbf{\Pi}}_{i,m,t}$ collects the parameter estimates for the set of methods in i. Note that the only requirement imposed on how the forecasts are produced is that they are measurable functions estimated over the same finite estimation window. This allows the forecasts to be produced using a wide range of methods. It also allows for different 'models' to be used for each variable and

[4]Note that as in Giacomini and White (2006), m can vary across forecasting systems. In that case we can redefine m as the maximum of the recent observations used across the two forecasting systems.

/ or horizon as long as they are estimated over the same window and only rely on information in $\mathbf{F}_t$.

Out-of-sample forecast evaluation is performed using a "rolling window" estimation scheme. Let $T+H$ be the total sample size. The forecasts are produced at time m using data indexed $1,\ldots,m$ and the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{m,H,i} = \mathbf{Y}_{m,H} - \widehat{\mathbf{Y}}^m_{m,i}(H) \,\forall\, i \in \{1,2\}$. The estimation window is rolled forward one observation and the second set of forecasts are obtained using observations $2,\ldots,m+1$ where the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{m+1,H,i} = \mathbf{Y}_{m+1,H} - \widehat{\mathbf{Y}}^m_{m+1,i}(H)$. The procedure is thus iterated, so that the last set of forecasts are obtained using observations $T-m,...,T$, so that the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{T,H,i} = \mathbf{Y}_{T,H} - \widehat{\mathbf{Y}}^m_{T,i}(H)$. This yields a sequence of $N = T-m+1$ path forecast-error observations.

This setup builds on Giacomini and White (2006) and Amisano and Giacomini (2007). The main innovation is that we are interested in a vector of path forecasts across multiple variables rather than forecasts for a single variable and horizon at a time. While the framework used here closely follows the unconditional predictive ability test from Giacomini and White (2006), using the joint predictive density as a general loss function also links our approach with the likelihood ratio test for density forecasts in Amisano and Giacomini (2007).[5] In this sense our framework is also related to Vuong (1989), Mitchell and Hall (2005), and Bao et al. (2007).

Another innovation in our analysis is the use of the joint density as a general loss function. We restrict our attention to the log score rule $S(f,U) = ln\{f(U)\}$ which coincides with the log profile predictive likelihood in (2.6). Unlike in Amisano and Giacomini (2007) where the density, $f(\cdot)$, is chosen by the forecaster, here the density is specified as a joint loss function and is the same across forecasters.[6] While for density forecasts, the forecaster chooses the density (often based on the data), here it is chosen by the forecast evaluator and is not necessarily directly related to the underlying data generating process but is instead associated with an acceptable loss profile. As a result, once the

[5]The unconditional predictive ability test is similar to Diebold and Mariano (1995) and West (1996). However, as discussed in Clark and McCracken (2013a), the former focuses on finite sample results while the latter focus on the population, which has implications for the null hypothesis.

[6]Note that there are no technical constraints to keep the forecast evaluator from choosing a different joint density for different forecasts if they so choose.

forecasts are filtered through the joint density then it is also possible to appeal to other methods such as those proposed by Giacomini and Rossi (2010) and Hansen et al. (2011).

For two alternative sets of forecasting methods let

$$LR_{m,t,H} = \Delta ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H}\right)\right\} = ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)\right\} - ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)\right\}. \tag{3.1}$$

We can extend (3.1) to allow for different weights at each horizon. To see this, note that the joint density is equal to the product of the conditional and marginal densities. This conditioning has a unique temporal ordering such that the joint density for each method i can be decomposed as

$$f_t\left(\widetilde{\mathbf{U}}^m_{t,H,i}\right) = \prod_{h=1}^{H} f_t\left(\tilde{\mathbf{u}}^m_{t,i}(h) \mid \tilde{\mathbf{u}}^m_{t,i}(0),\ldots,\tilde{\mathbf{u}}^m_{t,i}(h-1)\right), \tag{3.2}$$

where $\tilde{\mathbf{u}}^m_{t,i}(0) = \mathbf{0}$ and the forecast error density at each horizon is conditional on all previous horizons. Given this decomposition, and allowing for different weights at each horizon, then (3.1) becomes

$$WLR_{m,t,H} = \sum_{h=1}^{H} w_h \Delta ln\left\{f_t\left(\tilde{\mathbf{u}}^m_t(h) \mid \tilde{\mathbf{u}}^m_t(1),\ldots,\tilde{\mathbf{u}}^m_t(h-1)\right)\right\}, \tag{3.3}$$

where w_h is the weight assigned to each horizon. In practice, this could allow longer horizons to receive less weight than shorter horizons while accounting for the dependence between them. Note that re-weighting the conditional and marginal error densities differs from Amisano and Giacomini (2007) who re-weight the density forecast to focus on different aspects of the distribution.

A test for equal performance of the weighted path forecasts can be formulated as

$$H_0 : \mathbb{E}\left[WLR_{m,t,H}\right] = 0, \;\; t = 1,2,.... \text{ against} \tag{3.4}$$

$$H_A : \mathbb{E}\left[\overline{WLR}_{m,N,H}\right] \neq 0 \quad \text{for all } N \text{ sufficiently large,} \tag{3.5}$$

where $\overline{WLR}_{m,N,H} = \frac{1}{N}\sum_{t=m}^{T} WLR_{m,t,H}$. Then the test is based on the following statistic

$$t_{m,N,H} = \frac{\sqrt{N} * \overline{WLR}_{m,N,H}}{\hat{\sigma}_N}, \tag{3.6}$$

where $\hat{\sigma}_N^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of asymptotic variance $\sigma^2 = var\left[\sqrt{N} * \overline{WLR}_{m,N,H}\right]$; see Andrews (1991) and Lazarus et al. (2017).

A level α test rejects the null hypothesis of equal performance of the path forecasts of systems 1 and 2 whenever $abs\left(t_{m,N,H}\right) > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution. The following theorem provides asymptotic justification:

Theorem 1. Given a finite estimation window $m < \infty$, with fixed weights $0 \leq w_h < \infty$, suppose

A1. $\{z_t\}$ is a mixing sequence with ϕ of size $-r/(2r-2)$, $r \geq 2$, or α of size $-r/(r-2)$, $r > 2$;

A2. $\mathbb{E}\left[ln\left\{f_t\left(\widetilde{\mathbf{U}}_{t,H,i}^{m}\right)\right\}\right]^{2r} < \infty$ for all t and $i \in \{1,2\}$;

A3. $\sigma^2 = var\left[\sqrt{N} * \overline{WLR}_{m,N,H}\right] > 0$ for all N sufficiently large.

Then (a) under H_0 in (3.4), $t_{m,N,H} \xrightarrow{D} N(0,1)$ as $N \to \infty$ and (b) under H_A in (3.5), for some constant $c \in \mathbb{R}$, $P\left[abs\left(t_{m,N,H}\right) > c\right] \to 1$ as $N \to \infty$.

The proof of Theorem 1 follows directly from Giacomini and White (2006, Proof of Theorem 4) and Amisano and Giacomini (2007) where $\mathbf{W}_t \equiv \mathbf{z}_t$ and $\Delta L_{m,t} \equiv WLR_{m,t,H}$. Choosing $f_t(\cdot)$ in (3.1)-(3.3) from the class of elliptical densities illustrates the link with the GFESM. Any number of multivariate densities could be chosen depending on the choice of a loss function. However, elliptical densities are symmetric and closely related to the standard MSE loss function. Given this and assuming that $w_h = 1$ for all $h \in H$, then (3.3) becomes

$$WLR_{m,t,H} = -\frac{1}{2}\left[\Delta ln\left(\left|\widehat{\mathbf{\Phi}}_{N,H}^{m}\right|\right) + \Delta g_t\left(\widetilde{\mathbf{U}}_{t,H}^{m\prime}\left\{\widehat{\mathbf{\Phi}}_{N,H}^{m}\right\}^{-1}\widetilde{\mathbf{U}}_{t,H}^{m}\right)\right] \tag{3.7}$$

so that (3.6) can be reformulated as

$$t_{m,N,H} = -\frac{\sqrt{N}}{2\hat{\sigma}_N}\left[\Delta ln\left(\left|\widehat{\mathbf{\Phi}}_{N,H}^{m}\right|\right) + \Delta o_{m,N,H}\right], \tag{3.8}$$

where $\Delta o_{m,N,H} = \frac{2}{N}\sum_{t=m}^{T-1}\Delta g_t\left(\widetilde{\mathbf{U}}_{t,H}^{m\prime}\left\{\widehat{\mathbf{\Phi}}_{N,H}^{m}\right\}^{-1}\widetilde{\mathbf{U}}_{t,H}^{m}\right)$. Alternatively, choosing the multivariate normal density so that $\Delta o_{m,N,H} = 0$ and allowing for alternative weights, (3.6) becomes

$$t_{m,N,H} = -\frac{\sqrt{N}}{2\hat{\sigma}_N}\sum_{h=1}^{H} w_h \Delta ln\left(\left|\widehat{\mathbf{\Sigma}}_{N,h|(0,\ldots,h-1)}^{m}\right|\right), \tag{3.9}$$

where $ln\left|\widehat{\mathbf{\Sigma}}_{N,h|(0,\ldots,h-1)}^{m}\right|$ is the log determinant of the MSE matrix at each horizon, see (2.2), which is orthogonalized by all prior horizons.[7] So (3.9) is a weighted test for differences in the conditional log determinants of the MSE matrix at each horizon. Note that when $w_h = 1\ \forall\ h \in H$, (3.9) is a test for differences in the log GFESM since $ln\left(\left|\widehat{\mathbf{\Phi}}_{N,H}^{m}\right|\right) = \sum_{h=1}^{H} ln\left(\left|\widehat{\mathbf{\Sigma}}_{N,h|(0,\ldots,h-1)}^{m}\right|\right)$.

3.1 Comparisons with Other Joint Multi-Horizon / Multivariate Approaches

The difference between the proposed test statistic and alternative joint multi-horizon approaches by Capistrán (2006) and Quaedvlieg (2019) is seen most clearly when using the multivariate normal density as a loss function. We can expand (3.9) as:

$$\begin{aligned} t_{m,N,H} = &-\frac{\sqrt{N}}{2\hat{\sigma}_N}\left[\sum_{h=1}^{H} w_h \Delta\left\{ln\left(\left|\widehat{\mathbf{\Sigma}}_{N,h}^{m}\right|\right) - ln\left(\frac{\left|\widehat{\mathbf{\Sigma}}_{N,h}^{m}\right|}{\left|\widehat{\mathbf{\Sigma}}_{N,h|(0,\ldots,h-1)}^{m}\right|}\right)\right\}\right. \\ &+\frac{1}{N}\sum_{t=m}^{T}\Delta\left\{\widetilde{\mathbf{U}}_{t,H}^{m\prime}W_{H,K}^{\frac{1}{2}\prime}\left(\left\{\widehat{\mathbf{\Phi}}_{N,H}^{m}\right\}^{-1} - \mathbf{I}_{HK}\right)W_{H,K}^{\frac{1}{2}}\widetilde{\mathbf{U}}_{t,H}^{m} + C\left(W_{H,K}\right)\right\} \\ &\left.+\sum_{h=1}^{H} w_h \sum_{k=1}^{K}\Delta\left\{\frac{1}{N}\sum_{t=m}^{T}\tilde{u}_t^m(h,k)^2\right\}\right], \end{aligned} \tag{3.10}$$

where $\tilde{u}_t^m(h,k)$ is the kth element in $\tilde{\mathbf{u}}_t^m(h)$ as defined in (2.1), $W_{H,K} = diag(\{w_1,\ldots,w_H\})\otimes\mathbf{I}_K$, and where $C\left(W_{H,K}\right)$ captures interactions between the weights and covariance terms. Note that $C\left(W_{H,K}\right) = 0$ when $w_h = \gamma\ \forall h \in H$ for some constant γ or when $diag\left(\widehat{\mathbf{\Phi}}_{N,H}^{m}\right) = \widehat{\mathbf{\Phi}}_{N,H}^{m}$. The first two lines of (3.10) capture the weighted differences in the forecast-error covariances and dynamics (i.e. differences in the Gaussian copulas). The third line captures the weighted differences in the

[7] For example, the conditional MSE matrix at the second horizon is $\widehat{\mathbf{\Sigma}}_{2|1} = \widehat{\mathbf{\Sigma}}_2 - \widehat{\mathbf{\Sigma}}_{2.1}\widehat{\mathbf{\Sigma}}_1^{-1}\widehat{\mathbf{\Sigma}}_{1.2}$, where $\widehat{\mathbf{\Sigma}}_{1.2} = \widehat{\mathbf{\Sigma}}_{2.1}^{\prime}$ is the co-movement between the forecast errors at the first and second horizons. Also note that $\widehat{\mathbf{\Sigma}}_{1|0} = \widehat{\mathbf{\Sigma}}_1$.

marginal densities across variables and horizons; i.e. differences in the mean squared errors.

Equation (3.10) illustrates how the path test relates to other joint multi-horizon/variate tests. The trace test statistic proposed by Capistrán (2006) or the average superior predictive ability (aSPA) test discussed in Quaedvlieg (2019) is captured in the last line of (3.10). On the other hand, Doan et al. (1984)'s log determinant metric is captured by the first term in the first line of (3.10). As a result, the path test incorporates information from these approaches while also including additional information that captures differences in how the forecasts incorporate dependence across variables and horizons. This is important, especially when large differences in the forecast-error covariances can offset or exacerbate the differences between individual variables at longer horizons.

3.2 Some Comments on Nested Path Forecasts

There is an extensive literature examining the link between tests of nested and non-nested forecasts. In particular, Clark and McCracken (2001) demonstrate the link between the Diebold and Mariano (1995) and West (1996) frameworks for nested models. While the approach we take above allows for both nested and non-nested models in sample, it relies on the assumption that they are not nested in the population so that $\sigma^2 = var\left[\sqrt{N} * \overline{WLR}_{m,N,H}\right] > 0$; see Giacomini and White (2006, p. 1546).

If this does not hold, i.e. when $\sigma^2 = 0$, then the true forecast-error densities are identical and the test statistic will converge to a non-standard distribution; see Clark and McCracken (2013a). However, Hansen and Timmermann (2015) show that when $HK = 1$ this test statistic is related to the difference between Wald test statistics from the forecast sample (N) and estimation sample (m). Thus, following Engle (1984) and Stewart (1995), it is also related to the difference between likelihood ratio test statistics from the forecast and estimation samples. This suggests that it may be possible to formulate a test, following Vuong (1989), of σ^2 in order to determine whether a nested or non-nested asymptotic distribution should be used. However, this task is beyond the scope of the current analysis and so is left to future research.

There are several of other out-of-sample nested forecasting tests in the literature. For example,

Gelper and Croux (2007) propose a likelihood ratio test for testing nested 1-step-ahead forecasts in multivariate systems using a bootstrapping procedure. This could also be extended to path forecasts, but requires direct knowledge of and the ability to estimate the underlying model as well as a bootstrap procedure that can account for a large amount of autocorrelation. More recently, Hungnes (2018) extends the forecast-encompassing tests originally proposed in Chong and Hendry (1986) to path forecasts. This is a very general framework since forecast-encompassing tests are effectively out-of-sample tests of nested forecast models. However, it is not obvious how this relates back to the rest of the nested forecast model literature. Thus, there is considerable work that needs to be done to fully explore the link between nested and non-nested models in the context of path forecast evaluation.

4 A Special Case of the Path Forecast Accuracy Test

While the previous section provides a general path forecast accuracy test, this section derives more explicit results for the forecast path test under stronger assumptions. This provides further insight into the properties of path forecast accuracy tests and allows us to move away from HAC estimates of the variance, which have poor finite sample performance; see Müller (2014). We start by laying out the main result from which everything else follows with the following theorem:

Theorem 2. (Asymptotic distribution) Suppose that

A1. the forecast errors follow a $MA(p)$ process: $\widetilde{\mathbf{u}}_t(h) = \sum_{i=0}^{p} \mathbf{\Pi}^i \mathbf{v}_{t+h-i}$ for $p \leq h-1$,

A2. where $\mathbf{v}_t \overset{\text{iid}}{\sim} N_K[\mathbf{0}, \mathbf{\Omega}]$ and $\mathbf{\Omega}$ is positive definite,

A3. and that $H > 0$ and $K < N$ where both H and K are fixed.[8]

Under these assumptions, the estimated log GFESM satisfies

$$\frac{ln\left|\widehat{\mathbf{\Phi}}_{N,H}\right| - ln\left|\mathbf{\Phi}_H\right| - \frac{HK(K+1)}{2N}}{H\sqrt{2K/N}} \xrightarrow{D} N(0,1) \; as \; N \to \infty. \tag{4.1}$$

This extends Cai et al. (2015)'s central limit theorem for log determinants to $MA(p)$ processes. The sketch of the proof starts with **A1**, where if the point forecast errors are $MA(h-1)$, then it is

[8] This assumption can be relaxed by allowing K to vary with N. See Cai et al. (2015).

possible to construct a lower triangular parameter matrix for the path forecast errors

$$\widetilde{\mathbf{U}}_{t,H} = \mathbf{\Psi}_{\Pi,H}\mathbf{V}_{t,H} = \begin{pmatrix} \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Pi} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{\Pi}^{H-1} & \cdots & \mathbf{\Pi} & \mathbf{I}_K \end{pmatrix} \begin{pmatrix} \mathbf{v}_{t+1} \\ \mathbf{v}_{t+2} \\ \vdots \\ \mathbf{v}_{t+H} \end{pmatrix}, \tag{4.2}$$

which drops out when taking the determinant as long as $p \leq h-1$. Given this, the problem can also be thought of as the determinant of the covariance of an asymmetric Hankel matrix where the number of variables (K) governs the degree of asymmetry. From **A1** and the iid assumption in **A2** then

$$ln\left|\widehat{\mathbf{\Phi}}_{N,H}\right| = ln\left|\frac{1}{N}\sum_{t=1}^{N}\mathbf{V}_{t,H}\mathbf{V}_{t,H}^{'}\right| \approx ln\begin{vmatrix} \widehat{\mathbf{\Omega}}_{N+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\mathbf{\Omega}}_{N+H} \end{vmatrix} = \sum_{h=1}^{H} ln\left|\widehat{\mathbf{\Omega}}_{N+h}\right|, \tag{4.3}$$

where we can treat each $\left|\widehat{\mathbf{\Omega}}_{N+h}\right|$ independently. The rest of the proof is in Technical Appendix A.1 and follows directly from Cai et al. (2015) under the assumptions in **A2** and **A3**. We can also interpret this as a general central limit theorem for the log determinant of the covariance of independent ($H = 1;\ K > 1$) and Hankel ($H > 1;\ K = 1$) matrices. As far as we know, this is the first study to consider the distribution of the covariance of asymmetric Hankel matrices; see Ghodsi et al. (2015).

We now formulate tests for differences in the GFESM, as described by the following corollaries:

Corollary 1 (GFESM Test: zero mean). If the assumptions of Theorem 2 are satisfied by two path forecasting systems and the stacked forecast error shocks are uncorrelated for all variables and horizons across forecasting systems so that $\mathbb{E}\left(V_{H,1}V_{H,2}'\right) = \mathbf{0}$, where each $V_{H,i}$ is the $HK \times N$ matrix of stacked forecast error shocks for the respective system, then a test can be constructed which, under the null hypothesis of equal path forecast performance, satisfies

$$-\frac{\sqrt{N}}{2H\sqrt{K}}\Delta ln\left(\left|\widehat{\mathbf{\Phi}}_{N,H}\right|\right) \xrightarrow{D} N(0,1)\ as\ N \to \infty. \tag{4.4}$$

This result follows from the properties of a two-sample t-test under the normality and uncorrelated shocks assumptions. It is also closely related to Anderson (2003, Theorem 7.5.4). However, (4.4) is smaller than Clements and Hendry (1993, equation 38) for the simulated GFESM by a factor of $1/\sqrt{H}$. Corollary 1 coincides with the general test in (3.8) when $\sigma^2 = H^2K$. Alternatively, it coincides with (3.9) for any $w_h \geq 0$ as long as $\sum_{h=1}^{H} w_h > 0$ where $\sigma^2 = \left(\sum_{h=1}^{H} w_h\right)^2 K$. Furthermore, if the assumptions for Corollary 1 are satisfied, then it is the efficient test since σ^2 is not estimated.

Corollary 2 (GFESM Test: non-zero mean). If we relax the assumption **A2** of Theorem 2 for both forecasting systems so that $\mathbf{v}_{t,i} \overset{\text{iid}}{\sim} N_K\left[\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i\right]$ for any $\boldsymbol{\mu}_i$ but still require that the stacked forecast error shocks are uncorrelated for all variables and horizons across forecasting systems so that $\mathbb{E}\left(V_{H,1}V_{H,2}'\right) = \mathbf{0}$, then a test can be constructed which, under the null hypothesis of equal path forecast performance, satisfies

$$\sqrt{N}\frac{-\Delta ln\left(\left|\widehat{\boldsymbol{\Phi}}_{N,H}\right|\right)}{\sqrt{2H\left\{\sum_{i=1}^{2} Tr\left[\left(\mathbf{I}_{HK}+2\boldsymbol{\eta}_i\right)\left(\mathbf{I}_{HK}+\boldsymbol{\eta}_i\right)^{-2}\right]\right\}}} \xrightarrow{D} N(0,1) \; as \; N \to \infty, \tag{4.5}$$

where $\boldsymbol{\eta}_i$ is the squared non-centrality across variables and horizons.

See Technical Appendix A.2 for the proof. Note that when $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = \mathbf{0}$ then (4.5) collapses to (4.4). If $\boldsymbol{\eta}_i$ is unknown, it is necessary to use a consistent estimator. The uncorrelated shocks assumption could also be relaxed following the approach by Iliopoulos (2008) or by Jafari (2012).

One implication of this result is that the conditional MSEs are broadly stable across horizons. This is a very restrictive condition. Chong and Hendry (1986, p. 685) argue that for stationary processes, conditional MSEs are "not [...] monotonic in the forecast horizon". This implies that there is either heteroskedasticity across horizons, large outliers, or some other form of misspecification. Therefore, tests based on Corollaries 1 and 2, which average across horizons, will likely understate the uncertainty so as to over reject the null hypothesis.

It is possible to allow for higher order dependence in the forecast errors. For example, if the forecast errors follow a $MA(p)$ process, then for any $p \geq 0$ for $H \geq 2$ the log GFESM is

$$ln\left|\widehat{\mathbf{\Phi}}_{N,H}\right| \approx \sum_{h=2}^{H} ln\left|\widehat{\mathbf{\Omega}}_{N+h}\right| + ln\left|\sum_{j=0}^{\max(p-H+1,0)} \mathbf{\Pi}^{j}\widehat{\mathbf{\Omega}}_{N+1-j}\mathbf{\Pi}'^{j}\right|. \tag{4.6}$$

When p is close to $H-1$ and $|\mathbf{\Pi}| < 1$ this additional dependence has little impact. In general it is necessary to account for the additional forecast-error variance. Focusing on the first term in (4.6) one can condition the path on the first-step-ahead forecast (or nowcast) errors and then compute the test statistic using conditional densities for the second through H-step-ahead forecast errors while adjusting for the now shorter path. The test statistic for the first horizon can then be computed using the general test described above and recombined with the rest of the path. This should produce a test statistic that is valid for any amount of dependence.

It may also be possible to extend this special case to relax the iid and normality assumptions. In order to do so, it is necessary to have an estimator of the GMFESM that is heteroskedasticity consistent (see White, 1980) and robust to outliers or fat-tails (see Rousseeuw, 1985). The approach described in Chapter 2 could provide one potential avenue for doing this. However, it is beyond the scope of this analysis and so is left for future research.

5 Simulations

In this section we use Monte Carlo experiments to compare the properties of the path forecast accuracy tests with existing joint tests. We start by describing the path forecast-error generating process. We then describe alternative tests and present the results. All numerical results were obtained using OxMetrics Version 7.10; see Doornik (2013).

5.1 The Path Forecast-Error Generating Process

The path forecast errors are generated as follows:

$$\widetilde{\mathbf{U}}_{t,H,\{b,v,c_k,c_h\}} \equiv \boldsymbol{\theta}_b + \mathbf{\Psi}_{\Pi,H}\mathbf{\Sigma}^{1/2}_{v,c_k,c_h}\mathbf{V}_{t,H}, \tag{5.1}$$

where $\mathbf{V}_{t,H}$ is an asymmetric Hankel matrix where each unique element is $\mathbf{v}_{t,h} \overset{\text{iid}}{\sim} N_K(\boldsymbol{\mu}, \mathbf{I}_K)$. For simplicity we set $\boldsymbol{\mu} = \mathbf{0}$. The forecast errors have a model-specific bias through $\boldsymbol{\theta}_b$, are serially correlated through $\boldsymbol{\Psi}_H$, and have model-specific variances and correlations across horizons and variables through $\boldsymbol{\Sigma}_{v,c_k,c_h}$. Thus, the forecast errors follow a $MA(h-1)$ process and exhibit dependence and biases across variables and horizons. Next, we discuss the individual parameters.

First, we define the covariance structure across horizons and variables, at a single point in time. Since forecast errors generally converge to the unconditional mean when the horizon is large, then the correlations should get larger for adjacent horizons when h is large, and smaller for shorter horizons. We use the correlation matrix C, with elements ρ_{g,h,j,k,c_k,c_h}:

$$\rho_{g,h,j,k,c_k,c_h} = \begin{cases} 1 & \text{if } g = h,\ j = k \\ \exp(-1.2 + 0.025max(g,h) - 0.125abs(h-g)) + c_h & \text{if } g \neq h,\ j = k \\ \exp(-1.8) + c_k & \text{if } g = h,\ j \neq k \\ \exp\left(-1 - \sqrt{abs((k-j)(h-g))}\right) + \frac{c_k + c_h}{2} & \text{if } g \neq h,\ j \neq k \end{cases} \tag{5.2}$$

where c_k governs the differences in how errors are correlated across variables while c_h governs differences in how errors are correlated across horizons. Higher values for either increases the overall correlation. This plays a prominent role in our analysis where under the null, c_k and c_h are set equal to zero while under the alternative they vary across models. We allow the variance to change across horizons so that $\sigma_{v,h,k} = v\left(1 + \frac{\sqrt{h-1}}{2}\right)$ where we set $v = 1$. The variance and correlation are combined so that $\Sigma_{v,c_k,c_h} = diag(\sigma_v) C_{c_k,c_h} diag(\sigma_v)$.

Next, we define the model-specific bias as

$$\boldsymbol{\theta}_{b,h,k} = b\left(1 + \sqrt{h-1}\right), \tag{5.3}$$

which implies that the bias increases with the horizon but is similar across variables. b governs the

degree of bias across models where in the baseline case $b = 1$. Changes in $\boldsymbol{\theta}_b$ shift all horizons or variables up or down based on a fixed spread, while changes in $\boldsymbol{\mu}$ increase the spread across horizons.

Finally, we define the serial correlation as a matrix of HK elements where

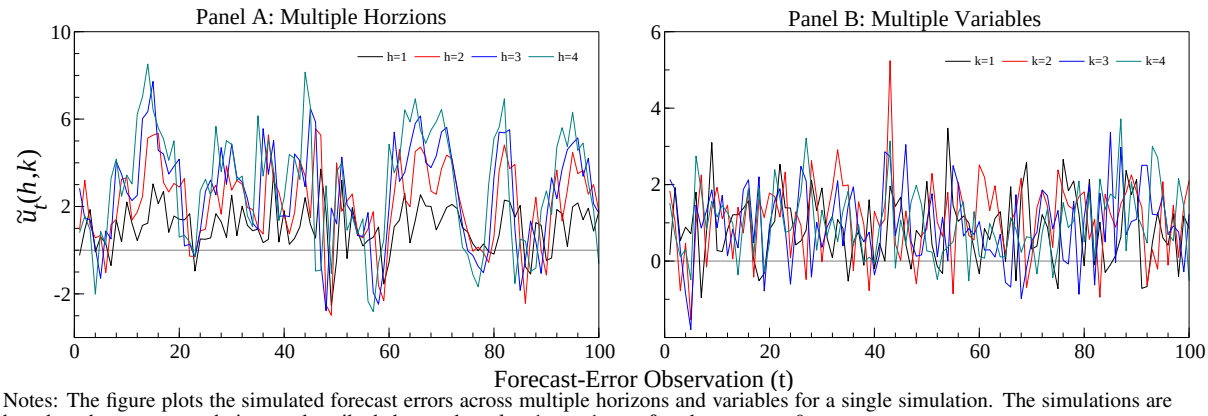
$$\boldsymbol{\Psi}_{\Pi,H} = \begin{pmatrix} \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Pi} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \boldsymbol{\Pi}^{H-1} & \cdots & \boldsymbol{\Pi} & \mathbf{I}_K \end{pmatrix}, \tag{5.4}$$

so that serial correlation accumulates across horizons just as a $MA(h-1)$ process. We use the following $\pi_{j,k}$ elements for each $K \times K$ matrix $\boldsymbol{\Pi}$:

$$\pi_{j,k} = \begin{cases} 0.4 + min\left(\frac{k}{10}, 0.5\right) & j = k \\ 0.2 & j \neq k \end{cases}. \tag{5.5}$$

Note that $\boldsymbol{\Pi}$ only plays a role when $\boldsymbol{\theta}_b \neq \mathbf{0}$ since $\left|\boldsymbol{\Psi}_{\Pi,H}\right| = 1$. We visualize the choice of the DGP by plotting the forecast errors across 100 observations in Figure 5.1. The pattern of the forecast errors across horizons is similar to that of the U.S. Congressional Budget Office's forecast errors for the gross federal debt; see Martinez (2015). It is also possible to use alternative parameterizations so that the simulated forecast errors follow patterns similar to the Federal Reserve's Greenbook forecasts errors for GDP growth and inflation. However, this does not change the overall simulation results.

We compare the performance of four test statistics: (1) the equally weighted univariate aSPA (i.e. the trace across horizons and variables), (2) the equally weighted multivariate aSPA (i.e. the single horizon version of (3.8)), (3) the equally weighted general path test from equation (3.8), and (4) the special case of the path test in (4.5). The first represents existing approaches in the literature as seen in Capistrán (2006) and Quaedvlieg (2019) that do not explicitly capture differences in covariances or dynamics. The second is effectively the log determinant measure from Doan et al. (1984) applied to



Notes: The figure plots the simulated forecast errors across multiple horizons and variables for a single simulation. The simulations are based on the parameter choices as described above where $b = 1$, $v = 1$, $\boldsymbol{\mu} = \mathbf{0}$ and $c_k = c_h = 0$.

Figure 5.1: Illustration of Forecast-Error Data Generating Process

the test described in Quaedvlieg (2019) which accounts for covariances but not dynamics. Standard errors are computed using the HAC estimator in Andrews (1991) for the first three test statistics.[9] The standard errors for the fourth test statistic are computed based on estimates of the bias.

The comparisons allow for an assessment of the strengths and weaknesses of the loss functions. The univariate test does not explicitly capture differences in covariances or dynamics. The multivariate test explicitly captures differences in covariances but not differences in dynamics. The general path test and the special case test capture differences in covariances and dynamics. This allows us to illustrate the advantages and disadvantages of the loss functions under different scenarios.

Performance is evaluated by examining rejection frequencies under the null and various alternatives for a given critical value. Section 5.2 examines the null rejection frequency across alternative sample sizes and dimensions. Section 5.3 considers the non-null rejection frequency across various differences between the forecasts. This allows for the determination of how well different tests or loss functions detect alternative forms of misspecification. Finally, section 5.4 analyzes the null rejection frequency when the number of path forecast-error observations is small relative to the number of variables or the length of the forecast path.

[9]The standard errors could also be computed using bootstrap procedures as in Hansen et al. (2011) or Quaedvlieg (2019). However, simulations indicate that these procedures only perform well when there is an abundance of forecast-error observations (i.e. > 200).

5.2 Null Rejection Frequency

We first consider the case where the forecast errors from the two forecasting systems are essentially the same. In this case, since we are conducting our tests using a nominal size of 5%, we expect that the null rejection frequency for each test statistic should be approximately 5%. We assess how this changes with the number of forecast-error observations as well as the dimension of the system and the length of the forecast horizon. Table 5.1 presents the results.

Table 5.1: Null Rejection Frequencies

	Univariate Test				**Multivariate Test**				**General Path Test**				**Special Path Test**			
N↓ ‖ H,K→	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2
32	3.07	3.23	0.10	0.25	4.21	4.57	0.82	1.13	4.21	4.57	0.59	0.83	5.39	5.80	5.60	5.81
64	4.26	4.54	1.05	1.50	4.89	5.42	3.04	3.36	4.89	5.42	2.43	2.51	5.16	5.93	6.12	5.09
128	4.54	4.51	2.73	3.39	4.54	5.11	4.02	4.57	4.54	5.11	3.82	4.27	4.71	5.54	5.56	5.26
256	4.80	4.94	4.00	4.16	4.80	4.87	4.85	5.00	4.80	4.87	4.54	5.04	4.80	5.31	5.28	5.15
512	4.89	4.94	4.58	4.49	4.89	4.84	4.97	4.97	4.89	4.84	5.09	4.91	5.11	5.07	6.05	4.98
1000	4.72	4.84	4.68	4.81	4.72	4.99	5.29	4.99	4.72	4.99	4.96	5.11	4.79	5.34	5.77	5.34

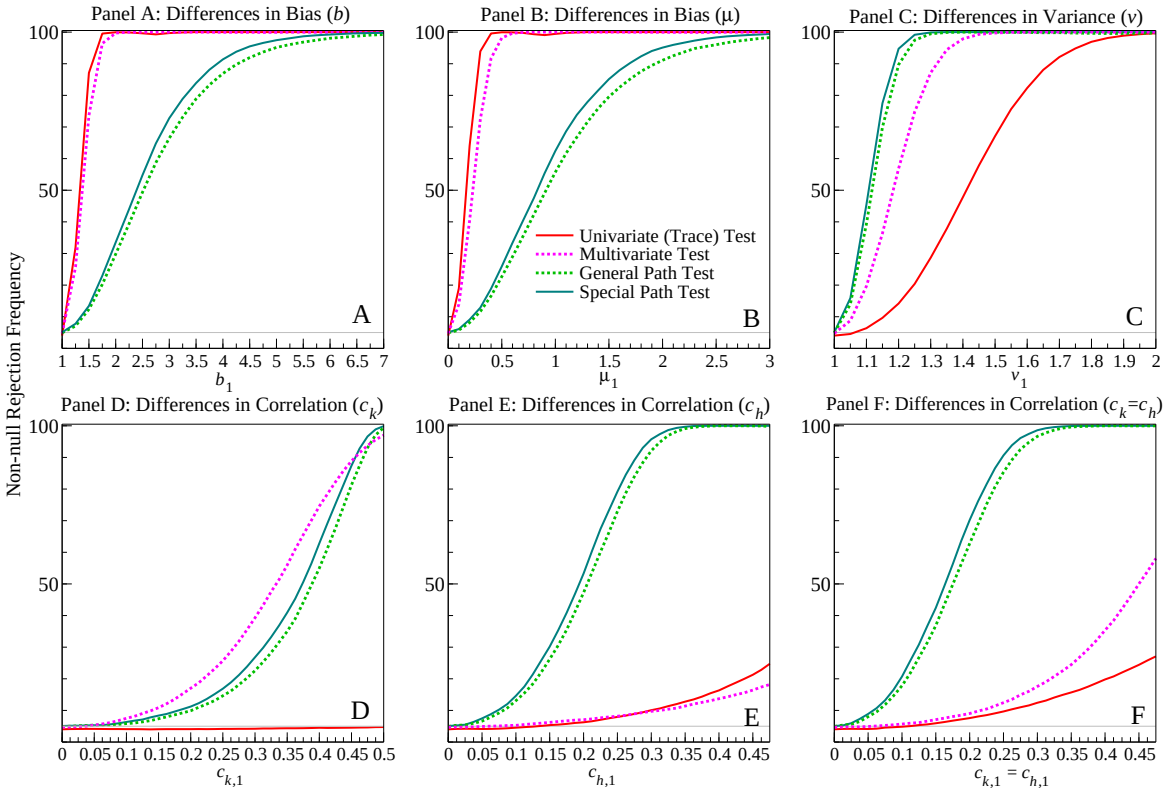
Notes: The nominal size is 5%. The Univariate test is the average trace statistic across all variables and horizons. The Multivariate test averages across horizons. General path test refers to the general likelihood ratio test statistic and the special path test refers to Corollary 2. 10,000 Replications.

Table 5.1 illustrates that while most test statistics start off undersized, when $N = 32$ seriously so, for all but the special path test when $N \leq 64$ and $H, K = 4,2$, they all generally converge to the correct nominal size as the number of forecast-error observations increases. This dynamic is even more pronounced for larger systems and longer horizons. When $H = 1$ the multivariate and general path tests are identical, which stems from the fact that they both rely on the multivariate normal density as their underlying loss function. When $K = H = 1$ the univariate test, the multivariate test, and the general path test are identical for $N > 100$.

The special path test from Corollary 2 performs well. Its null rejection frequency is slightly above 5%, particularly when $H > 4$, which may stem from the fact that the bias in the forecast errors does not follow the assumed structure, i.e. $\boldsymbol{\eta}_i \neq \mathbf{I}_H \otimes \boldsymbol{\mu}_i \boldsymbol{\mu}_i'$. However, despite this, the null rejection frequency remains close to the nominal size, especially when $K > 1$. The advantages of this test are seen in small sample sizes where most other test statistics tend to be severely undersized, whereas the special path test is not. This stems from the fact that the other test statistics rely on HAC estimates of the standard errors which are known to have poor small-sample properties.

5.3 Non-Null Rejection Frequency

Next we assess the ability to capture differences across forecasting systems. While only differences in biases are typically considered, we examine differences in biases, variances, and correlations across variables and horizons. This allows us to assess how well alternative loss functions capture various differences in the forecast errors. For simplicity, we focus on the case where $K = 2$, $H = 4$ and $N = 200$. Figure 5.2 presents the results across model differences.



Notes: The figure plots the rejection frequencies when the null is false for different statistics across different degrees of model differences for fixed $N = 200$ when $K = 2$ and $H = 4$. The nominal size is 5%. General path test refers to the test statistic in (3.8). 10,000 replications.

Figure 5.2: Non-Null Rejection Frequencies by Type of Difference

Figure 5.2 illustrates an important difference between the path forecast accuracy tests and the alternative tests. Panels A and B indicate that the path forecast tests are less able to distinguish differences in forecast bias when compared against other joint accuracy tests. This is likely due to the fact that the bias persists across multiple variables and horizons. This persistence implies an increase in the forecast-error covariance, which can offset increases in forecast biases at individual horizons. Therefore in order to detect differences in bias, path forecast accuracy tests require either that these differences do not persist across horizons or that they are very large.

Path forecast accuracy tests are better at detecting most other forecast-error differences; see panels C, D, E and F of Figure 5.2. Panel C shows that path tests are better at detecting differences in the forecast-error variances. However, the real strength of the path tests stems from their ability to detect differences in dynamics / correlations (panel E) across horizons. Even the multivariate aSPA test has difficulty capturing differences in dynamics. This illustrates that the strength of a multi-horizon loss function, and therefore its associated tests, is the ability to capture differences in forecast-error covariances (i.e. dynamics) across horizons. These advantages are greater when there is a larger degree of forecast error covariance across variables and horizons.

5.4 Extensions to Large H and K

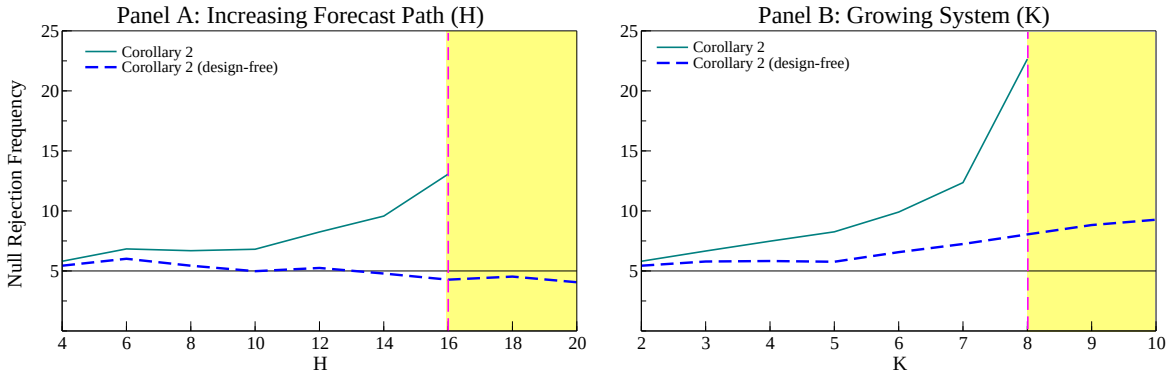
It is possible extend the tests to allow for higher dimensional systems and longer forecast horizons even when there are a relatively small number of forecast-error observations. Estimates of the general matrix of the forecast-error second-moment, i.e. the GMFESM in equation (2.4), become ill-conditioned when the dimension of the system times the length of the forecast path is larger than the number of forecast-error observations ($HK > N$). In practice, this limits the usefulness of these metrics since the reliability of estimates of the GFESM can deteriorate rapidly. This also has important implications for test statistics based on these metrics.

As described in Chapter 2, Hendry and Martinez (2017) propose a procedure that extends these limitations and produces reliable estimates of the GFESM even when $HK > N$. Their procedure provides a model-free extension of Abadir et al. (2014)'s "design-free" methods to non-iid data by using sub-samples of the eigenvectors to improve estimates of the eigenvalues and thereby ensure that the GMFESM remains positive definite. Hendry and Martinez (2017) show that the approach does well in a variety of simulations and produces reliable estimates of the GFESM when $HK \geq N$.

Here we show how the test statistics introduced above perform when using these "design-free" estimates.[10] For simplicity we focus on the properties of the test statistic associated with Corollary

[10]This is closely related to the literature on testing for differences in high dimensional matrices. See Li and Chen (2012), Cai et al. (2013) and Cho and Phillips (2018).

2. These results can also be extended to the general test statistic. However, there are important complications when using the HAC estimator. Figure 5.3 presents the null rejection frequencies for long forecast horizons and large systems.



Notes: The figure plots the null rejection frequency for different methods of computing Corollary 2 where either H or K is expanding for fixed $N = 32$. When H is expanding, K is fixed at $K = 2$ and when K is expanding, H is fixed at $H = 4$. The shaded region indicates where the standard GMFESM is singular. The nominal size is 5%. 10,000 replications.

Figure 5.3: Null Rejection Frequency for Larger Dimensions

Figure 5.3 shows that increasing the forecast horizon or the size of the forecast system relative to the number of forecast-error observations causes the test to be substantially oversized. The GMFESM is ill-conditioned when $H > 16$ in Panel A or $K > 8$ in Panel B and so cannot be computed in these cases. However, the "design-free" methods overcome these limitations by enabling computation of the GFESM even when $HK > N$. Panel A demonstrates the benefits of these methods even when extending the forecast paths out to twenty-steps-ahead, the modified Corollary 2 test remains close to the correct nominal size. The benefits are less pronounced in Panel B for increases in K, however there is still a substantial reduction in the null rejection frequency relative to the original test statistic. Thus, there are large gains when using path forecast accuracy tests in large systems and long horizons.

6 Evaluating FRB Greenbook and DSGE Model Path Forecasts

This section uses path tests to compare the accuracy of the Federal Reserve Board's Greenbook path forecasts with the path forecasts from four DSGE models for GDP growth, inflation and interest rates. While it is natural to focus on the path of the level of GDP, analysis of the path of GDP growth rates and levels are identical when using the GFESM, since it is invariant to this transformation. We

describe the data and forecasts. Next we analyze the point and path forecast accuracy and test for differences in the path forecasts using the general path test from Section 3. Finally, we decompose the path test statistic to understand which aspect of the forecast errors is driving the relative performance.

The data and forecasts that are used in this analysis are real GNP/GDP growth (hereafter referred to as GDP), the GDP deflator (hereafter referred to as inflation) and the federal funds rate (hereafter referred to as interest rate) spanning 1985Q4-2000Q4. The DSGE model forecasts are generated in Wolters (2015).[11] While Wolters (2015) utilizes every forecast release per quarter to evaluate forecast accuracy, we use one forecast release per quarter. This provides a consistent set of path forecasts to evaluate each quarter. Note that earlier in the sample, several quarters do not contain forecast releases. In these cases, the forecast made in the previous quarter which corresponds to the missing quarter's forecast is used instead.[12] The 'actual' values against which we compare the forecasts are those published in the Greenbook two quarters after the quarter to which data refer. Thus, we follow the common practice of using the third (final) data estimates in our analysis.

We compare the Greenbook (GB) forecasts against four DSGE model forecasts. The models come from: Del Negro and Schorfheide (2004, "DS"), Edge et al. (2008, "EDO"), Fuhrer (1997, "FM"), and Smets and Wouters (2007, "SW"). They range in size from small (DS and FM) and medium (SW) to large (EDO) and were chosen because of their academic and policy relevance as well as their wide-spread use before the Great Recession. The forecasts were generated using the real-time dataset from Faust and Wright (2009).

Figure 6.1 plots the actual values and the paths of the forecasts made in every quarter for each model and variable. The rows represent forecasts for different variables, while the columns present the forecasts for different models. They are commonly referred to as 'hedgehog takeoff plots' since the forecasts often point away from the actual series like the quills on a hedgehog. They illustrate that the forecasts consistently fail to capture the correct future path of the series and often converge to a long-run equilibrium that is different from the short-run fluctuations in the data.

[11] Special thanks to Maik Wolters for sharing his data and forecasts.

[12] For example, the previous quarter's 1-step-ahead forecast becomes the current quarter's nowcast and so on.

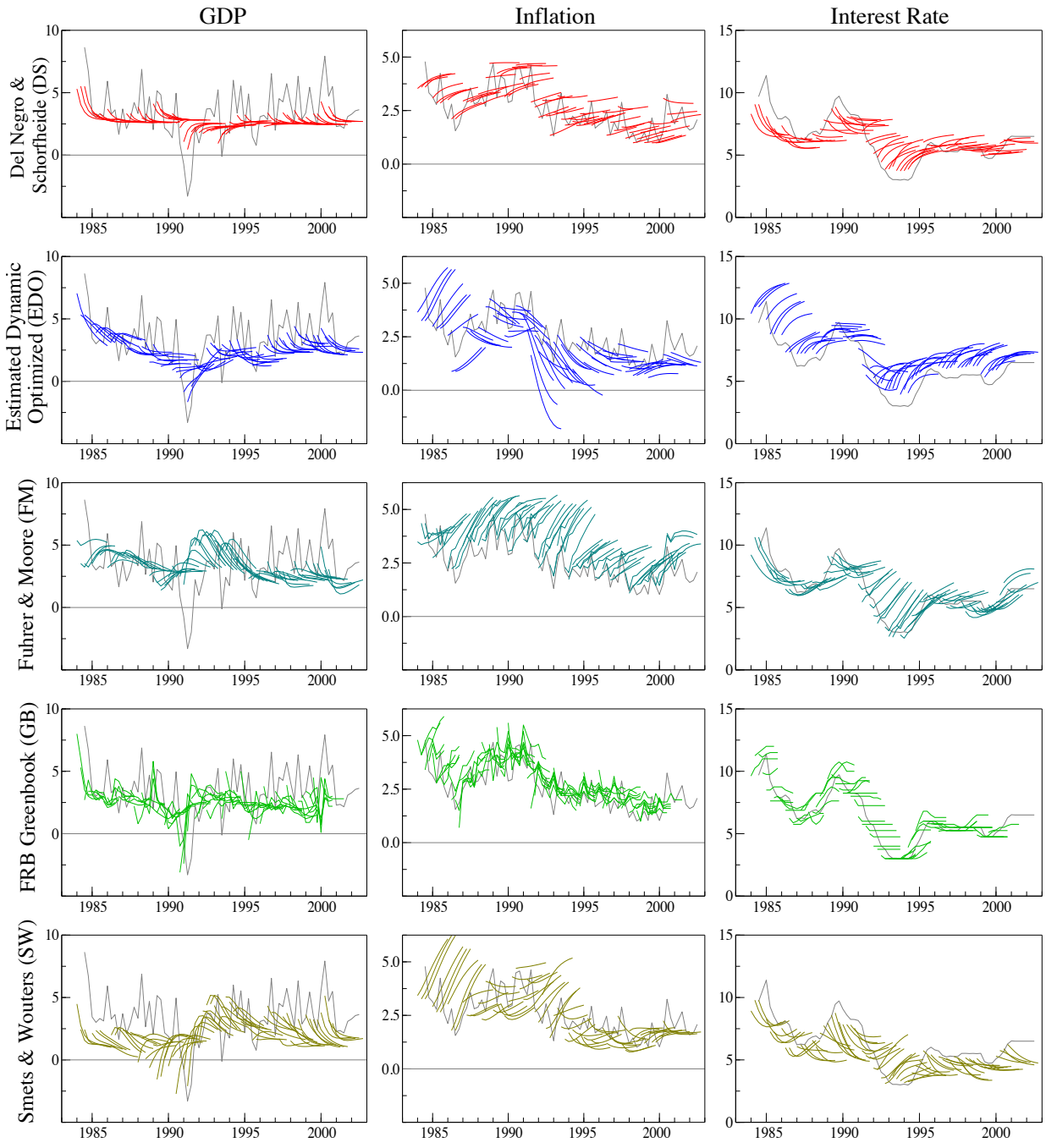


Figure 6.1: 'Path' Forecasts and Actuals, 1984-2002

There are both similarities and differences in the path forecasts across models. Looking first at the forecasts for GDP, the DSGE models generally do not capture the dynamics of the series but instead adhere closely to the long-run mean of the series. This is particularly true for forecasts from the DS model but also to some extent for the others. On the other hand, the Greenbook path forecasts do seem to capture some of the dynamics away from the mean.

These results are more pronounced for the inflation forecasts. The DSGE models tend to have linear path forecast trajectories, especially when compared with the Greenbook. That being said, the properties of the forecasts are very different across alternative models. The DS model's inflation forecasts are generally fairly constant across horizons. On the other hand, the FM model's inflation forecasts generally trend upwards towards a much higher equilibrium, which may be influenced by the high inflation period prior to the start of the forecast sample. The EDO and SW inflation forecasts tend to be somewhat more reactive to the business cycle despite having periods of both substantial over and under prediction.

Forecasts of the interest rate appear to be the most similar across models. This is particularly interesting since the Greenbook 'forecasts' of the interest rate over this period were not forecasts in the traditional sense but instead represent the interest rate path upon which the Greenbook forecasts are conditioned. In fact, the Greenbook appears to have been conditioned on a simple no change rule for at least part of the sample. Despite this, the DS model's forecasts tend to vary the least and adhere closely to the mean of the series, while the EDO and FM forecasts tend to consistently over predict the interest rate which is consistent with the fact that they also over-predict inflation and implement some form of Taylor rule when setting the interest rate. The SW model's forecasts are best at capturing the dynamics despite also under-predicting the interest rate over long periods.

Although the DSGE model forecasts extend out through eight-quarters-ahead, we only evaluate them up through four-quarters-ahead due to data limitations. We also construct a simple average of each of the model forecasts (AV1). We exclude the nowcasts from our tests to allow for more consistent comparisons (although this does not change the results). Thus, four forecast horizons are

Table 6.1: Nowcast Error Properties Across Forecast Models

	Nowcast Bias						**Standard Deviation**					
	GB	DS	EDO	FM	SW	AV1	GB	DS	EDO	FM	SW	AV1
GDP	0.72	0.44	0.70	-0.25	1.06	0.48	1.46	2.06	1.89	2.18	2.19	1.86
INF	-0.21	-0.02	0.40	-0.32	0.21	0.07	0.63	0.79	0.76	0.88	0.76	0.72
INT	0.01	-0.03	-0.66	0.08	0.34	-0.07	0.09	0.61	0.47	0.40	0.46	0.41

Notes: See Figure 6.1 and text for model descriptions. INF refers to inflation. INT refers to the interest rate.

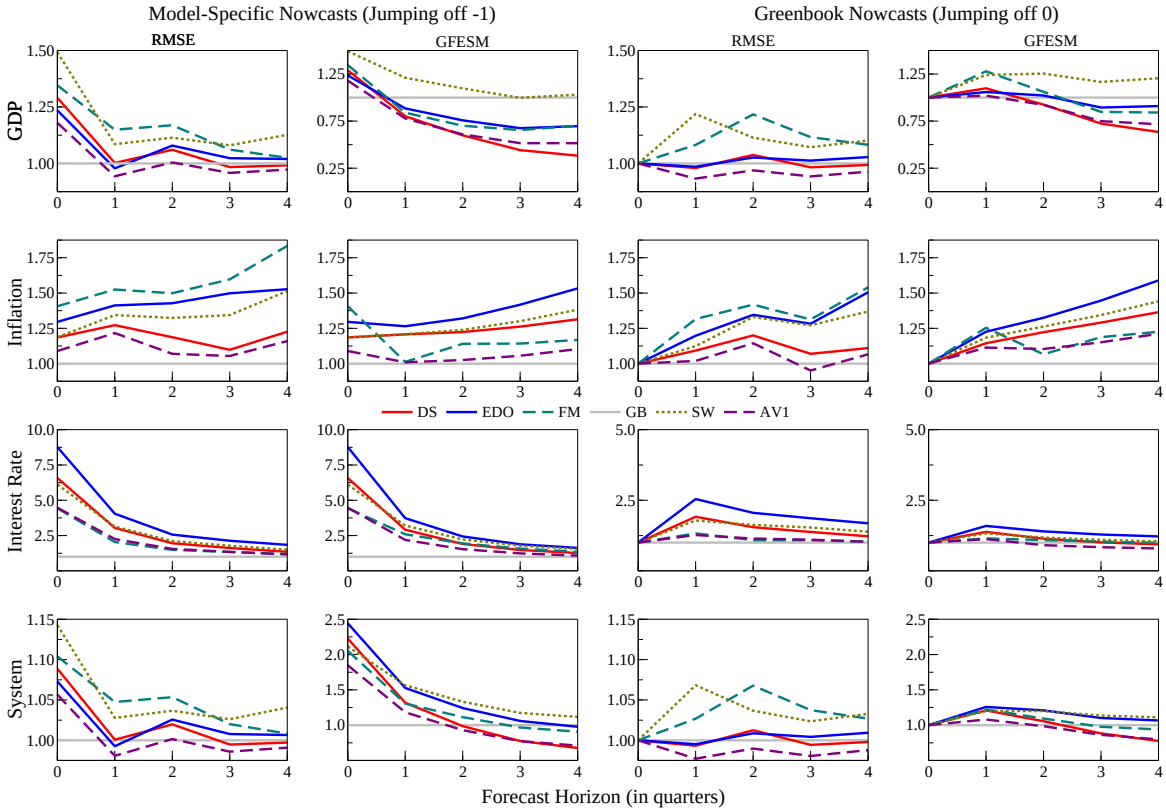
evaluated along with a maximum of three variables at a time so that $HK = 12$. Since there are approximately 70 forecast-error observations, there is no concern of too few forecast-error observations.

Assessing Point and Path Forecast Accuracy

We start by evaluating the point forecast accuracy using the root mean squared forecast errors (RMSE). We also compute the system's point forecast accuracy using the root standardized determinant of the MSE matrix, which is similar to the root geometric mean of the sequentially conditioned RMSEs. Path forecast accuracy is evaluated using the root standardized GFESM. Together they provide a comprehensive view of point and path forecast accuracy.

The DSGE forecasts face an important disadvantage relative to the Greenbook. They are constructed so that they only use data from the previous quarter, which ignores the fact that nowcasts are often augmented with higher frequency data. As a result, Wolters (2015) generates two different DSGE forecasts. The first (i.e. Jumping off -1) uses data from the previous quarter to generate the nowcast and forecasts. The second (i.e. Jumping off 0) uses the Greenbook nowcasts as an additional observation of the sample from which the forecasts are generated.

Augmenting the DSGE forecasts with the Greenbook nowcasts should improve the DSGE forecasts. Recent work by Faust and Wright (2013) and Schorfheide and Song (2015) shows that improvements in the nowcasts can result in large improvements in the forecasts. However, it may not be the case here since the DSGE 'nowcasts' and the Greenbook nowcasts are used differently and have very different properties. The DSGE nowcasts are essentially 1-quarter-ahead forecasts while

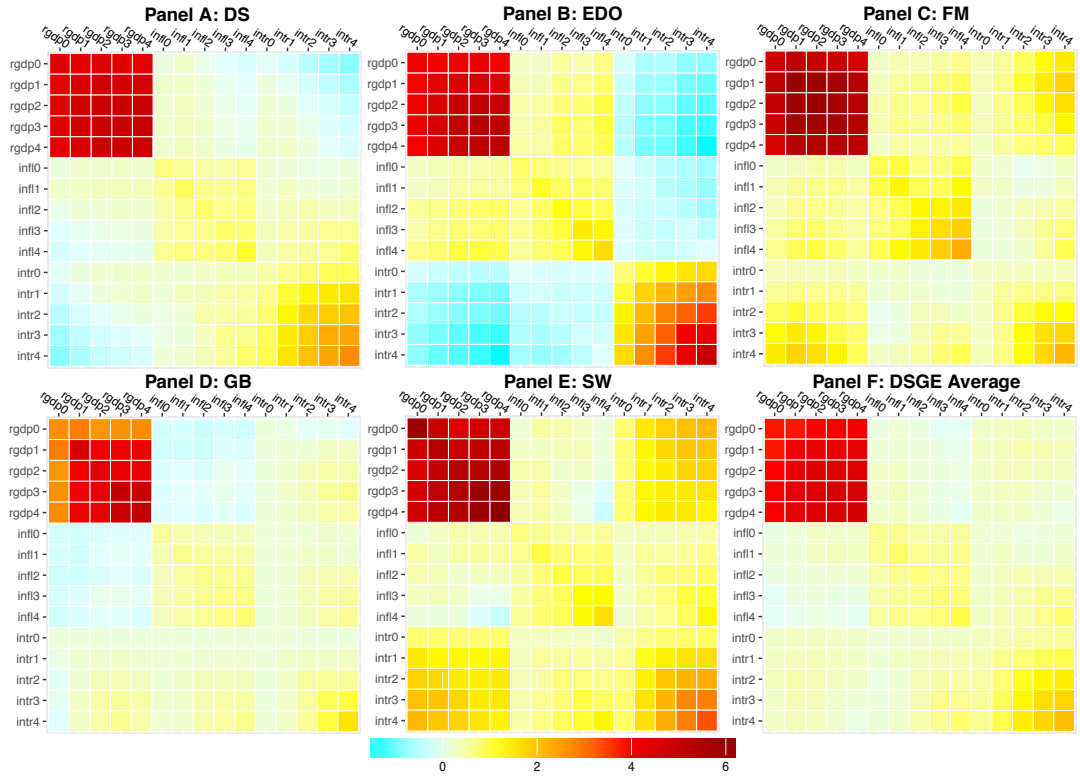


Notes: The Greenbook is identical for both jumping off points. The GFESM is computed as the root of the standardized determinant where the determinant is raised to the power of $1/2HK$. System includes all three variables. See Figure 6.1 and text for model definitions.

Figure 6.2: Relative Point and Path Forecast Accuracy

the Greenbook nowcasts are treated as data without accounting for any measurement errors. Table 6.1 presents summary statistics for the nowcast errors by model and variable. The Greenbook nowcast errors exhibit a higher bias than the DSGE models, especially for GDP. In fact, the Greenbook nowcasts only have a lower bias for the interest rate. The Greenbook makes up for the higher bias with much lower standard deviations. However, this suggests that using Greenbook nowcasts as data will not necessarily improve the DSGE model forecasts at longer horizons.

Relative forecast accuracy results are presented in Figure 6.2. It shows that the Greenbook forecasts of inflation and interest rates have lower RMSEs at all horizons while the DS model and AV1 forecasts of GDP have lower RMSEs than the Greenbook. This implies that forecasts of GDP growth, which capture the long-run mean of the series are likely to do well. However, the sample coincides with the Great Moderation, so it is unclear if this extends to more volatile periods; see Del Negro and Schorfheide (2013), Hendry and Mizon (2014) and Hendry and Muellbauer (2018).



Notes: rgdp = GDP, infl = inflation, intr = interest rate and the numbers denote the forecast horizon from 0 to 4 quarters ahead. The matrices are symmetric by construction where the mean square forecast errors for each variable and horizon fall along the main diagonal of each matrix. See Figure 6.1 and text for model definitions.

Figure 6.3: Forecast-Error Second-Moment Matrices for Different Forecast Models

To evaluate the path forecast accuracy we consider the joint density. This is summarized using all of the information in the forecast-error second moment matrix rather than just the diagonal as is common with Trace MSE. Figure 6.3 presents a graphical depiction of the forecast-error second-moment matrices for each model as well as for the data autocovariance and for forecasts based on the in-sample mean. It illustrates the rich covariance structure in the forecast-errors. It also indicates that there are large differences in this structure across each of the models.

We summarize the information in the forecast-error second-moment matrices using the GFESM. Figure 6.2 shows that based on this measure, the Greenbook path forecasts of inflation do best even when accounting for differences in the nowcasts. The DSGE model forecast average performs better for interest rates, which is not surprising since its RMSEs are not much different from the Greenbook's. Although the DSGE model forecast average has low RMSEs for GDP, the DS model has the lowest GFESM. This suggests that although the model average has a lower bias than the DS model

at each horizon, it comes at a cost of not being offset by the forecast-error dynamics. This is seen in Figure 6.3 where the covariances across horizons and variables are clearly smaller for the DSGE model average than for the DS model. Thus in the context of the path forecasts, the benefits of reducing the forecast bias with averaging, which is common for point forecasts, depends on whether it is outweighed by reductions in forecast-error dynamics which offset the bias.

The path forecast accuracy rankings indicate that there are few improvements at longer horizons when using the Greenbook nowcasts. In fact, for GDP and inflation the DSGE model relative path forecast accuracy declines at the longest horizons when using the Greenbook nowcasts versus their own. The interest rate forecasts are an exception in that there are improvements in path forecast accuracy at longer horizons. These findings go against the point forecast accuracy results where using the Greenbook nowcasts lowers the RMSEs across all horizons.

Differences in the Path Forecasts

It is not clear if the differences in path forecast accuracy are statistically significant or what might be driving them. We can answer these questions by testing for differences in path forecast accuracy using the general path test from section 3.[13] We take the Greenbook as the baseline against which we compare all other forecasts. Next, we decompose this test-statistic into its different components in order to better understand what aspect of the path forecast is driving the differences between the forecasts and whether there is any information beyond what is available in point forecast metrics.

Table 6.2 presents the results. We focus specifically on the statistical significance of the test. Larger differences between the metrics do not necessarily always translate into greater statistical significance since the standard errors may be more or less impacted by different correlations between the forecasts. Therefore, it is important to focus on the test statistic.

The test statistics and their associated p-values in Table 6.2 provide fairly consistent results. Overall, the Greenbook path forecasts of GDP perform significantly worse than the DSGE forecasts with

[13] We also considered the special path test from section 4, however the conditional log determinants are not very stable across the forecast horizons indicating that the required assumptions are not satisfied.

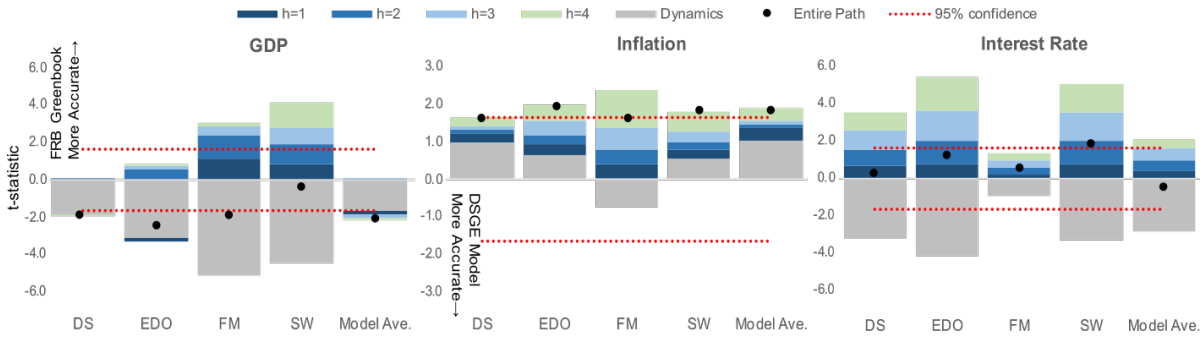
Table 6.2: Testing for Differences in Path Forecast Accuracy (1985 Q4 - 2000 Q4)

		Model-Specific Nowcasts						Greenbook Nowcasts					
		GB	DS	EDO	FM	SW	AV1	GB	DS	EDO	FM	SW	AV1
GDP	GFESM	0.96	**0.33**	0.65	0.70	0.93	0.47	0.96	**0.52**	0.85	0.71	1.17	0.62
	Path test		1.75*	2.38**	2.02*	0.29	2.11*		2.15*	1.41	2.23*	1.66*	2.40**
			[0.040]	[0.009]	[0.022]	[0.388]	[0.017]		[0.016]	[0.080]	[0.013]	[0.048]	[0.008]
INF	GFESM	**0.47**	0.64	0.77	0.65	0.69	0.55	**0.47**	0.67	0.79	0.59	0.70	0.57
	Path test		1.64	2.00*	1.60	1.80*	1.89*		1.87*	1.94*	1.87*	1.92*	1.97*
			[0.051]	[0.023]	[0.055]	[0.036]	[0.029]		[0.031]	[0.026]	[0.031]	[0.027]	[0.024]
INT	GFESM	0.51	0.54	0.70	0.56	0.66	**0.47**	0.51	0.47	0.66	0.51	0.54	**0.38**
	Path test		0.34	1.29	0.466	1.74*	0.67		0.56	1.17	0.08	0.52	1.44
			[0.366]	[0.099]	[0.321]	[0.041]	[0.253]		[0.287]	[0.121]	[0.468]	[0.302]	[0.075]
SYS	GFESM	0.57	**0.35**	0.52	0.50	0.58	0.37	0.57	**0.41**	0.61	0.52	0.63	0.42
	Path test		1.67*	1.37	1.12	0.24	1.81*		1.79*	1.17	1.19	1.54	1.87*
			[0.048]	[0.086]	[0.132]	[0.407]	[0.035]		[0.036]	[0.122]	[0.118]	[0.062]	[0.030]

Notes: Only horizons 1-4 are considered. INF refers to inflation. INT refers to the interest rate. SYS refers to the system of all three variables. See text and Figure 6.1 for model definitions. AV1 is an average of the DSGE model forecasts. The GFESM is standardized by $1/2KH$. Bold values indicate the smallest GFESM for each variable and nowcast. Path test refers to the general path test statistic. The tail probability associated with the null hypothesis of equal path forecast accuracy is in square brackets: *p< 0.1 **p< 0.05 ***p< 0.01.

the exception of the EDO and SW models (especially after controlling for starting point differences). On the other hand, the Greenbook path forecasts of inflation perform significantly better than the DSGE forecasts and their averages. The differences in path forecasts of the interest rate are generally not significant, especially after controlling for differences in the nowcasts.

This presents a more nuanced and comprehensive view of differences in forecast accuracy between the Greenbook and DSGE models. The Greenbook inflation forecasts dominate all other forecasts considered here. This is true for both the point and path forecasts. However, the DSGE models' ability to track the long-run mean does well for the GDP forecasts over this period. Furthermore, while the Greenbook has a solid advantage in point forecasts of interest rates (despite them not being actual forecasts), the DSGE forecasts demonstrate some value in the path forecasts even after accounting for differences in the nowcasts. Comparing across DSGE model forecasts, we find that both point and path forecast accuracy metrics (and tests) indicate that smaller DSGE models do better (particularly DS) than medium and large-scale DSGE models (i.e. SW and EDO). While this is an interesting finding, it is beyond the scope of this analysis to speculate as to why this may be the case.



Notes: See text and Figure 6.1 for model definitions. The decomposition of the general path test follows from (3.10).

Figure 6.4: Decomposition of Differences in Path Accuracy (Relative to FRB Greenbook)

The differences in the results between the point and path forecast metrics indicate that forecast dynamics play an important role in determining path forecast accuracy. We illustrate this by decomposing the general path test statistic as in (3.10) to get the relative contribution of each horizon and as well as dynamics across horizons to the overall path test statistic.

Figure 6.4 presents this decomposition for each of the DSGE model forecasts across all three variables. It indicates that the dynamics (or common error components across horizons) play a large role across all models and variables. The bars show that differences in forecast dynamics can exacerbate differences between the models (e.g. inflation) or ameliorate or even flip the differences (e.g. GDP or interest rates). For example, the FM model performs worst across all horizons in terms of the GDP point forecasts. However, when accounting for differences in forecast dynamics, then FM's path forecast is significantly more accurate than the Greenbook path forecast.

7 Evaluating Hurricane Irma's Path Forecast Accuracy

In this section we illustrate the general applicability of path forecast accuracy tests to subject areas beyond the field of economics. In particular, we use the tests to evaluate alternative forecasts of the path of Hurricane Irma in 2017. Irma initially formed from a tropical wave near Cape Verde in late August and strengthened to become one of the most intense hurricanes in terms of maximum sustained wind speeds ever observed in the Atlantic Ocean. Irma made its way across the Atlantic

in early September causing an estimated 65 billion dollars in damages in several Caribbean islands, Cuba, and along the west coast of Florida before moving up the gulf coast of Florida and into the panhandle where it ultimately dissipated further inland.

More than 100 different models produced short-term forecasts of Hurricane Irma's path. Forecasts from these models were generated every six hours for the life of the storm with forecast horizons ranging from as short as 6-hours-ahead and extending all the way out to longer than 168-hours-ahead (or 7-days-ahead). The National Hurricane Center maintains a database of all of the publicly available forecasts along with the 'best track' observations of the actual path of the storm. Although this database contains 116 forecast models or methods for Hurricane Irma, we can narrow this down to 91 by eliminating models that are identical or are missing a large number of observations.[14] We compute the forecast errors at every horizon for each model using methods for calculating the distance on the surface of a spheroid by Vincenty (1975). Following the National Hurricane Center, we limit our analysis to seven horizons: 12, 24, 36, 48, 72, 96 and 120-hours-ahead. This gives 37 forecast error observations for each model across seven horizons.

We further narrow down the number of models in our analysis by estimating a Model Confidence Set (MCS) for all 91 models following Hansen et al. (2011). The MCS seeks to find the set of models that have the best joint performance for a given loss function at a given horizon and for a chosen probability. While typical analyses rely on the individual MSE loss function, we also include the joint multivariate normal loss function. We estimate the MCS for each loss function and at each horizon up through 120-hours-ahead. While it is common to estimate an MCS that includes models that have a relatively low probability, in this exercise we focus on models that are included in the MCS with a high probability so as to limit the total number of models included in our analysis.

Table 7.1 shows that 35 of the 91 models fall within the 50% model confidence set for at least one horizon and measure of loss, where a higher p-value implies a higher probability of falling within the MCS. The same set of models also falls within the 20% MCS for at least one horizon when only

[14]For all other models, missing forecast observations are imputed linearly.

Table 7.1: Hurricane Irma Track Forecasts: Model Confidence Set Results

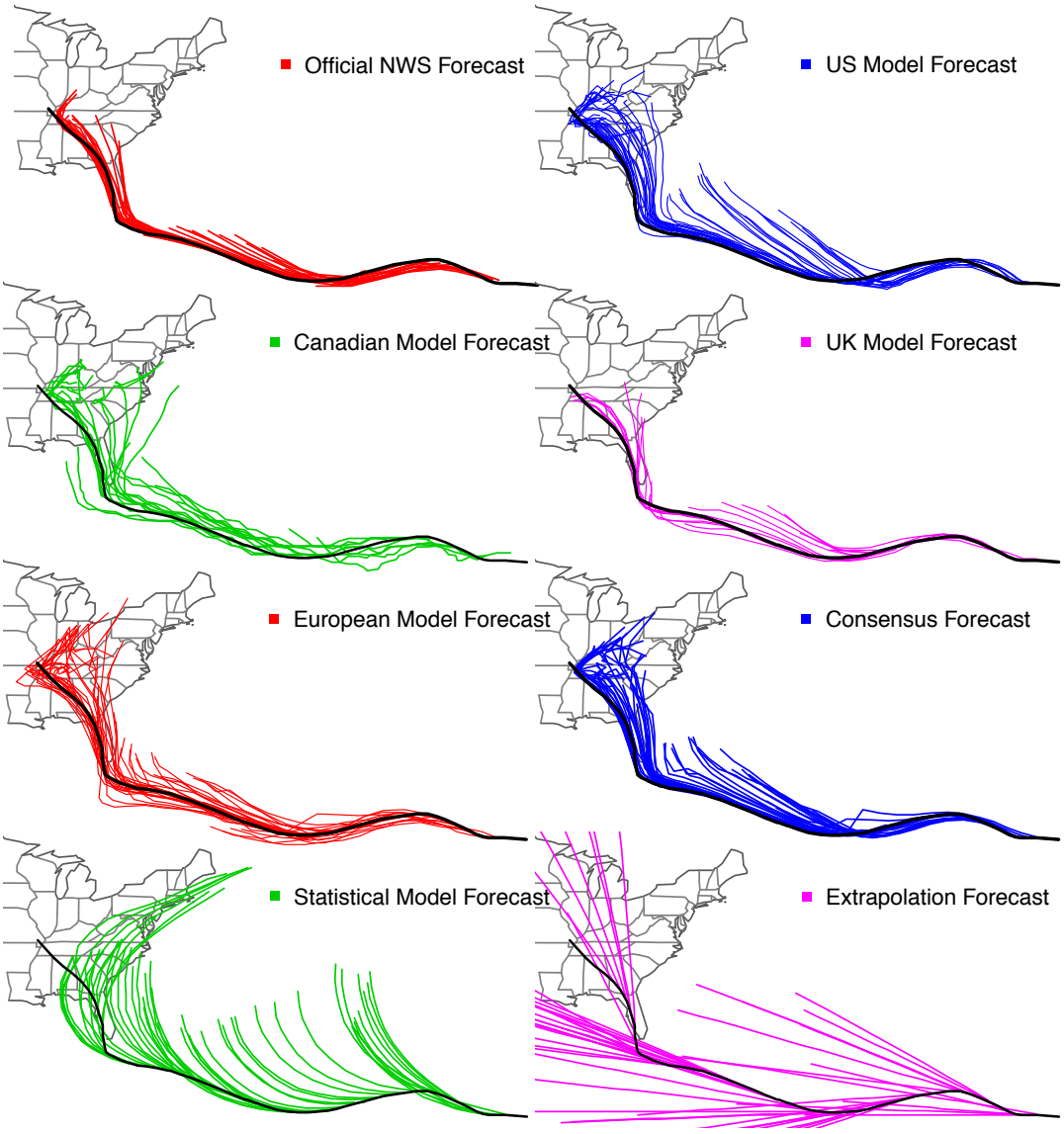
	Path							Points						
H:	12	24	36	48	72	96	120	12	24	36	48	72	96	120
U.S. Global Models:														
AEMI	0.00	0.38	0.24	**0.83**	**0.94**	**0.66**	0.19	0.00	0.04	0.05	0.08	0.00	0.00	0.00
AEMN	0.19	0.07	0.04	0.27	**0.73**	**0.61**	0.19	0.36	0.25	0.32	0.43	0.09	0.05	0.07
AP14	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.28	**0.62**	0.07	0.12	0.07
HMON	0.12	0.20	0.01	0.00	0.00	0.00	0.00	0.28	0.40	**0.87**	**0.76**	0.12	0.08	0.20
HWRF	0.01	0.03	0.24	**0.83**	**0.94**	**0.93**	**0.59**	0.02	0.00	0.00	0.03	0.20	0.22	**0.51**
HWRFI	0.00	0.08	0.32	**0.83**	**0.94**	0.39	0.02	0.00	0.00	0.00	0.01	0.17	0.21	0.21
NGX	**0.82**	0.28	0.00	0.00	0.00	0.00	0.00	**0.82**	**0.95**	**0.97**	**0.95**	0.16	0.11	0.20
NGX2	0.49	0.28	0.00	0.00	0.00	0.00	0.00	**0.53**	**0.56**	0.27	0.17	0.04	0.05	0.14
NGXI	**0.97**	**0.71**	0.02	0.04	0.09	0.16	0.24	**0.97**	**0.85**	0.32	0.36	0.11	0.08	0.19
NVGM	0.31	0.03	0.00	0.00	0.00	0.00	0.00	0.36	0.46	0.45	**0.76**	0.23	0.13	0.20
Canadian Global Models:														
CC00	0.28	0.27	0.03	0.01	0.00	0.00	0.00	0.46	**0.95**	**0.97**	**0.99**	**1.00**	**1.00**	**1.00**
CMC	0.05	0.02	0.02	0.02	0.08	0.00	0.00	0.20	0.67	**0.91**	**0.98**	**0.53**	0.22	0.20
CMC2	0.00	0.08	0.02	**0.59**	0.40	0.01	0.00	0.21	0.24	0.33	0.49	0.24	0.12	0.06
CP01	0.12	0.04	0.00	0.00	0.00	0.00	0.00	0.31	0.24	0.31	**0.62**	0.33	0.22	0.20
CP03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.24	0.57	**0.62**	0.22	0.05	0.07
CP07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.24	**0.82**	**0.98**	0.24	0.05	0.04
CP15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.47	0.31	**0.58**	0.15	0.06	0.10
CP19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.24	0.26	**0.62**	0.29	0.13	0.10
CTCX	**0.67**	0.01	0.00	0.00	0.00	0.00	0.00	**0.70**	0.15	0.05	0.02	0.00	0.00	0.00
U.K. Global Models:														
EGRR	**0.67**	**0.71**	0.08	0.15	0.24	0.01	0.05	**0.70**	**0.95**	**1.00**	**0.99**	0.29	0.08	0.03
UKM	**0.52**	**0.64**	0.24	0.27	**0.52**	0.07	0.14	**0.58**	**1.00**	**0.97**	**0.98**	0.29	0.07	0.03
UKX	0.01	0.07	0.00	0.00	0.01	0.00	0.03	0.04	**0.95**	0.28	**0.62**	0.29	0.07	0.04
UKX2	0.28	0.19	0.02	0.26	**0.70**	0.03	0.02	0.45	0.10	0.03	0.06	0.01	0.00	0.00
UKXI	**1.00**	**0.51**	0.13	0.26	**0.94**	0.31	0.34	**1.00**	**0.95**	0.30	0.43	0.08	0.00	0.00
European Global Models:														
EMXI	0.39	0.01	0.00	0.00	0.00	0.00	0.00	0.48	**0.69**	**0.91**	**1.00**	0.38	0.22	**0.51**
Consensus Models:														
RVCN	0.00	0.40	0.21	**0.83**	**0.94**	**1.00**	1.00	0.03	0.24	0.45	**0.69**	0.27	0.22	**0.51**
TVCN	0.04	0.40	0.32	**0.83**	**0.94**	**0.81**	0.19	0.11	0.05	0.01	0.00	0.00	0.00	0.00
TVCX	**0.97**	**1.00**	**1.00**	**1.00**	**1.00**	**0.94**	**0.78**	**0.97**	**0.95**	**0.97**	**0.98**	0.02	0.02	0.10
TVCY	0.26	0.59	0.32	**0.83**	**0.94**	**0.93**	**0.79**	0.32	0.06	0.10	0.43	0.00	0.01	0.00
Other Physical, Statistical, and Judgemental Methods:														
DRCL	0.10	0.59	0.04	0.02	**0.51**	0.21	0.20	0.33	0.24	**0.52**	**0.62**	0.24	0.22	0.21
DSHP	0.08	0.40	0.08	0.21	**0.51**	**0.51**	0.19	0.33	0.02	0.01	0.42	0.17	0.20	0.20
LGEM	0.08	0.40	0.08	0.21	**0.51**	**0.51**	0.19	0.33	0.02	0.01	0.42	0.17	0.20	0.20
OFCL	**0.97**	**0.71**	0.21	0.27	**0.94**	**0.94**	**0.78**	**0.97**	**0.95**	**0.97**	**0.99**	**0.53**	**0.81**	**0.80**
SHIP	0.08	0.40	0.08	0.21	**0.51**	**0.51**	0.19	0.33	0.02	0.01	0.42	0.17	0.20	0.20
TABM	0.39	0.02	0.00	0.00	0.00	0.00	0.00	0.45	0.24	0.21	**0.62**	0.02	0.03	0.01

Notes: This table presents the p-values for the model confidence set. Only those models that are part of the 50% MCS for at least one horizon and loss are included. Boldface denotes the model is part of the 50% MCS for a given loss and horizon. The definition and a description of each model can be found at `http://hurricanecity.com/models/models.cgi?page=models`

considering the general path loss function. Although the point and path MCS's are similar when $H = 12$, there are few similarities between them at longer horizons with only the 'Official National Weather Service' forecast (OFCL) included with a high probability in most confidence sets.

We can classify the models in the MCS into several distinct groups: (1) 'US Models' produced by the National Oceanic and Atmospheric Administration and the U.S. Navy; (2) 'Canadian Models' from the Canadian Meteorological Center; (3) 'UK Models' from the U.K. Met Office; 'European Models' from the European Center for Medium-range Weather Forecasting (ECMWF); (4) 'Consen-

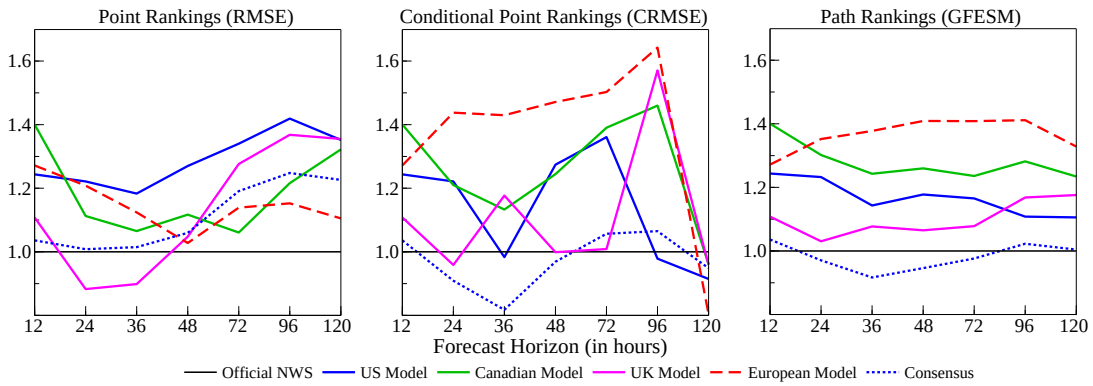
sus Models' which combine the different national model forecasts; and (5) 'Statistical and Judgmental Models' which includes the Official National Weather Service forecasts. To simplify the analysis further, we choose one model from each of these groups along with two naive forecasts.



Notes: The Official NWS forecast (OFCL) is from the National Hurricane Center. The US model (AEMN) is from NOAA's Global Forecasting System (GFS). The Canadian model (CMC) is from the Canadian Meterological Center's global model. The UK Model (EGRR) is the UK Met's global model. The Consensus model (TVCN) a simple average of the US, Canadian, UK, and European models where available. The European model (EMXI) is the ECMWF model forecast. Note that it is not publicly available but was derived from the other forecasts. All plots are restricted to a maximum forecast horizon of 168-hours-ahead where available.

Figure 7.1: Alternative Path Forecasts for Hurricane Irma

Figure 7.1 plots the forecasts generated by each of these models for the entire life of Hurricane Irma for all available horizons up through 168-hours-ahead (7 days) where available. The individual plots provide an initial view of the sometimes very large differences between the various forecast



Notes: CRMSE: Conditional Root MSE. All metrics are estimated based on 37 observations. See notes for Figure 7.1.

Figure 7.2: Relative Forecast Performance for Hurricane Irma

models. It is clear that the statistical model and the naive extrapolation forecast perform considerably worse than any of the other models, particularly at longer horizons. However, many of the forecasts including the Official, US, Canadian, UK, Consensus, and statistical model have a north easterly bias.

Computing the forecast accuracy for each of the forecast models provides another perspective on the relative performance. Figure 7.2 depicts the relative forecast rankings based on point (RMSE), conditional point (CRMSE), and path (GFESM) forecast accuracy. The point rankings indicate that only the UK model does better than the official forecasts and even then only at the 24-hour and 36-hour-ahead horizons. The Consensus forecast performs about as well as the official forecasts initially but then its performance deteriorates beyond 2-days-ahead. The US model consistently underperforms for horizons beyond 24-hours-ahead whereas the European model recovers from a relatively poor short-horizon forecast performance to be the second-best forecast model at 5-days-ahead.

The conditional point rankings estimate forecast performance conditional on the forecast errors at previous horizons. These rankings illustrate that only the Consensus forecast outperforms the official forecasts at shorter horizons. However, by 5-days-ahead all of the forecasts outperform the official forecast where the European model has the best performance. Unsurprisingly, the path rankings are much more persistent. In this case, only the consensus forecasts outperform the official forecasts for the horizons from 24-hours-ahead through to 72-hours-ahead. The UK model is the next best followed by the US model and the Canadian Model. The European model never recovers from its

Table 7.2: Alternative Weights for Hurricane Irma

Horizon	Radius	Equal Weights	Short Weighted	Long Weighted
12	29	1	2.282	0.293
24	45	1	1.471	0.455
36	63	1	1.050	0.637
48	78	1	0.848	0.789
72	107	1	0.618	1.082
96	159	1	0.416	1.608
120	211	1	0.314	2.134

Notes: Radius is in miles. Weights sum to the number of horizons. The radius comes from the National Hurricane Center in 2017: `https://web.archive.org/web/20170906033729/https://www.nhc.noaa.gov/aboutcone.shtml`

relatively poor short-horizon forecasts and remains the worst performing model throughout despite relative improvements at the longest horizon. This indicates that none of the individual model forecasts consistently outperforms the official forecasts across all horizons.

We can test whether differences in the path are statistically significant. This can be done using the general path test and the special case path test since the log conditional MSE's are relatively stable, indicating that the necessary assumptions are satisfied. The general path test can also be used to explore the implications of alternative weighting schemes. While it is common to weight all of the horizons along the path equally, there may be greater interest in either the longer or shorter horizons. The National Hurricane Center produces an ex-ante measure of the uncertainty surrounding the official forecasts at each horizon; i.e. the 'cone of uncertainty'. It is possible to use this expected uncertainty to assign weights across horizons. To discount horizons with more uncertainty we can set weights based on the inverse of the radius of the cone, giving more weight to shorter horizons; see Column 4 in Table 7.2. Alternatively, to focus on horizons with more uncertainty, we can set weights based on the length of the radius, giving more weight to longer horizons; see Column 5 in Table 7.2.

Table 7.3 shows the test results where the entire path is considered and the official forecast is used as a baseline. The equally weighted general path test indicates that all of the forecasts are significantly different from the official forecast except for the Consensus model. Assuming that the nominal test size is correct, then the unweighted special case path test shows that most forecasts are not significantly different. Differences in the general path test are larger and more significant when focusing on shorter horizons and are smaller and less significant when focusing on longer horizons.

Table 7.3: Testing for Differences in Irma's Path Accuracy

	OFCL	US	CAN	UK	EU	Con
Equally weighted GFESM (in miles):	18.4	20.3	22.7	21.6	24.4	18.5
General path test statistic		1.92**	2.27**	1.63*	2.54***	0.11
		[0.027]	[0.012]	[0.051]	[0.006]	[0.457]
Special case path test statistic		0.66	1.38*	1.07	1.87**	0.03
		[0.255]	[0.083]	[0.143]	[0.031]	[0.489]
Short weighted GFESM (in miles):	16.8	19.9	20.8	19.4	22.5	17.0
General path test statistic		2.61***	2.59***	1.39*	2.59***	0.23
		[0.005]	[0.005]	[0.082]	[0.005]	[0.409]
Long weighted GFESM (in miles):	21.7	22.6	26.5	25.8	27.3	21.9
General path test statistic		0.77	1.70**	1.44*	2.43***	0.20
		[0.220]	[0.045]	[0.075]	[0.008]	[0.422]

Notes: See notes for Figure 7.1. The tests consider all horizons through H=120. The p-value associated with the tail probability of the null hypothesis is in brackets: *p< 0.1 **p< 0.05 ***p< 0.01.

This is consistent with the conditional rankings and their spread in Figure 7.2. In fact, we would have to focus just on the longest horizon to find that the European model is more accurate than / no different from the official forecast.

8 Conclusions

This chapter proposes a new approach for testing differences in path forecast accuracy. We argue that the focus on point forecasts ignores forecast dynamics which are crucial to understanding the future trajectory of a series. Instead, additional attention should be paid to the accuracy of the entire forecast path. We show that the GFESM can be thought of as a path forecast accuracy metric and is equivalent to choosing a joint multivariate normal loss function.

We then construct a general test for differences in path forecast accuracy using the link between the joint density and the loss function. We also demonstrate that when using a multivariate normal density, this test statistic is closely linked with other joint multi-horizon tests. Furthermore, we illustrate that it explicitly captures differences in error covariances and dynamics. We also examine and derive the properties of this test for an important special case.

Monte Carlo simulations are used to illustrate the trade-offs associated with our test. We find that although our path forecast accuracy test has lower power to detect differences in biases, it has much higher power to capture differences in variances, covariances and dynamics across forecast

models. This is particularly true for differences in forecast-error dynamics, which other tests are unable to capture. Thus, our test captures a new aspect of differences in forecast accuracy. We also demonstrate that the test statistic can be extended to higher dimensional settings.

Finally we apply the path test in two very different settings. We start by comparing the Federal Reserve Board's Greenbook point and path forecasts with four DSGE model forecasts. We find that while the Greenbook dominates with its point and path inflation forecasts, the DSGE models' ability to track the long-run mean does well for the GDP growth. Decomposing the differences along the path, we find that forecast dynamics play an important role and that there are common differences across models. In the second application we evaluate the path forecast accuracy of models for Hurricane Irma in 2017. We narrow down the more than 100 available methods into a handful of best performing models. While the official forecasts tend to dominate all methods, the path tests show that other models can match its accuracy depending on the specific weights assigned to the path.

Overall the results indicate that differences in path forecast accuracy provide a new and unique perspective. This is because, unlike point forecasts, the path captures dependencies across horizons. The applications show that these dependencies can play a very important role in determining whether there are persistent differences across horizons. These findings coincide with the results from the simulation exercise and demonstrates the value of using path forecast accuracy measures and tests to assess differences in multi-horizon forecasting systems.

A Technical Appendix

A.1 Proof of Theorem 2

Lemma 1. Let $\boldsymbol{X}_{H,T+1},\ldots,\boldsymbol{X}_{H,T+N} \sim N_{HK}(\mathbf{0},\Phi_H)$ where Φ_H has dimension $KH \times KH$ with $KH \leq N$ and $\boldsymbol{X}_{H,T+n} = \left\{X'_{T+n},\ldots,X'_{T+H+n-1}\right\}'$ where each $X_{t+n+h-1}$ is of dimension $K \times 1$ and follows a MA$(h-1)$ process. Denote the sample covariance matrix by $\widehat{\Phi}_{N,H}$. Then

$$ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - ln\left\{det\left(\Phi_H\right)\right\} = ln\left\{det\left(\hat{\mathbf{I}}_{N,KH}\right)\right\},$$

where $\hat{\mathbf{I}}_{H,N,K} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{Y}_{H,T+n}\boldsymbol{Y}'_{H,T+n}$ is the sample matrix of two stacked vectors where $\boldsymbol{Y}_{H,T+n} = \left\{Y'_{T+n},\ldots,Y'_{T+H+n-1}\right\}'$, $Y_{T+1},\ldots,Y_{T+H+N-1} \sim IN_K(\mathbf{0},\mathbf{I}_K)$ and $\mathbf{I}_K$ is an identity matrix.

Proof of Lemma 1. Define $\boldsymbol{Y}_{H,T+n} = \Phi_H^{-1/2}\boldsymbol{X}_{H,T+n}$ which removes any covariance and most of the dependence given the assumed structure so that

$$\begin{aligned}
ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - ln\left\{det\left(\Phi_H\right)\right\} &= ln\left\{det\left(\Phi_H^{-1/2}\widehat{\Phi}_{N,H}\Phi_H^{-1/2}\right)\right\} \\
&= ln\left\{det\left(\frac{1}{N}\sum_{n=1}^{N}\Phi_H^{-1/2}\boldsymbol{X}_{H,T+n}\boldsymbol{X}'_{H,T+n}\Phi_H^{-1/2}\right)\right\} \\
&= ln\left\{det\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{Y}_{H,T+n}\boldsymbol{Y}'_{H,T+n}\right)\right\} \\
&= ln\left\{det\left(\hat{\mathbf{I}}_{N,KH}\right)\right\}.
\end{aligned}$$

Lemma 2. The distribution of $ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\}$ is the same as H times the sum of K-independent log χ^2 distributions such that

$$ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} \overset{D}{=} H\sum_{k=1}^{K} ln\left(\chi^2_{N-k+1}\right),$$

where $\chi^2_N,\ldots,\chi^2_{N-K+1}$ are independent χ^2 distributions with $N-k+1$ degrees of freedom.

Proof of Lemma 2. Define $\mathbf{C}_{h,N,K} = \left\{\frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y'_{T+n},\ldots,\frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y'_{T+n+h-2}\right\}'$

where $\mathbf{C}_{1,N,K} = \mathbf{0}$ and $\hat{\mathbf{I}}_{T,N,K,h} = \frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y_{T+n+h-1}^{'}$. Then by the properties of block determinants it is possible to decompose as

$$ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} = \sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,h}\right)\right\} + \sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K - \hat{\mathbf{I}}_{T,N,K,h}^{-1}\left[\mathbf{C}_{h,N,K}^{'}\hat{\mathbf{I}}_{h-1,N,K}^{-1}\mathbf{C}_{h,N,K}\right]\right)\right\},$$

where $\hat{\mathbf{I}}_{h-1,N,K}$ is invertible as long as $K(h-1) \leq N$. The intuition behind this decomposition follows from the properties of determinants of block matrices; see Silvester (2000). To illustrate these properties, consider a 2×2 block matrix where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are $K(H-1)\times K(H-1)$, $K(H-1)\times K$, $K\times K(H-1)$ and $K\times K$ matrices respectively for $H > 1$ and $K \geq 1$. So the log determinant is

$$\begin{aligned} ln\left\{det\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\right\} &= ln\left\{det\left(\begin{Bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{Bmatrix}\begin{Bmatrix}\mathbf{I}_{K(H-1)} & -\mathbf{A}^{-1}\mathbf{B}\\ \mathbf{0} & \mathbf{I}_K\end{Bmatrix}\right)\right\} \\ &= ln\left\{det\begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{C} & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\end{pmatrix}\right\} \\ &= ln\left\{det\left(\mathbf{A}\right)\right\} + ln\left\{det\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)\right\} \\ &= ln\left\{det\left(\mathbf{A}\right)\right\} + ln\left\{det\left(\mathbf{D}\right)\right\} + ln\left\{det\left(\mathbf{I}_K - \mathbf{D}^{-1}\left[\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right]\right)\right\}, \end{aligned}$$

where $lndet\left(\mathbf{A}\right)$ can also be decomposed so that only deteriminants of $K\times K$ matrices are left.

From the Law of Large Numbers for iid variables we know that as $N \to \infty$ then $\hat{\mathbf{I}}_{T,N,K,h} \xrightarrow{P} \mathbf{I}_K$ and $\frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h}Y_{T+n+j}^{'} \xrightarrow{P} \mathbf{0}\ \forall j \neq h$ so that $\hat{\mathbf{I}}_{h-1,N,K} \xrightarrow{P} \mathbf{I}_{K(h-1)}$ and $\mathbf{C}_{h,N,K} \xrightarrow{P} \mathbf{0}$. Since the log determinant is a continuous function, then by the Continuous Mapping Theorem

$$\sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K - \hat{\mathbf{I}}_{T,N,K,h}^{-1}\left[\mathbf{C}_{h,N,K}^{'}\hat{\mathbf{I}}_{h-1,N,K}^{-1}\mathbf{C}_{h,N,K}\right]\right)\right\} \xrightarrow{P} 0\ as\ N \to \infty.$$

The first term can be decomposed further as

$$\begin{aligned}\sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,h}\right)\right\} &= \sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1} + N\hat{\mathbf{I}}_{T,N,K,h} - N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \\ &= H * ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \\ &\quad + \sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K + \hat{\mathbf{I}}^{-1}_{T,N,K,1}\left[\hat{\mathbf{I}}_{T,N,K,h} - \hat{\mathbf{I}}_{T,N,K,1}\right]\right)\right\}.\end{aligned}$$

Repeatedly applying the Law of Large Numbers for iid variables to each of the individual terms, then by the Continuous Mapping Theorem

$$\sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K + \hat{\mathbf{I}}^{-1}_{T,N,K,1}\left[\hat{\mathbf{I}}_{T,N,K,h} - \hat{\mathbf{I}}_{T,N,K,1}\right]\right)\right\} \xrightarrow{P} 0 \; as \; N \to \infty.$$

Finally, for fixed H as $N \to \infty$

$$H * ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \overset{D}{=} H \sum_{k=1}^{K} ln\left(\chi^2_{N-k+1}\right),$$

which follows from the Bartlett decomposition where each variable is independent as $N \to \infty$; see Anderson (2003) and Muirhead (2005). Combining each of these results using the Continuous Mapping Theorem for fixed H and K as $N \to \infty$ proves the Lemma.

Theorem (Asymptotic Distribution). Let $\boldsymbol{X}_{H,T+1},\ldots,\boldsymbol{X}_{H,T+N} \sim N_{HK}(0,\Phi_H)$ where Φ_H has dimension $KH \times KH$ with $KH \leq N$ and $\boldsymbol{X}_{H,T+n} = \{X_{T+n},\ldots,X_{T+H+n-1}\}'$ where each $X_{t+n+h-1}$ is of dimension $K \times 1$ and follows a MA$(h-1)$ process. Suppose that $N \to \infty$. Then the log determinant of the sample second-moment matrix $\widehat{\Phi}_H$ satisfies

$$\frac{ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - \tau_{N,K,H} - ln\left\{det\left(\Phi_H\right)\right\}}{\sigma_{N,K,H}} \to N(0,1) \; as \; N \to \infty,$$

where constants $\tau_{N,K,H}$ and $\sigma_{N,K,H}$ are given as

$$\tau_{N,K,H} := \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) - ln\left(\frac{N}{2}\right)\right],$$

where $\Psi(x) = \frac{\partial}{\partial z} log\left(\Gamma(z)\right)|_{z=x}$, where $\Gamma(\cdot)$ is the gamma function and

$$\sigma_{N,K,H} := H\left(\sum_{k=1}^{K} \frac{2}{N-k+1}\right)^{\frac{1}{2}}.$$

Proof of Theorem. It follows from **Lemma 1** and **Lemma 2** that

$$\begin{aligned} z = ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - \tau_{N,K,H} - ln\left\{det\left(\Phi_H\right)\right\} &= ln\left\{det\left(\hat{\mathbf{I}}_{N,KH}\right)\right\} - \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) - ln\left(\frac{N}{2}\right)\right] \\ &= ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} - \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) + ln(2)\right] \\ &\overset{D}{=} H\sum_{k=1}^{K}\left[ln\left(\chi^2_{N-k+1}\right) - \Psi\left(\frac{N-k+1}{2}\right) + ln(2)\right]. \end{aligned}$$

Using the characteristic function of the log χ^2 distribution gives

$$\begin{aligned} \phi_z(t) &= \prod_{k=1}^{K} \phi_{H*ln\left(\chi^2_{N-k+1}\right)}(t)\, \frac{1}{exp\left\{it*H\left[\Psi\left(\frac{N-k+1}{2}\right) + ln(2)\right]\right\}} \\ &= \prod_{k=1}^{K} \mathbb{E}\left(\chi^2_{N-k+1}\right)^{it*H} \frac{1}{exp\left\{it*H\left[\Psi\left(\frac{N-k+1}{2}\right)\right]\right\} 2^{itH}}. \end{aligned}$$

Then applying the formula for moments from the moment generating function of the χ^2 distribution:

$$\begin{aligned} \phi_z(t) &= \prod_{k=1}^{K}\left[\frac{\Gamma\left(\frac{1}{2}(N-k+1) + it*H\right)}{\Gamma\left(\frac{1}{2}(N-k+1)\right)}\right] 2^{it*H} \frac{1}{exp\left\{it*H\left[\Psi\left(\frac{N-k+1}{2}\right)\right]\right\} 2^{it*H}} \\ &= \prod_{k=1}^{K}\left[\frac{\Gamma\left(\frac{1}{2}(N-k+1) + it*H\right)}{\Gamma\left(\frac{1}{2}(N-k+1)\right)}\right] \frac{1}{exp\left(it*H\left(\Psi\left(\frac{N-k+1}{2}\right)\right)\right)}. \end{aligned}$$

Taking logs gives

$$ln\left\{\phi_z(t)\right\} \overset{D}{=} \sum_{k=1}^{K}\left\{ln\left\{\Gamma\left(\frac{1}{2}(N-k+1)+it*H\right)\right\}-ln\left\{\Gamma\left(\frac{1}{2}(N-k+1)\right)\right\}-it*H\left(\Psi\left(\frac{N-k+1}{2}\right)\right)\right\}.$$

Cai et al. (2015) show that when $H = 1$ this can be approximated as

$$ln\left\{\phi_z(t)\right\} \overset{D}{\approx} (it)^2\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left((t)^2\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right).$$

Extending this approximation to $H \geq 1$ gives

$$\begin{aligned}
\phi_o(t) &= \phi_z\left(\frac{t}{\sigma_{N,K,H}}\right)\\
&\overset{D}{\approx} exp\left\{\frac{(it*H)^2}{\sigma^2_{N,K,H}}\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left(\frac{(t*H)^2}{\sigma^2_{N,K,H}}\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right)\right\}\\
&= exp\left\{\frac{(it)^2}{\sum_{k=1}^{K}\frac{2}{(N-k+1)}}\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left(\frac{t^2}{\sum_{k=1}^{K}\frac{2}{(N-k+1)}}\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right)\right\}\\
&= exp\left\{\frac{(it)^2}{2}+O\left(\frac{t^2}{2}\frac{\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}}{\sum_{k=1}^{K}\frac{1}{(N-k+1)}}\right)\right\}\\
&= exp\left\{\frac{(it)^2}{2}+O\left(\frac{t^2}{2}r_{N,K}\right)\right\},
\end{aligned}$$

where Cai et al. (2015, Lemma 3) shows that $r_{N,K} \to 0$ as $N \to \infty$ when $K \leq N$. As a result, when $N \to \infty$ then this gives the characteristic function of the standard normal distribution:

$$\phi_o(tH) \to e^{\frac{(it)^2}{2}}.$$

A.2 Proof of Corollary 2

Let $\boldsymbol{X}_{H,T+1},\ldots,\boldsymbol{X}_{H,T+N} \sim N_{HK}(\boldsymbol{\Theta_H},\Phi_H)$ where Φ_H and $\boldsymbol{\Theta_H}$ have dimension $KH \times KH$ with $KH \leq N$ and $\boldsymbol{X}_{H,T+n} = \left\{X'_{T+n},\ldots,X'_{T+H+n-1}\right\}'$ where each $X_{t+n+h-1}$ is of dimension $K \times 1$ and follows a MA$(h-1)$ process. Denote the sample second-moment matrix by $\widehat{\Phi}_{N,H}$. Following Muirhead (2005, Chapter 10), the sample second-moment matrix has a non-central Wishart distribution where $N\widehat{\Phi}_{N,H} \sim W_{HK}(N,\Phi_H,\boldsymbol{\eta})$ where $\boldsymbol{\eta} = \Phi_H^{-1}\Theta_H$ is the non-centrality.

Fujikoshi (1968, Theorem 1) shows for $N\boldsymbol{\eta} = O(N)$ and $lim\sqrt{N}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = 0$

$$\frac{\sqrt{N}\left\{\left(\left|\widehat{\Phi}_{N,H}\right|/|\Phi_H|\right) - |\mathbf{I}_{HK} + \boldsymbol{\Omega}_H|\right\}}{\sqrt{2 * H * |\mathbf{I}_{HK} + \boldsymbol{\eta}|^2 Tr\left\{(\mathbf{I}_{HK} + 2\boldsymbol{\eta})(\mathbf{I}_K + \boldsymbol{\eta})^{-2}\right\}}} \rightarrow N(0,1).$$

Now consider the delta-method from Casella and Berger (2002, Theorem 5.5.24) where X is a sequence of random variables that satisfies $\sqrt{N}(X-\mu) \rightarrow N\left(0,\sigma^2\right)$ in distribution. For a continuous function $g(\cdot)$ and a specific value of μ, suppose that $g'(\mu)$ exists and is non-zero. The delta-method gives that $\sqrt{N}(g(X) - g(\mu)) \rightarrow N\left(0,\sigma^2 g'(\mu)^2\right)$.

So if $X = \left|\widehat{\Phi}_{N,H}\right|/|\Phi_H|$, $\sigma^2 = 2H|\mathbf{I}_{HK} + \boldsymbol{\eta}|^2 Tr(\mathbf{I}_{HK} + 2\boldsymbol{\eta})(\mathbf{I}_{HK} + \boldsymbol{\eta})^{-2}$, $\mu = |\mathbf{I}_{HK} + \boldsymbol{\eta}|$, and $g(\cdot) = ln(\cdot)$ so that $g'(\mu)^2 = |\mathbf{I}_{HK} + \boldsymbol{\Omega_H}|^{-2}$, then applying the delta-method gives

$$\frac{\sqrt{N}\left\{ln\left(\left|\widehat{\Phi}_H\right|\right) - ln(|\Phi_H|) - ln(|\mathbf{I}_{HK} + \boldsymbol{\eta}|)\right\}}{\sqrt{2H * Tr\left[(\mathbf{I}_{HK} + 2\boldsymbol{\eta})(\mathbf{I}_{HK} + \boldsymbol{\eta})^{-2}\right]}} \rightarrow N(0,1),$$

This is similar to Theorem 2 except that it excludes the finite sample bias correction terms and allows for non-centralities. Applying this result to a two-sample test with the additional assumption that the underlying shocks are uncorrelated across forecast systems, i.e. $\mathbb{E}\left(\boldsymbol{X}_{H,T+n,1}\boldsymbol{X}'_{H,T+n,2}\right) = \mathbf{0}$, then

$$\frac{-\sqrt{N}\nabla ln\left(\left|\widehat{\Phi}_{N,H}\right|\right)}{2\sqrt{H * Tr\left[(\mathbf{I}_{HK} + 2\boldsymbol{\eta})(\mathbf{I}_{HK} + \boldsymbol{\eta})^{-2}\right]}} \rightarrow N(0,1) \; as \; N \rightarrow \infty,$$

which simplifies to Corollary 1 when $\boldsymbol{\eta} = \mathbf{0}$. If the true non-centrality is unobserved an estimator, $\widehat{\boldsymbol{\eta}}_N$, can be used so that the denominator becomes $\sqrt{2H * \sum_{i=1}^{2} Tr\left[\left(\mathbf{I}_{HK} + 2\widehat{\boldsymbol{\eta}}_{\mathbf{N,i}}\right)\left(\mathbf{I}_{HK} + \widehat{\boldsymbol{\eta}}_{\mathbf{N,i}}\right)^{-2}\right]}$.

Chapter 4

The Impact of Forecast Accuracy on Hurricane Damages

Can forecasts of natural disasters alter their destructiveness? Poor forecasts increase damages when individuals do not mitigate risks based on the false belief that they will be unaffected. We test this hypothesis by examining the impact of 12-hour-ahead forecasts on hurricane damages and find that larger errors in a storm's predicted landfall location lead to higher damages. We estimate that the cumulative reduction in damages from forecast improvements since 1970 is about $82 billion. This exceeds the U.S. government's spending on the forecasts and private willingness to pay for them. The benefits from forecast improvements are underestimated and individual adaptation decisions matter.

1 Introduction

Damages from natural disasters in the United States, driven in part by several hurricanes, reached a record high of $313 billion in 2017 (NOAA NCEI, 2018). Hurricanes now account for seven of the top ten costliest disasters in the United States since 1980. They have substantial impacts on local economic growth (Strobl, 2011), fiscal outlays (Deryugina, 2017), lending and borrowing patterns (Gallagher and Hartley, 2017), and on where people live and work (Deryugina et al., 2018). Despite their widespread effects, natural disasters are localized events. Their costs are determined in part by individuals' decisions about how and when to protect their property. Decisions about whether to board up windows, stack sandbags, harvest crops, relocate property etc. have to be made in advance. They rely on forecasts of the event's occurrence, location, and severity produced up to several days ahead. These forecasts, despite dramatic improvements over time, are far from perfect and can exhibit large errors. Large and unexpected forecast errors, even up to just a few hours ahead, may lead people in the disaster area to protect their property less than they would have otherwise. Individuals who place too much faith in the forecasts may make poor decisions which result in higher damages.

This chapter seeks to quantify how the accuracy of short-term forecasts leading up to a natural disaster affects the resulting destruction. While a considerable amount of research examines changes in the natural hazards and vulnerabilities associated with natural disasters, little attention has been paid to the effectiveness of short-term damage mitigation decisions just prior to a disaster's occurrence. Even less research focuses on how early warning systems based on short-term forecasts of the event affect these decisions. Quantifying the extent to which short-term forecasts of natural disasters help individuals make better damage mitigation decisions is important for understanding the effectiveness of short-term adaptation efforts, for illuminating the drivers of the rising costs of natural disasters, and for understanding the speed of the ultimate economic recovery.

We start by showing that, in an expected utility framework, errors in the forecasts of a natural disaster can affect damages if individuals place too much faith in the accuracy of these forecasts

when formulating their beliefs about the benefits of damage mitigation. Next, using our theoretical framework as a guide we test for evidence of these effects using a newly constructed database of all hurricanes to strike the continental United States from 1955-2014. We formulate an empirical model of damages with many determinants including the 12-hour-ahead landfall-forecast errors and estimate the model using ordinary least squares. Then, we use model selection methods to determine the best model specification and also whether forecast errors are among the most important determinants of damages. The robustness of our results is confirmed using a variety of selection procedures and model specifications. Finally, we conduct a counterfactual exercise in order to quantify how much improvements in short-term forecast accuracy since 1970 have altered hurricane damages.

Short-term forecast errors of a storm's location, together with a handful of other variables, explain most of the variation in aggregate hurricane damages over the past sixty years. A one standard deviation increase in the distance between where a hurricane is expected to strike and where it actually strikes is associated with up to $3,000 in additional damages per household affected by the hurricane. Interpreting this through the lens of our theoretical framework indicates that short-term forecasts guide individual damage mitigation decisions. In aggregate, individual decisions to protect and relocate property in the face of a disaster have a significant impact on the overall costs.

The U.S. government devotes considerable resources to improving its hurricane forecasts despite few studies of their value beyond reductions in fatalities. We quantify the static reduction in hurricane damages due to short-term forecast improvements. The predicted cumulative damages prevented due to improvements in forecast accuracy since 1970 is about $82 billion. This means that the cumulative net benefit is between $\$30-71$ billion after accounting for what the U.S. federal government spends on hurricane operations and research. The benefits of further forecast improvements also outweigh measures of society's 'willingness to pay' for them; see Katz and Lazo (2011).

This chapter contributes to several strands of the literature. The first is related to the measurement of forecast uncertainty. Analyses of uncertainty go back at least as far as Knight (1921), who distinguished between risk, which is measurable ex-ante, and 'Knightian' uncertainty which is not. Recent

work defines uncertainty predominantly in terms of second moments; see Bloom (2009, 2014). This is less true for forecast uncertainty, since as Jurado et al. (2015, p. 1178) argue, "what matters for economic decision making is [...] whether the economy has become more or less predictable; that is, less or more uncertain." We extend Rossi and Sekhposyan (2015)'s measure of forecast error uncertainty to allow for time-varying densities. We show how and under what conditions we can approximate this measure as the ratio of forecast errors to their ex-ante risk. This allows us to test for whether forecast accuracy plays a separate role from ex-ante uncertainty in altering damages from natural disasters; see Kruttli et al. (2019).

We also contribute to the literature on the effectiveness of adaptation to natural disasters. Our work is related to Bakkensen and Mendelsohn (2016), who estimate a model of global tropical cyclone damages to understand the relationship between income and adaptation in natural disasters. Although they find evidence of adaptation globally, they do not find any in the United States. We extend their approach by allowing forecast errors to alter adaptation decisions. This allows us to measure short-term adaptation efforts separately from income. We find evidence of adaptation in the United States while also finding that higher income is not associated with lower damages.

Finally, we also contribute to the literature on the value of environmental forecasts. Our work relates to a large number of studies including Krzysztofowicz and Davis (1983), Carsell et al. (2004), Regnier (2008) and Pappenberger et al. (2015). When assessing the value of hurricane forecasts, previous studies typically focus on the value of improved evacuation decisions and reduced fatalities. Our analysis illustrates that the benefits of accurate forecasts also play an important role in short-term damage mitigation decisions. This allows us to consider the cost of forecast improvements and the private willingness to pay for those improvements in a different context and we show that the benefits to forecast improvements are even higher than previously considered.

The rest of the chapter is structured as follows: Section 2 explores the link between natural disasters, damages, and uncertainty. Section 3 proposes a theoretical model of how forecast uncertainty can impact damages. The rest of the chapter applies this model to an application of damages

from hurricane strikes where section 4 describes the statistical methods and the data used. Section 5 presents the results while section 6 assesses their robustness. Section 7 discusses the implications of improving the forecasts and section 8 concludes.

2 Natural Disasters: Damages and Imperfect Forecasts

The destruction from a wide range of natural disasters, including wild fires, hurricanes, tornadoes, droughts, and floods, reached a record breaking $313 billion in the United States in 2017 (NOAA NCEI, 2018). This likely underestimates the total cost as natural disasters have both persistent and transitory economic impacts. Strobl (2011) argues that a one percent increase in the direct cost of a natural disaster is associated with a transitory decline in local economic growth by 0.45 percentage points. Baker and Bloom (2013) also find that natural disasters precede declines in economic activity. Deryugina (2017) finds that natural disasters are associated with additional widespread direct and indirect fiscal costs. Gallagher and Hartley (2017) find that Hurricane Katrina [2005] altered borrowing patterns and spurred efforts to deleverage, while Deryugina et al. (2018) finds that it led to lasting changes on where people live, but only temporary changes otherwise. This illustrates that natural disasters can have widespread and often lasting economic impacts.

There are many determinants of the destructiveness of a natural disaster. Natural hazards such as the maximum wind speed of a hurricane, its storm surge, rainfall, and minimum central pressure are considered to be the most important; see Nordhaus (2010), Murnane and Elsner (2012) and Chavas et al. (2017). Damages are also determined by the vulnerabilities of a location. This is often measured by how much income, housing or capital stock there is in an area; see Pielke Jr and Landsea (1998), Pielke Jr et al. (2008), and Neumayer and Barthel (2011). There is also a growing literature on medium and longer-term efforts to mitigate damages. This is captured through incomes, building codes, and government spending on damage mitigation programs; see Bakkensen and Mendelsohn (2016), Geiger et al. (2016), Dehring and Halek (2013) and Davlasheridze et al. (2017).

Little attention has been paid to the role that early warning systems for natural disasters and their

forecasts could play in altering the destructiveness. For example, Deryugina et al. (2018, p. 202) claim that Hurricane Katrina [2005] "struck with essentially no warning", which ignores that there were warnings in place several days before the storm struck. On the other hand, Letson et al. (2007) argue that there is a trade-off between damage mitigation efforts and forecast improvements in the medium-term. In their view, earlier warnings and better forecasts of a natural disaster may lead to moral hazard concerns as people are more likely to move into higher risk locations due to the declining risk of fatalities. This in turn would lead to increased vulnerabilities and higher damages. Sadowski and Sutter (2005) claim that the negative bivariate correlation between damages and fatalities and between damages and forecast errors (see Appendix Table A.1) provides evidence in support of this argument. However, these simple relationships are most likely confounded by the longer-term trends in technological change and economic growth.

We focus instead on the short-term impact of forecasts on damages. We argue that forecasts matter because they are used for planning and mitigation decisions. Behavioral response surveys conducted by the U.S. Army Corps of Engineers (2004) illustrate this link:

> "Many people believe the storm will miss their location, sometimes placing too much faith in the forecast track of the storm, and sometimes those misconceptions are reinforced by similar misconceptions by emergency management officials. In some cases, 40% of the respondents said they have never spent anything to make their homes safer in hurricanes".

This finding, which is reaffirmed in more recent surveys (Milch et al., 2018) and in the purchases of emergency supplies (Beatty et al., 2019), suggests that individuals make decisions using the forecasts. They form their beliefs and make damage mitigation decisions based on imperfect forecast information. This provides a basis through which poor forecasts can increase damages. The next section formalizes this relationship within a theoretical model of damages from natural disasters.

3 Theoretical Framework

We start by developing a theoretical framework to illustrate how damage mitigation decisions can provide a link between forecasts and damages. Consider the expected expenditures when facing the risk of a natural disaster. Damages, $d(\cdot)$, occur with probability $p(\cdot)$, and a cost $c(\cdot)$ of mitigating them. Damages depend on the event's location and severity, F, the vulnerable assets, V, and any adaptation efforts undertaken, A.

We assume that the sequence of actions are: (a) nature determines the location and severity of the event, F; (b) individuals form imperfect beliefs about the event, $\hat{F}$, and the probability $p(\hat{F})$ of damages $d(\cdot)$; (c) they then choose a level of adaptation, A, based on their beliefs, the likely damages, and costs; (d) the event occurs based on nature's predetermined location and severity. Since individuals are not observed directly, we assume that they face similar choices based on publicly available forecast information and so the aggregate decisions across individuals can be modeled as a single representative agent. Future extensions could relax this assumption.

We can write the choice in (c) as an expenditure minimization problem under uncertainty:

$$\min_{A} p(\hat{F})d(V,\hat{F},A) + c(A). \tag{3.1}$$

This formulation dates back to Von Neumann and Morgenstern (1945) where the probability of the event is known. Modifications adapt this formulation to allow for the probability to be unknown; see Kahneman and Tversky (1979) among others. The typical setup allows for individuals to be ex-ante uncertain (or ambiguous) about probabilities by integrating over all possible future states. A false sense of security occurs when there is no ex-ante uncertainty and where probabilities are determined exclusively by point forecasts of the event.

We denote the expenditure function in (3.1) by $E(A)$. Taking the derivative w.r.t A gives

$$\frac{\partial E(A)}{\partial A} = E'(A) = p(\hat{F})d'(V,\hat{F},A) + c'(A). \tag{3.2}$$

Assuming that $d(\cdot)$ is convex in A is necessary and sufficient for a solution. Then the optimal level of adaptation A^* (conditional on ex-ante beliefs about $\hat{F}$) satisfies

$$d'(V,\hat{F},A^*) = -\frac{c'(A^*)}{p(\hat{F})}, \tag{3.3}$$

where we assume that $0 < c'(A) < \infty$ and $p(\hat{F}) > 0$. This implies that individuals should mitigate up until the marginal benefit of adaptation is equal to the ratio between the marginal cost of adaptation and the perceived probability of damages occurring.[1] Thus, adaptation choice is a function of individuals' beliefs about the event which we denote as: $A(\hat{F})$. When $p(\hat{F})$ is small ex-ante, whatever the outcome ex-post, individuals will not undertake much adaptation even facing large potential damages.

Even when beliefs are formed without any ex-ante uncertainty, they can be wrong or uncertain ex-post. We can assess the impact of this ex-post uncertainty by examining the difference between the ex-ante and ex-post optimal marginal benefit of adaptation. If adaptation is optimal in both cases, then with some algebraic manipulation this is written as:

$$d'(V,F,A(F)) - d'(V,\hat{F},A(\hat{F})) = \frac{1}{p(\hat{F})}\left[c'(A(\hat{F})) - c'(A(F))\right] + \left[\frac{p(F)-p(\hat{F})}{p(\hat{F})}\right]\frac{c'(A(F))}{p(F)}. \tag{3.4}$$

The first term in (3.4) represents the difference in the ex-post and ex-ante marginal costs and the second term is driven by the error in the perceived probability of damages.

From (3.4), the marginal reduction in damages is lower ex-post than it is ex-ante if and only if $\frac{c'(A(\hat{F}))-c'(A(F))}{c'(A(F))} > \frac{p(\hat{F})-p(F)}{p(F)}$. When marginal costs are constant or increasing then $p(\hat{F}) < p(F)$ is required. If this holds, then the marginal benefit of additional adaptation is greater ex-ante than it is ex-post. Thus, when $\hat{F} < F$, adaptation will be lower, $A(\hat{F}) < A(F)$, and damages higher, $d(V,F,A(F)) \leq d(V,F,A(\hat{F}))$, than they otherwise would have been.

If adaptation choices are sub-optimal, then the marginal cost is not a binding constraint and (3.4)

[1] This is analogous to the classic optimal demand for insurance problem.

does not hold. However, a second order Taylor series expansion of $d(V,F,A(\hat{F}))$ around F illustrates that as long as damages are convex in adaptation then $d(V,F,A(F)) < d(V,F,A(\hat{F}))$ if $\hat{F} < F$. Thus, irrespective of the optimality of adaptation, our theoretical framework predicts that higher ex-post uncertainty about an event in the context of an ex-ante optimistic bias is associated with lower levels of adaptation and higher damages.

This framework provides a guide for thinking about a realistic model of damages from natural disasters. We assume, as is common in the literature, e.g. see Bakkensen and Mendelsohn (2016), that damages can be represented by a Cobb-Douglas power function. Taking logarithms gives the expression:

$$ln(d_i) = c + \boldsymbol{\alpha} ln(\boldsymbol{V}_i) + \boldsymbol{\beta} ln(\boldsymbol{F}_i) - \boldsymbol{\delta} ln\left(A_i(\hat{\boldsymbol{F}}_i)\right), \tag{3.5}$$

where bold terms are vectors, logs of bold terms are vectors of variables in logs, and damages are convex in adaptation if $\boldsymbol{\delta} > 0$. This extends Bakkensen and Mendelsohn (2016) to allow for adaptation based on imperfect information. We assume that adaptation is described by the log-linear relationship in Bakkensen and Mendelsohn (2016): $ln(A_i(\boldsymbol{F}_i)) = \boldsymbol{\gamma}_1 ln(\boldsymbol{V}_i) + \boldsymbol{\gamma}_2 ln(\boldsymbol{F}_i)$ and reparameterize (3.5) by adding and subtracting adaptation under perfect information to get

$$ln(d_i) = c + (\boldsymbol{\alpha} - \boldsymbol{\delta\gamma_1}) ln(\boldsymbol{V}_i) + (\boldsymbol{\beta} - \boldsymbol{\delta\gamma_2}) ln(\boldsymbol{F}_i) + \boldsymbol{\delta\gamma_2}\left[ln(\boldsymbol{F}_i) - ln\left(\hat{\boldsymbol{F}}_i\right)\right], \tag{3.6}$$

where the final term in (3.6) captures the distance between the actual and predicted intensity and/or location of the storm. This corresponds directly to the interpretation of (3.4) above where an optimistic bias in private beliefs feeds into ex-ante adaptation decisions: $P_i(F_i) - P_i(\hat{F}_i) \approx A_i(F_i) - A_i(\hat{F}_i)$ and then into higher damages.

The formulation in (3.6) does not account for uncertainty around the forecast. To capture the joint impact of forecast accuracy and uncertainty around the forecast, we can replace the final term in (3.6)

with a general measure of forecast uncertainty:

$$ln(d_i) = c + (\boldsymbol{\alpha} - \boldsymbol{\eta\gamma}_1)ln(\boldsymbol{V}_i) + (\boldsymbol{\beta} - \boldsymbol{\eta\gamma}_2)ln(\boldsymbol{F}_i) + \boldsymbol{\eta} ln\left(U_i(\boldsymbol{F}_i, \hat{\boldsymbol{F}}_i)\right) + \epsilon_i, \tag{3.7}$$

where $U_i(\cdot)$ jointly captures accuracy and uncertainty about or surrounding the forecast and ϵ_i indicates the residual or approximation error when going from (3.6) to (3.7).

There are many measures of forecast uncertainty. A popular measure for point forecasts is the mean square forecast error (MSE); see Ericsson (2001). While the MSE produces a fixed measure over time, Jurado et al. (2015) propose a time-varying measure which combines MSE's across variables using a dynamic factor model with stochastic volatility. A popular measure for density forecasts is the log score (Mitchell and Wallis, 2011) which evaluates the predicted density, $\hat{g}(\cdot)$, at $\boldsymbol{F}_i$ conditional on the prediction $\hat{\boldsymbol{F}}_i$. When $\boldsymbol{F}_i$ falls in the tails of $\hat{g}(\cdot)$, it has a lower probability and so is associated with higher forecast uncertainty. Another measure is the continuous ranked probability score (CRPS), which compares observations against the predicted cumulative distribution function. However, these measures do not distinguish between alternative tails of the distribution.

Rossi and Sekhposyan (2015) propose a measure of forecast uncertainty based on the unconditional likelihood of the observed outcome. Their measure, in the context of a single forecast $\hat{F}_{1i}$, is computed by evaluating the predicted cumulative distribution function at its true value F_{1i}

$$U_{1,i}\left(F_{1i}, \hat{F}_{1i}\right) = \int_{-\infty}^{F_{1i}} \tilde{g}_i\left(x_1 | \hat{F}_{1i}\right) dx_1, \tag{3.8}$$

which captures how likely it is to observe F_{1i} given the predicted distribution. The innovation is that $\tilde{g}_i(\cdot)$ is computed using historical forecast errors. An important distinction between (3.8) and Rossi and Sekhposyan (2015) is that we allow $\tilde{g}_i(\cdot)$ to change across events (or time) so as to capture changes in the distribution (Hendry and Mizon, 2014). This ensures that $\tilde{g}_i(\cdot)$ is measured using information most relevant for adaptation decisions. If F_{1i} is large relative to $\hat{F}_{1i}$ and the difference is big compared to past prediction errors, then there is more uncertainty than when F_{1i} is small relative

to $\hat{F}_{1i}$. Thus, (3.8) captures both forecast accuracy and the uncertainty surrounding the forecast.

This measure extends to multiple variables and/or forecast horizons through the joint predictive distribution function. If we define $\boldsymbol{F}_i = \{F_{1i}, \ldots, F_{Ji}\}'$ for up to a total of J variables and or horizons, then the multivariate extension of (3.8) is

$$U_{J,i}\left(\boldsymbol{F}_i, \hat{\boldsymbol{F}}_i\right) = \int_{-\infty}^{F_{1i}} \ldots \int_{-\infty}^{F_{j,i}} \tilde{g}_i\left(\left\{x_{1i,}, \ldots, x_{Ji}\right\}' | \hat{\boldsymbol{F}}_i\right) dx_1 \ldots dx_J, \tag{3.9}$$

which captures both variable specific and common aspects of uncertainty across variables. The measure can also be decomposed to illustrate what information would be lost if multiple variables or horizons actually mattered but only a single variable and horizon was considered:

$$ln\left(U_{J,i}(\boldsymbol{F}_i, \hat{\boldsymbol{F}}_i)\right) = ln\left(U_{1,i}(F_{1i}, \hat{F}_{1i})\right) + ln\left(\int_{-\infty}^{F_{2i}} \ldots \int_{-\infty}^{F_{j,i}} \tilde{g}_i\left(\left\{x_{2i,}, \ldots, x_{ji}\right\}' | F_{1i}, \hat{\boldsymbol{F}}_i\right) dx_2 \ldots dx_J\right), \tag{3.10}$$

where the second term in (3.10) represents information loss. If the unobserved terms are highly correlated with the observed terms, then the variation of the conditional density will be limited. In this case, focusing only on the observed terms results in limited information loss.

Further simplification occurs if $\tilde{g}_i(\cdot)$ follows a normal distribution. Technical Appendix B.1 shows that for small distances between F_i and $\hat{F}_i$, (3.8) can be approximated as

$$U_{1,i}(F_{1i}, \hat{F}_{1i}) \approx \frac{F_{1i} - \hat{F}_{1i}}{\hat{\sigma}_{i,(F_{1_}, \hat{F}_{1_})}}, \tag{3.11}$$

where $\hat{\sigma}_{i,(F_{1_}, \hat{F}_{1_})}$ is the standard deviation of the predicted distribution based on historical forecast errors. It represents the ex-ante risk (in a Knightian sense) ascribed to the forecast at the time of the forecast. When (3.11) has an absolute value greater than one, then the forecast error falls outside of its expected mid-range and is associated with greater uncertainty. An absolute value less than one indicates there is less uncertainty since the forecast error is within the expected range. This is also related to previous comparisons of ex-post and ex-ante forecast uncertainty; see Chow (1960), Clements (2014) and Rossi et al. (2018).

We can generalize this measure further by taking logs and relaxing the fixed 1-to-1 relationship between forecast accuracy and ex-ante risk

$$ln\left(U_{1,i}(F_{1i},\hat{F}_{1i})\right) \approx \frac{\eta_1}{\eta}\ln(\left|F_{1i}-\hat{F}_{1i}\right|)+\frac{\eta_2}{\eta}\ln(\hat{\sigma}_{i,(F_{1_},\hat{F}_{1_})}), \tag{3.12}$$

which allows us to assess their relative importance for damages.[2] In the context of natural disasters, forecast accuracy captures the errors in beliefs that individuals have about the location or severity of a disaster while ex-ante risk modulates the strength of those beliefs. Plugging (3.12) back into (3.7), allows us to test the hypothesis that forecast accuracy matters after accounting for the ex-ante uncertainty. The next section discusses how to assess this relationship in the context of short-term forecasts and damages from hurricane strikes.

4 Application: Short-term Hurricane Forecasts and Damages

Tropical cyclones are powerful natural events that occur intermittently around the globe. They are an intrinsic part of Earth's climate system in that they are fueled by large air-sea surface temperature differentials and mix different ocean layers to help distribute heat (Emanuel, 2001). Tropical cyclones are also among the most destructive climate events, accounting for six of the top ten costliest global natural disasters since 1980 (MunichRe, 2018). Hurricanes, which are tropical cyclones that occur in the Atlantic and northeastern Pacific oceans, account for seven of the top ten costliest weather and natural disasters in the United States over the same period (NOAA NCEI, 2018). Therefore, they provide an important application on which to test the implications of our framework.

Hurricane forecasts have a long history in the United States. The U.S. government has produced hurricane forecasts since at least the 1850s. The National Hurricane Center (NHC) took over this responsibility when it was established in the 1950s. In the 60 years since, there have been large changes in the forecasts due to improvements in the ability to observe and measure hurricanes through

[2]While the forecast errors are constrained to be positive, indicators could be added to capture negative forecast errors.

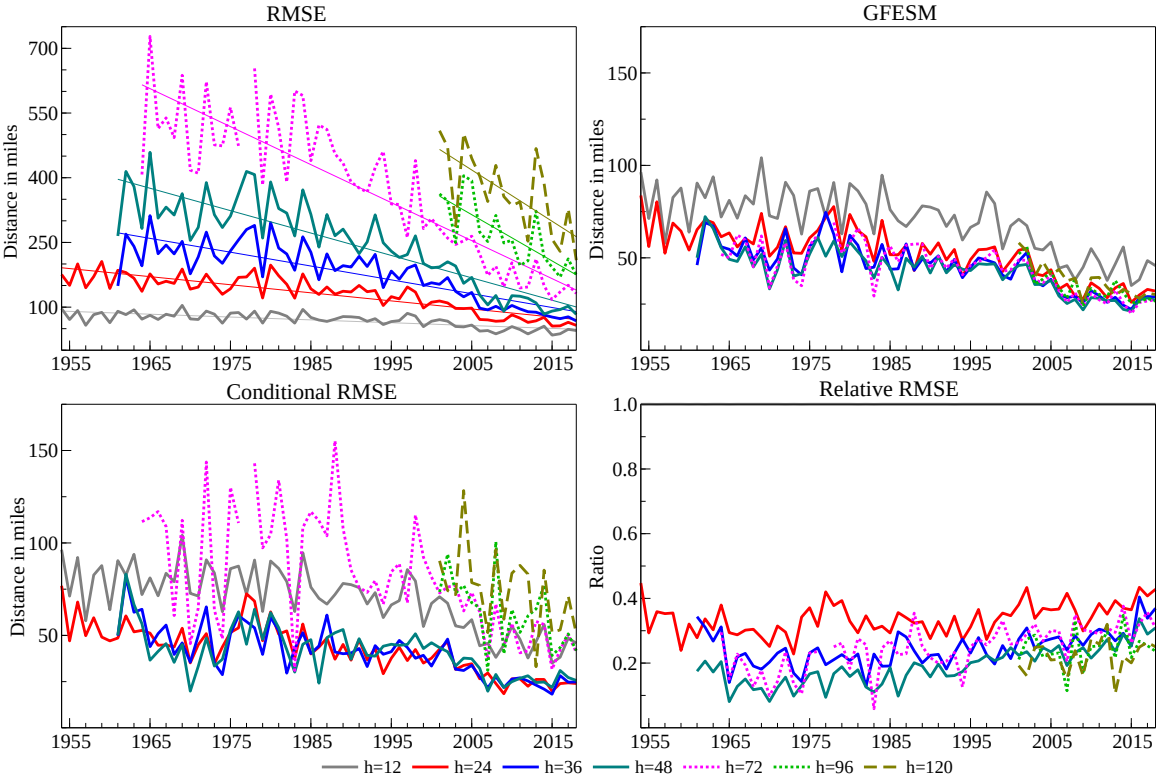
the use of satellites, as well as methods and models to analyze them using state of the art super computers; see Shuman (1989), Sheets (1990) and Rappaport et al. (2009).

We can examine changes in hurricane forecast accuracy over time by computing metrics based on the forecast errors for every tropical storm in the Atlantic since 1954. In total, we analyze the forecast errors from all NHC track forecasts for 676 storms in the Atlantic between 1954 and 2018. Looking at the evolution of several different forecast-error metrics provides insight into the different ways in which these forecasts have changed over time. The four measures that we consider are the root mean square error (RMSE), the general forecast-error second moment (GFESM), the RMSE conditional on the forecast errors at previous horizons, and the conditional RMSE relative to the overall RMSE.

Figure 4.1 presents the RMSE, the GFESM, the conditional RMSE, and the relative RMSE averaged across each storm for every year since 1954. The top left-hand panel illustrates the steady decline in the root mean square errors across all horizons over time. These trends are well documented and illustrate that the longer horizons have benefited most from these improvements. The GFESM, which measures the geometric average of forecast accuracy across multiple horizons, offers a different perspective on the evolution of the forecasts. It shows that the average forecast error for horizons beyond 12-hours-ahead are very similar after accounting for the errors at earlier horizons. This suggests a large degree of persistence in the errors across horizons so that longer-horizon forecasts benefit from improvements at shorter horizons.

The conditional RMSE in the bottom left-hand panel of Figure 4.1 reinforces this view. After conditioning on errors at earlier horizons, the trends at longer horizons are less pronounced than they were for the total RMSE. The relative RMSE in the bottom right-hand panel shows that the individual contribution to forecast errors at each horizon has increased over time as the idiosyncratic components of the forecast errors at each horizon have declined less than systemic components. This indicates that improvements in the forecasts have come predominantly from capturing systematic patterns in storm movements (cycles, trends, etc.) and less from capturing idiosyncratic movements.

Despite these large improvements in forecasting ability, hurricanes continue to cause incredible



Notes: Each year is an average of the estimated forecast accuracy for all the storms in the year. For a given storm, a forecast accuracy metric is only calculated if there are at least as many observations available as horizons. The 12-hour-ahead measure is identical for the RMSE, the GFESM and the conditional RMSE. The RMSE is computed as an average of root mean square forecast errors of all storms for a given year and horizon. The GFESM is computed as the average of the square root of the determinant of the forecast-error second-moment matrix of all storms for all horizons up to the current one in a given year. The conditional RMSE is computed as an average of the ratio of the GFESM for the current horizon divided by the GFESM for the previous horizon for all storms for a given year and horizon. The relative RMSE is computed as an average of the ratio of the conditional RMSE for a given horizon divided by the RMSE for that horizon across all storms for a given year and horizon.

Figure 4.1: Tropical Storm Forecast Performance in the Atlantic Basin since 1954

amounts of destruction. This raises questions about the value of improving these forecasts when it is hard to disentangle the potential benefits from other changes that may be driving costs. In order to answer this, it is necessary to capture the different drivers of hurricane damages along with any potential changes to forecast accuracy and uncertainty surrounding these forecasts over time.

4.1 Methods

The damage model in (3.7) encompasses many of the existing models of hurricane damages. Emanuel (2005), Nordhaus (2010) and Strobl (2011) implicitly set $\eta \equiv 0$ and $\alpha \equiv 1$ to examine the relationship between damages and natural hazards. Others set $\eta \equiv 0$ to investigate the relationship between damages and vulnerabilities; see Kellenberg and Mobarak (2008) and Geiger et al. (2016). Bakkensen

and Mendelsohn (2016) allow for $\eta \neq 0$ but implicitly assume $\hat{F}_i \equiv F_i$.

We are interested in understanding the relationship between damages and forecast accuracy and uncertainty. Instead of imposing restrictions on the other determinants, we focus on the general framework in (3.7). This allows us to test the implications of the theoretical framework while controlling for alternative variables and explaining existing results. We then use model selection to simplify the model and understand which variables are the most important drivers of damages. Our approach is broadly defined within a general-to-specific modeling framework since we start with a general model and then simplify it while using our theoretical framework as a guide.

The general-to-specific (*Gets*) modeling framework is described in detail by Campos et al. (2005) and Hendry and Doornik (2014). Recent developments, by Hendry et al. (2008), Castle et al. (2015), Hendry and Johansen (2015) and Pretis et al. (2016), illustrate its usefulness across a range of applications. *Gets* modeling provides a way to simultaneously summarize and extend the literature. This contrasts with the approach of focusing exclusively on individual determinants of damages. We describe it in the current context as follows.

We formulate a general unrestricted model (GUM), which includes all potentially relevant (theory-based or otherwise) determinants of hurricane damages. It is assumed that the residuals of the GUM are iid and follow a normal distribution.[3] It is also assumed that the GUM is potentially sparse and nests the local data generating process (LDGP). Under these assumptions, *Gets* consistently recovers the same model as if selection began from the LDGP. This helps ensure valid post-selection inference; see Chernozhukov et al. (2015). Formulation of the initial GUM is an integral part of the modeling process and requires effort to ensure that all potentially important variables are included.

Including a large number of explanatory variables can result in spurious correlations and misleading inference. *Gets* tackles this by simplifying the GUM based on the 'encompassing principle' (Mizon and Richard, 1986) so that each reduction exhibits minimal information loss based on a user-specified 'target gauge'. The target gauge plays the same role as regularization in other machine

[3]Hurricane damages approximate a log-normal distribution; see Willoughby (2012), Blackwell (2014) and Appendix Figure A.1.

learning or model selection procedures (Mullainathan and Spiess, 2017) in that it seeks to control the loss of information in the selection procedure. While regularization parameters are typically chosen empirically based on model performance, the target gauge is chosen beforehand and has a theoretical interpretation in that it determines the false-retention rate of variables in expectation, i.e. the 'gauge' (Castle et al., 2011; Johansen and Nielsen, 2016). In practice, the target gauge is set based on the number of variables being selected over, so that *on average* a single irrelevant variable is kept.

The final model is chosen so that it provides a parsimonious explanation of the GUM conditional on the acceptable amount of information loss. This approach can be used to search across many different model reduction paths in order to minimize potential path dependencies; see Hendry and Doornik (2014). If multiple models are retained, information criteria are used to select between otherwise equally valid models. Alternatively, 'thick modeling', as proposed by Granger and Jeon (2004) and discussed in a *Gets* framework by Castle (2017), can be used to pool selected models.

There are many different model selection algorithms. We use the multi-path block search algorithm known as 'Autometrics' available in PcGive; see Doornik (2009) and Doornik and Hendry (2013a). Another multi-path search algorithm is implemented using the 'gets' package in R; see Pretis et al. (2018). We assess the robustness of these results by using regularized regression methods (i.e. Lasso) as implemented in the 'glmnet' package in R; see Friedman et al. (2010).

Regardless of which selection procedure is used, there is a concern that post-selection inference is misleading since it does not capture the uncertainty in the selection procedure. The literature has proposed several solutions to this. Selection can be restricted so that it does not take place over the variables of interest (e.g. in this case forecast accuracy); see Belloni et al. (2014), Hendry and Johansen (2015), and Chernozhukov et al. (2018). Alternatively, the standard errors (Van De Geer et al., 2014) or the critical values (Berk et al., 2013) can be adjusted to capture additional selection uncertainty. Instead of adjusting the standard errors or critical values, we choose not to select over the forecast accuracy variable.

4.2 Data

Although data on hurricane strikes goes back to the 1850s, we focus on hurricane strikes in the Atlantic basin of the continental United States since 1955 for which a continuous database of hurricane forecasts exists. We start by describing the number of hurricane strikes and the sources for damages used in our analysis. Next, we describe the forecasts, the errors, and how forecast uncertainty is measured. Finally, in the remainder of this section we describe the additional variables.

The hurricane research division of the U.S. National Oceanic and Atmospheric Administration (NOAA) maintains a list of every storm with hurricane force winds to make landfall in the continental United States since 1851. 192 hurricanes made landfall in the Atlantic basin between 1900 and 2015.[4] Of these, 88 occurred between 1955 and 2015. Accounting for the fact that some hurricanes struck in multiple locations, e.g. Katrina [2005] first crossed the Florida panhandle and then moved into the Gulf and struck Louisiana several days later, there were 101 unique strikes between 1955 and 2015. We focus on 98 of these strikes after removing storms for which forecasts were not produced.

We collate damages for each strike from multiple sources. Damages are taken from annual Atlantic Hurricane Season reports (1955-2015) following Pielke Jr and Landsea (1998).[5] These are supplemented using individual tropical cyclone reports (1955-2015) and are updated using data from NOAA's hurricane research division; see Blake et al. (2011).

While there are many datasets on hurricane damages, their values are not entirely reliable. Pielke Jr and Landsea (1998) (updated and extended by Pielke Jr et al., 2008 and the ICAT database) compile damages from 1900-2012. NOAA's 'Storm Events' database, which the SHELDUS database relies on, catalogs damages associated with each storm event at the U.S. county level since 1959.[6] However, Smith and Katz (2013) find that there is a tendency to underestimate the most damaging storms.

[4]The hurricane research division maintains two lists of U.S. Atlantic landfalls. We use most up-to-date where the differences since 1955 are that it includes Helene [1958] and Ophelia [2005] but excludes Diane [1955].

[5]Reports were published in the Monthly Weather Review through 2011 and are available from the Hurricane Research Division until 2011. The National Hurricane Center maintains the annual summaries since 2012.

[6]We do not use the Sheldus database due to concerns of under reporting and because damages are allocated equally across counties for each storm. A common solution (see Davlasheridze et al., 2017) is to redistribute damages based on wind speed which would bias the analysis in favor of wind speed as an explanatory variable for damages. Addressing these issues is beyond the scope of our analysis and is left for future research.

Table 4.1: Comparing Hurricane Damages by Source

Source	Obs	Median		Std. Dev.		Min		Max		Corr.	Similar
		dif	(%)	dif	(%)	dif	(%)	dif	(%)	(%)	(%)
Pielke Jr and Landsea (1998)	52	-	-	561	28	-3,225	-95	739	100	99.6	75.0
Pielke Jr et al. (2008)	79	-	-	3,907	322	-33,291	-95	5,533	2,186	99.2	64.6
ICAT	86	-	-	4,643	52	-33,291	-95	1,159	388	99.3	66.3
NOAA: Storm Events	81	-52	-27	11,860	46	-91,736	-96	2,117	180	88.2	9.9
NOAA: Billion-Dollar	27	373	4	4,713	28	-6,899	-32	20,961	117	99.2	7.4

Notes: All external sources are expressed relative to the current dataset. Dif is calculated as external minus current damages. All values are in millions of 2017 dollars. % dif is computed by dividing dif by current damages to get a percentage difference. A positive value implies that the external source has higher damages, whereas a negative number implies that current damages are higher. Corr. is the correlation between external and current damages. Similar is the share of observations for which the absolute percentage difference is less than or equal to 1 percent.

As a result, NOAA established the 'Billion-Dollar' database which provides damages for climate and weather disasters that caused at least $1 billion in damages since 1980.

Table 4.1 shows that the current dataset is closest to the Pielke Jr et al. (2008) and ICAT datasets. However, there are important differences. Damages are revised for several hurricanes, notably Celia [1970]. There are also some hurricanes for which the damages are lower. For example, Agnes [1972] initially struck Florida as a hurricane. It then weakened and later re-intensified into a tropical storm causing damage in Pennsylvania, New Jersey and New York. We only include damages associated with the initial strike whereas other datasets include the damages for the entire storm.

The current dataset is also comparable with damages from the 'Storm Events' and 'Billion-Dollar' databases. Damages tend to be higher than the 'Storm Events' database, which suffers from under-reporting, but lower than the 'Billion-Dollar' event database. However, they typically fall within the upper and lower confidence intervals.

Hurricane Forecasts, Errors and Uncertainty

The NHC's 'official' forecasts form the basis for hurricane watches, warnings and evacuation orders and are widely distributed to, and used by, news outlets. The forecasts are not based on a single model and should not be considered the same across storms. In fact, they are a combination of many different models and forecaster judgment; see Broad et al. (2007).

The forecasts are produced every 6 hours for the entire history of a storm. However, not all time

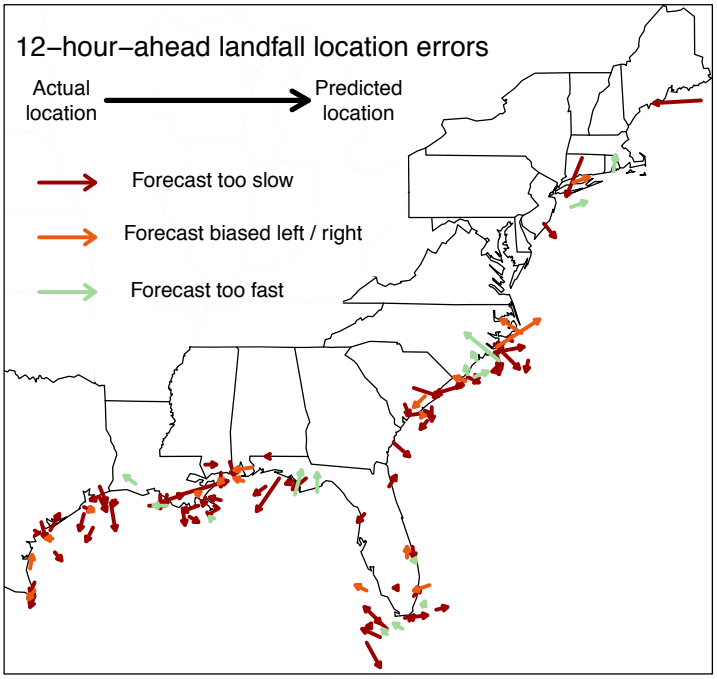
periods are relevant for our analysis. This is especially true since hurricanes can be active for up to a month (e.g. Ginger [1971]) and transect the entire Atlantic ocean. In order to focus on the most relevant forecast for damages, we relabel each forecast in terms of the number of hours it was made before landfall. We start by rounding the timing of each landfall to the closest point in a 6-hour window. Thus, if a hurricane made landfall at 16:00 UTC then it is rounded to 18:00 UTC. Next, we subtract the length of the forecast horizon (h) from the landfall time to get the time at which the h-hour-ahead 'landfall forecast' was generated. So for example, the 12-hour-ahead landfall forecast of a storm that made landfall at 18:00 UTC was generated at 6:00 UTC.

We focus on the 12-hour-ahead landfall track forecasts for several reasons. First, the U.S. Army Corps of Engineers (2004), Milch et al. (2018) and Beatty et al. (2019) find that individuals often wait until the last minute and focus on the track. Second, the short-horizon track forecasts are available for virtually every U.S. hurricane strike going back to 1955, whereas other forecasts are not.[7] Third, the track is an integral part of the NHC's forecasts of hurricane intensity including rainfall (Kidder et al., 2005), wind speed (DeMaria et al., 2009), and storm surges (Resio et al., 2017). Thus, there is strong support for focusing exclusively on the 12-hour-ahead track forecasts.

Track forecast errors are computed differently from typical forecast errors. We calculate them as the distance between two points on the surface of a spheroid (Vincenty, 1975) to account for the curvature of the earth.[8] Thus, they are purely distance measures (i.e. absolute errors) and do not have a directional interpretation. This is problematic if the direction matters for damages. Forecasts that are too slow or biased to either side of the landfall location provide less warning time. Alternatively, forecasts that are too fast may underestimate the amount of rainfall and flooding and make individuals less likely to prepare for these damages. Despite these differences, each forecast error direction is associated with the belief that the hurricane will be less destructive than it ultimately is. Figure 4.2 plots the spatial distribution of the landfall locations and forecast error directions.

[7] No forecasts are available for Debra [1959] and Ethel [1960], likely due to their short duration. Longer horizon forecasts are available for even fewer storms and the intensity forecasts are only available since 1990.

[8] This calculation is more accurate than the more commonly used great-circle distance.



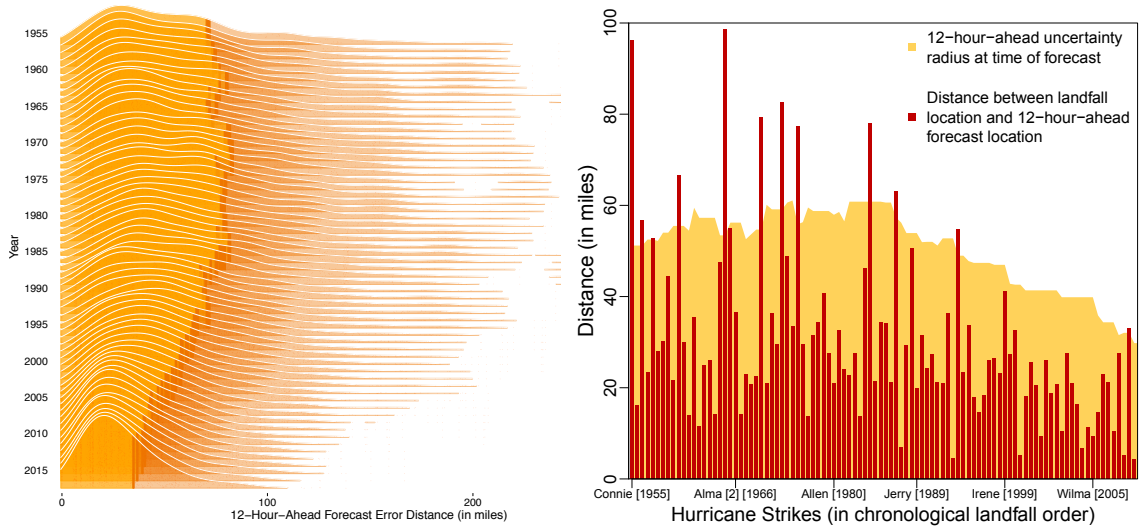
Notes: Coloring is based on the calculated angle of a triangle which is is calculated from the distance between the forecast and the actual (shown), the distance between the location 12 hours prior and the current location and the distance between the previous location and the forecast. Dark red is less than 90 degrees, orange is greater than 90 but less than 135 degrees and pale green is greater than 135 degrees.

Figure 4.2: Direction of 12-Hour-Ahead Landfall Errors in Space

We can evaluate the track forecast errors relative to the ex-ante risk associated with them at the time. To do this we estimate the time-varying historical forecast error densities since 1955 using every forecast error for all tropical storms at the 12-hour horizon in the previous five years. As Panel A of Figure 4.3 shows, there have been large changes in the error densities since 1955. Historically the densities have been skewed to the right with long tails driven by extreme outliers. However, since the 1990's, the skewness has declined and a normal distribution seems to be a good approximation. The 67th percentile corresponds directly with how the NHC computes the radius of the 'cone of uncertainty', which has accompanied every hurricane track forecast since 2002.[9]

Panel B of Figure 4.3 plots the 12-hour-ahead landfall forecast errors along with our reconstructed measure of the 12–hour-ahead radius of uncertainty going back to 1955. The figure illustrates that landfall forecast errors have declined by around 60 percent over the past 60 years. The implied ex-ante risk has also declined over the same period so that both the errors and the expected risk associated

[9]Since the forecast database only extends back to 1954, the radius for storms prior to 1959 is estimated using samples shorter than five years. Historically, the NHC measured the radius as the distance that captured all errors in the previous ten years. For more information see Broad et al. (2007).



Panel A: 12-Hour-Ahead Forecast Error Densities by Year

Panel B: 12-Hour-Ahead Landfall Errors Over Time

Notes: Errors are computed using Vincenty (1975)'s formula for the distance between two points on the surface of a spheroid. In panel A, each density is estimated using all 12-hour forecast errors for all tropical storms in the Atlantic Basin over the previous five years based on between 100-2000 observations. The dark orange bars represent the 67th percentile for each year. In panel B, the 12-hour-ahead landfall errors are computed such that the forecast was made 12 hours before the closest observed point of the hurricane at landfall. The shaded area is the implied radius of uncertainty and corresponds to the 67th percentile of Panel A for the year in which the strike occurred.

Figure 4.3: 12-Hour-Ahead Hurricane Landfall Track Errors

with the forecasts has declined. The figure also shows that the forecast errors at landfall exceeded their ex-ante risk in about 12 percent of all strikes; most recently for Sandy [2012].

Other Variables

While the focus of our analysis is on the link between damages and forecast uncertainty, we embed this relationship within a general model of damages. This ensures that we have a well specified model of hurricane damages. It also reduces the possibility that the relationship between forecast accuracy and damages is confounded by omitted variable bias.

We include several measures of vulnerabilities. We use county level population and personal income estimates from the Bureau of Economic Analysis (BEA) since 1969. Prior to 1969, county level population estimates are from the U.S. Census, available for each decade, along with state level population and personal income estimates available annually from the BEA. We compute county level population values prior to 1969 by interpolating county level population shares between decades and then distribute them using state level data. We do the same for land area and housing units.[10]

[10]Pielke Jr et al. (2008) use a similar approach. We aggregate counties using BEA's modifications to Census codes.

Table 4.2: Data Descriptions and Summary Statistics

Variable	Description	Years	Min	Average	Max	Source
		Damages (D)				
DAMAGE	Nominal Damage (U.S. $1,000)	1955-2015	$28	$3,990,506	$105,900,000	NOAA, NHC
		Socio-economic Vulnerabilities (V)				
PD	Population Density (persons per acre)	1955-2015	12	257	3,940	BEA, Census
IP	Income Per Capita ($ per person)	1955-2015	$864	$16,430	$60,213	BEA
HD	Housing Unit Density (houses per acre)	1955-2015	5	104	1,672	Census
IH	Income Per Housing Unit ($1,000 per unit)	1955-2015	$2	$37	$140	BEA, Census
FREQ	Historical Hurricane Frequency (Average per year)	1955-2015	0.01	0.09	0.32	NOAA, HRD
LEVEE	Levee Length Density (miles per acre)	1955-2015	0	0.03	0.22	USACE, NLD
CRS	FEMA Community Rating System (rank)	1990-2015	7	9	10	FEMA
HMGP	Hazard Mitigation Grant Program (U.S. $1,000)	1990-2015	0	$38,042	$396,102	FEMA
		Natural Forces (F)				
WIND	Max Sustained Wind Speed (kt)	1955-2015	65	90.3	150	NOAA, HRD
PRESS	Central Pressure at Landfall (mb)	1955-2015	909	965	1003	NOAA, HRD
RAIN	Max Rainfall (in)	1955-2015	4.8	13.75	38.5	NOAA, WPC
SURGE	Max Surge (ft)	1955-2015	0	8.5	27.8	NOAA, NHC
ACE	Accumulated Cyclone Energy (Seasonal)	1955-2015	17	135	250	NOAA, HRD
MOIST	Deviations from trend soil moisture (in)	1955-2015	-4.75	1	5.7	NOAA, ESRL
GST	Land, Air and Sea-Surface Temp. index	1955-2015	0.1	0.34	0.93	NASA, GISS
		Forecast Uncertainty (U)				
FORC12	12-Hour Official Track Error (nautical miles)	1955-2015	5	34	114	NOAA, NHC
RADII12	Implied 12-hour radius of uncertainty (nautical miles)	1955-2015	34	70	59	NOAA, NHC
NAIVE12	12-Hour Naïve Track Error (nautical miles)	1970-2015	5	31	97	NOAA, NHC
SKILL12	Ratio of 12-Hour naïve forecast error to FORC12	1970-2015	0.19	1.44	10.35	NOAA, NHC
WARN	Warning time over coast length (100 hours per mile)	1955-2015	0.7	11.33	45.78	NOAA, NHC

Notes: NOAA: National Oceanic and Atmospheric Administration; NHC: National Hurricane Center; HRD: Hurricane Research Division; WPC: Weather Prediction Center; ESRL: Earth System Research Laboratory; NASA: National Aeronautics and Space Administration; GISS: Goddard Institute of Space Studies; BEA: Bureau of Economic Analysis; Census: Census Bureau; FEMA: Federal Emergency Management Agency; USACE: US Army Corps of Engineers; NLD: National Levee Database.

Annual county level personal income prior to 1969 is estimated as follows. First, we assume that county level personal income shares were constant from 1955 to 1969. Second, we estimate annual income shares using a fixed effects panel data model and then starting in 1969, backcast the shares to 1955.[11] We combine the shares with state level income to get a county level estimate. We average these two approaches to get a robust measure.

We compute a real-time measure of historical strike frequency using county level hurricane strikes since 1900. The strike frequency for a county in a given year is computed over time by taking the number of hurricanes that struck that county since 1900 divided by the number of years that have passed. Strike level historical frequency is computed as an average of the strike frequencies of all counties struck by the hurricane at the time of the strike.

[11] The panel data model was estimated for all U.S. counties from 1969 to 1999 using leads of income shares and population shares as explanatory variables.

Since damages are measured at the strike level, we aggregate across impacted counties. This alleviates concerns about county level estimates but requires us to choose which counties are impacted. We focus on coastal counties (Jarrell et al., 1992), which over-weights the importance of coastal conditions but is less likely to over-weight the impact of wind damages, as in the approaches used by Strobl (2011), Hsiang and Narita (2012), and Deryugina (2017).

Next, we include measures of natural and climate forces. The maximum sustained (1-minute) surface (10 meter) wind speed, minimum central pressure at landfall, maximum storm surge height and accumulated seasonal cyclone energy are obtained from the NHC. Maximum rainfall comes from NOAA's Weather Prediction Center.

Monthly soil moisture estimates, derived using methods devised by van Den Dool et al. (2003), are obtained from NOAA's Earth System Research Laboratory. We link estimates in the nearest grid point to a county's centroid. County estimates are averaged across impact counties for each hurricane strike and then the smoothed estimate for the month prior to the strike is used.[12]

Storm-level estimates of sea surface air temperature are computed following Estrada et al. (2015) using data from NASA's global mean surface temperature index based on land-surface air temperature anomalies. The monthly series is smoothed using the Hodrick-Prescott filter with a smoothing parameter equal to 129,600 and the estimate for the month prior to the hurricane strike is used. Table 4.2 and Appendix Figure A.2 provide sources, summary statistics, and plots of the variables.

5 A Model of Hurricane Damages

This section presents the general hurricane damages model followed by its estimation and reduction. The model includes all major determinants of hurricane damages and several controls for spatial and temporal heterogeneity. It contains 37 explanatory variables and is estimated over a sample of 98

[12] We use the Hodrick-Prescott filter with a smoothing parameter equal to 129,600 following Ravn and Uhlig (2002).

observations. Lower case variables are in logs:

$$\begin{aligned} \text{damage}_i &= c + \alpha_1\ \text{hd}_i + \alpha_2\ \text{ih}_i + \alpha_3\ \text{FREQ}_i + \beta_1\ \text{rain}_i + \beta_2\ \text{surge}_i + \beta_3\ \text{npress}_i \\ &+ \beta_4\ \text{wind}_i + \beta_5\ \text{MOIST}_i + \beta_6\ \text{ace}_i + \beta_7\ \text{GST}_i + \eta_1\ \text{forc12}_i + \eta_2\ \text{radii12}_i \\ &+ \delta_1\ \text{STREND}_i + \delta_2\ \text{YTREND}_i + \psi\ \text{MONTH}_i + \kappa\ \text{HOUR}_i + \lambda\ \text{STATE}_i + \epsilon_i. \end{aligned} \tag{5.1}$$

The first line includes the vulnerabilities (V): housing unit density (hd), income per housing unit (ih) and 'real-time' hurricane strike location frequency (FREQ). We exclude population density since it is almost perfectly correlated with housing density and because housing has a more direct interpretation in this context.[13] The first two lines list the natural hazards (F): maximum rainfall (rain), storm surge (surge), negative minimum pressure (npress), maximum wind speed (wind), soil moisture relative to trend (MOIST), accumulated cyclone energy (ace) and global surface temperature (GST). The second line captures forecast accuracy and uncertainty (U): 12-hour-ahead forecast track errors (forc12) and the implied 12-hour-ahead radius of uncertainty (radii12). The last line lists additional spatial and temporal controls: strike and annual trends, month dummies, hour dummies (to control for the six-hour period in which the landfall occurred), and U.S. state dummy variables.

We estimate (5.1) using ordinary least squares (OLS). The estimated coefficients and their standard errors are reported in column (1) of Table 5.1. Several coefficients are statistically significantly different from zero. They include housing density, historical hurricane frequency, storm surge, central pressure, and the forecast errors. The coefficient on the forecast errors is positive, which is consistent with the prediction from the theoretical framework in (3.4) and (3.6). This supports the false sense of security hypothesis that larger forecast errors are associated with higher damages even after controlling for ex-ante uncertainty, vulnerabilities, and natural hazards.

Although the significant variables in the full model provide an initial sense of the most important drivers of hurricane damages, model selection provides a more systematic approach. Rather than impose restrictions to ensure that selection is consistent with our theoretical model, we start by se-

[13] The results are similar when population is included.

lecting over all variables without any restrictions. Since the full model (i.e. the GUM) has 38 unique parameters (i.e. 37 variables plus the variance), then we set the target gauge equal to $\frac{1}{37} \approx 0.03$. We can adjust this target to understand how sensitive the results are to this choice. This helps address concerns that model selection is unstable (Mullainathan and Spiess, 2017).

Table 5.1: Damage Models by Selection Method

	(1) OLS	*Gets*			(5) Lasso OLS	(6) Double Lasso
Selection Target:		(2) 1%	(3) 3%	(4) 5%	BIC (9%)	BIC (9%)
Housing density	0.53**		0.40***	0.40***	0.45***	0.40**
	(0.24)		(0.15)	(0.15)	(0.15)	(0.16)
Income per housing unit	0.70	1.58***	1.28***	1.28***	1.05***	1.35***
	(0.79)	(0.14)	(0.18)	(0.18)	(0.17)	(0.47)
Historical fequency	-6.63**				-3.72**	-3.13
	(2.94)				(2.23)	(2.22)
Max rainfall	0.51	1.18***	1.14***	1.14***	0.74**	0.89**
	(0.43)	(0.32)	(0.31)	(0.31)	(0.32)	(0.32)
Max storm surge	1.21**	1.37***	1.34***	1.34***	1.30**	1.39***
	(0.48)	(0.39)	(0.37)	(0.37)	(0.37)	(0.38)
Min central pressure (-)	52.6***	52.9***	52.0***	52.0***	51.1***	49.0***
	(18.2)	(8.72)	(8.43)	(8.43)	(8.46)	(8.44)
Max wind speed	-0.27					
	(1.54)					
Soil moisture	0.77					
	(1.42)					
Seasonal cyclone energy	0.42				0.43*	0.59**
	(0.32)				(0.23)	(0.25)
Sea surface temperature	0.94					-0.25
	(3.21)					(1.68)
12-hour forecast errors	0.54*	0.55**	0.48**	0.48**		0.50**
	(0.29)	(0.24)	(0.23)	(0.23)		(0.24)
12-hour radius	3.17					0.82
	(2.60)					(1.37)
Trends:	Yes	No	No	No	No	No
Hour fixed effects:	Yes	No	No	No	No	No
Month fixed effects:	Yes	No	No	No	No	No
U.S. state fixed effects:	Yes	No	No	No	Yes	Yes
k	37	5	6	6	8	11
$\hat{\sigma}$	1.300	1.304	1.261	1.261	1.250	1.228
R^2	0.876	0.806	0.821	0.821	0.828	0.840
BIC	4.710	3.658	3.626	3.626	3.680	3.749

Notes: Estimated using 98 observations including a constant and dummy variables for Gerda [1969] and Floyd [1987]. Standard errors are in parentheses. k is the number of selected regressors in the model. *p< 0.1 **p< 0.05 ***p< 0.01

There are 2^{37} (> 130 billion) possible model combinations if we allow for every variable to be selected over. For a target gauge of 3 percent, the selection algorithm narrows the search space to 2^{16} (< 70 thousand). In the process it eliminates entire branches of possible models and only estimates 335 candidate models. The algorithm finds that 9 terminal models are acceptable reductions of the GUM. The final model is selected from these terminal models using the Bayesian information criterion (BIC). It is also robust to alternative information criteria (see Appendix Table A.2).

Columns (2)-(4) in Table 5.1 present the selection results across a range of target gauges. Although the selected models are virtually identical, this masks large variation as the number of terminal models for each target range from $9-18$. In total, 48 unique terminal models are found across the different targets. Almost three quarters of these include some measure of forecast uncertainty, while almost 60 percent include the forecast errors themselves.

The selected variables are housing and income, (-) central pressure, rainfall, storm surge, and the forecast errors. This is consistent with (3.7) in that the selected model has at least one measure of vulnerabilities (V), natural hazards (F), and forecast uncertainty (U) and each of their signs are in the expected direction. Selection is identical across different target gauges with the exception of housing density, which is not retained when the target gauge is set to 1 percent. This is in line with the full model except that rainfall and income are selected but not historical hurricane frequency.

Wind speed, which is a common measure of natural hazards, does not appear in any of the selected models.[14] Instead, minimum central pressure is always included. This is supported by Bakkensen and Mendelsohn (2016), who find that central pressure provides a more reliable explanation of damages. In addition to central pressure, we also find that rainfall and storm surge are empirically relevant. Rainfall leads to inland flooding damages whereas storm surge causes damages along the coast. Thus, we find a more disaggregated measure of natural hazards than previous models.

The results are robust across alternative selection methods. We can perform model selection using Lasso, which shrinks the coefficients with a penalty (chosen here using BIC). We present the post-selection OLS coefficients (and standard errors), shown in column (5) of Table 5.1, to compare with the *Gets* results. More variables are retained, which is consistent with the fact that BIC corresponds roughly to a target gauge of 9% for the sample size and number of variables selected over; see Campos et al. (2003). The forecast errors are not retained, but this is very sensitive to the choice of the regularization hyper-parameter.[15]

[14] Even when it appears in the full GUM or one of the reduced GUMs, its coefficient has the wrong sign and is not significantly different from zero; see Appendix Table A.3.

[15] The forecast errors are included if cross-validation is used to determine the hyper-parameter.

In order to perform post-selection inference on the forecast errors we need to make sure that we account for any selection uncertainty. One way to do this in the context of *Gets* is to force the forecast errors into the model and select over all other regressors; see Hendry and Johansen (2015). The results for this approach are identical to column (3). Alternatively, assuming that the forecast errors are exogenous, conditional on the other regressors, then we can think of the forecast errors as being a treatment variable for damages. If so, then we can perform 'Double Lasso' selection proposed by Belloni et al. (2014). This procedure is described in three steps. First, a Lasso regression is run on the full model of damages excluding the forecast errors. Second, another Lasso regression is run on a model of the forecast errors using all other explanatory variables. Third, the variables selected in the first two steps along with forecast errors are combined into a final OLS regression on damages. This allows for valid inference on the coefficient of the forecast errors as long as the underlying process is sparse. The results of this procedure are shown in the final column of Table 5.1 where the coefficient and standard errors corresponding to the landfall location forecast error are broadly consistent with OLS and the *Gets* model selection results.

Overall the results indicate that a small subset of drivers explain much of the variation in hurricane damages. They provide further support for the false sense of security hypothesis in that the coefficient estimate on forecast errors variable is positive and significantly different from zero. This is in line with what the theoretical framework in (3.4) would predict. Furthermore, decomposing forecast uncertainty into forecast accuracy and ex-ante risk as in (3.12), illustrates that forecast accuracy matters but that uncertainty around the forecast does not.

6 Robustness of the Results

This section evaluates the robustness of the results. We start by checking for possible model misspecification. Next, we address concerns about omitted variable bias. Finally, we assess out-of-sample fit. Overall, the robustness checks confirm that forecast accuracy matters for damages.

6.1 Model Misspecification

Interpretation of the coefficients and their significance depends on whether the underlying assumptions about the model are satisfied. However, rerunning our analysis with a battery of diagnostic tests indicates that these assumptions may not be satisfied. The diagnostic tests shown in Table 6.1 are: the $\chi^2_{nd}(2)$ test for non-Normality (Doornik and Hansen, 2008), the F_{Het} / F_{Het-X} test for residual heteroskedasticity (with and without cross products; White, 1980), and the $F_{RESET23}$ test for incorrect model specification (Ramsey, 1969).[16] A rejection of the null hypothesis indicates that the assumption associated with that test is invalid. The diagnostic tests in column (1) indicate that the selected models have evidence of non-normal and heteroskedastic residuals. This provides evidence against the assumptions about the functional form and the log-linearity approximation.

Model misspecification can be dealt with in several ways. Heteroskedasticity of the residuals can be addressed by correcting the standard errors following Newey and West (1987). The model can also be extended by dropping outliers that may induce non-normality and adding squares of the explanatory variables to help capture possible nonlinearities that induce heteroskedasticity. These choices entail different trade-offs. Correcting the standard errors ensures consistent estimates without changing the model but does not address any underlying misspecification. Extending the model addresses misspecification but can make estimates less reliable if added variables are irrelevant.

Although the model can be expanded in different ways, we follow the approach advocated by Hendry and Johansen (2015). We embed the selected model into a more general model that includes impulse dummies for every observation and squares of the variables. Since there are more variables than observations it is not feasible to estimate the full GUM. Selection over the impulse dummies and nonlinearities is done by exploiting the algorithm's ability to search across multiple paths in subblocks. This is known as impulse indicator saturation (IIS); see Hendry et al. (2008) and Johansen and Nielsen (2016).[17] Since we have embedded our initially selected model, any further selection

[16] Since the data is irregularly spaced they are not really time series in a strict sense and so we do not report diagnostic tests for residual autocorrelation and time-varying variances.

[17] It is equivalent in this context to selecting over individual strike fixed effects.

we do will favor that model. However, doing so allows us to assess the robustness of the original selection results to outliers and non-linear terms.

Castle et al. (2018) advocate for expanding and then selecting over the original GUM. In practice, this requires a tighter target to control the number of impulses retained and a looser target to capture marginally relevant variables. Hendry (2018) uses a two-step procedure where selection is done first over the impulse dummies with a tight target gauge and then over the entire model jointly with a looser target. This ensures that not too many impulses are retained without limiting the search space.

The selected model is broadly robust to misspecification. The heteroskedasticity corrected standard errors in column (2) of Table 6.1 do not indicate major changes in the significance of the coefficients. Extending the model produces similar results despite retaining the square of income and several impulse dummy variables; see column (3).[18] The nonlinear relationship between income and damages is in line with Geiger et al. (2016). The outlying observations capture several issues including measurement concerns in the late 1950s and the glancing landfalls by Alex [2004] and Arthur [2014]. No further diagnostic concerns are indicated when both impulses and nonlinearities are selected over jointly.[19] Importantly, the standard errors of the coefficients in column (3) are smaller than in column (1), which indicates that estimates are more precise despite including more covariates.

Since the coefficients of the outlying storms are all negative and roughly the same magnitude, we can test whether they are of equal magnitude. Under the null hypothesis of equal magnitude, the likelihood ratio has a statistic of 5.76 which we compare against a χ^2-distribution with 6 degrees of freedom. The Wald test has a statistic of 0.96 which we compare against the F-distribution with $(6,82)$ degrees of freedom. In both cases we fail to reject the null hypothesis using any reasonable critical value. A single dummy variable for all outlying storms gives a more parsimonious model, column (4), with less uncertainty around the estimated coefficients; see Hendry and Santos (2005). It may also be possible to extend the model using a measure of the maximum distance traveled inland.

[18] Selection is done using a target gauge of 1 percent but the results are identical when using a target gauge as tight as 0.01 percent.

[19] The normality test should be regarded with caution in this context; see Berenguer Rico and Nielsen (2017).

Table 6.1: Robust Models of Hurricane Damages

	(1) *Gets* (3%)	(2) Gets+HCSE	(3) Extension (1%)	(4) Robust
Housing density	0.40***	0.40***	0.29***	0.27***
	(0.15)	(0.15)	(0.10)	(0.10)
Income per housing unit	1.28***	1.28***	1.40***	1.43***
	(0.18)	(0.21)	(0.13)	(0.15)
Min central pressure (-)	52.0***	52.0***	55.5***	56.6***
	(8.43)	(7.98)	(5.66)	(5.49)
Max rainfall	1.14***	1.14***	0.52**	0.57**
	(0.31)	(0.33)	(0.21)	(0.21)
Max storm surge	1.34***	1.34***	0.93***	0.99***
	(0.37)	(0.38)	(0.25)	(0.24)
12-hour forecast errors	0.48**	0.48*	0.30*	0.34**
	(0.23)	(0.27)	(0.16)	(0.15)
Income per housing unit sq.			0.45***	0.44***
			(0.09)	(0.09)
Outlying storms:				-3.44***
				(0.33)
Helene [1958]			-2.70***	
			(0.86)	
Cindy [1959]			-4.22***	
			(0.86)	
Gracie [1959]			-2.73***	
			(0.84)	
Alma [1] [1966]			-4.41***	
			(0.84)	
Bret [1999]			-2.48***	
			(0.87)	
Alex [2004]			-3.56***	
			(0.83)	
Arthur [2014]			-4.17***	
			(0.88)	
$\hat{\sigma}$	1.261	1.261	0.816	0.812
R^2	0.821	0.821	0.932	0.927
$\chi^2_{nd}(2)$	8.34**	8.34**	0.91	0.03
	[0.015]	[0.015]	[0.636]	[0.986]
F_{Het}	1.76*	1.76*	0.88	0.80
	[0.069]	[0.069]	[0.585]	[0.675]
F_{Het-X}	1.02	1.02	0.97	0.94
	[0.457]	[0.457]	[0.532]	[0.570]
$F_{RESET23}$	1.30	1.30	1.35	1.09
	[0.279]	[0.279]	[0.265]	[0.341]

Notes: All equations are estimated using 98 observations and include a constant and dummy variables for Gerda [1969] and Floyd [1987]. The standard errors are in parentheses. The tail probability associated with the null hypothesis of each diagnostic test statistic is in square brackets. Column (2) shows the heteroskedasticity corrected standard errors. Income per housing unit is demeaned to facilitate interpretability of the coefficients. *p< 0.1 **p< 0.05 ***p< 0.01

The coefficient on forecast errors remains positive and significant even after accounting for mis-specification. The estimate varies between $0.30 - 0.52$ for alternative specifications. Similar results are obtained if we first select over impulses while forcing the GUM and then select over the full model or if we augment the full model with the joint outlier dummy variable and the square of income and then select.[20] It is also robust to alternative measures of damages (Appendix Table A.4).

[20]One difference is the inclusion of historical hurricane frequency, which remains significant until outliers are accounted for but does not alter the results.

6.2 Alternative Measures of Forecast Accuracy and Uncertainty

We can also evaluate how sensitive the results are to alternative measures of forecast accuracy and uncertainty. Table 6.2 presents the results. Column (1) shows the robust model with forecast errors. Column (2) augments the model with the radius of uncertainty as in (3.12). Column (3) imposes a 1-to-1 relationship between the errors and the radius of uncertainty as in (3.11). We weakly reject that this 1-to-1 relationship is consistent with the data with a p-value of 1.14 percent. This indicates that the forecast errors and ex-ante risk are capturing different aspects of information.

We can also check if the implicit assumption about the distribution of the forecast error density matters. Column (4) applies Rossi and Sekhposyan (2015)'s measure of uncertainty in (3.8) allowing for a time-varying empirical forecast-error distribution using an estimation window based on of the previous five years of forecast errors. The results in columns (3) and (4) are similar, which suggests that the normal distribution approximation does not have a large impact. Finally, column (5) considers a weighted RMSE measure which reduces the impact of any given forecast error by also giving weight to errors made after the initial strike. This measure is more comparable to the forecast errors themselves although it has a slightly larger coefficient but is less significant which suggests that including other errors increases the noise around the estimate.

Overall, the results in Table 6.2 illustrate that there is a positive and statistically significant relationship between forecast accuracy and damages across the different measures. The size of this relationship is also fairly stable. In addition to this, uncertainty around the forecast has a positive and statistically significant relationship with damages after controlling for outliers. One interpretation for this is that it captures technological change. In that sense we can think of forecast accuracy capturing short-term adaptation efforts while uncertainty around the forecast captures long-term adaptation efforts. Under this interpretation, the restriction imposed in (3.11) assesses the relative importance of alternative measures of adaptation. The results in columns (2) and (3) of Table 6.2 indicate that short-term measures matter even after controlling for longer-term adaptation efforts.

Table 6.2: Alternative Measures of Forecast Accuracy and Uncertainty

	(1) Errors	(2) Errors & Radius	(3) Errors/Radius	(4) R&S(2015)	(5) wRMSE
Housing density	0.27***	0.24**	0.28***	0.28***	0.26***
	(0.10)	(0.09)	(0.10)	(0.10)	(0.10)
Income per housing unit	1.43***	1.76***	1.36***	1.36***	1.44****
	(0.12)	(0.19)	(0.12)	(0.12)	(0.13)
Income per housing unit sq.	0.44***	0.66***	0.41***	0.41***	0.45***
	(0.09)	(0.13)	(0.09)	(0.09)	(0.09)
Min central pressure (-)	56.6***	57.8***	56.8***	56.9***	57.3***
	(5.49)	(5.41)	(5.55)	(5.58)	(5.51)
Max rainfall	0.57***	0.56***	0.55***	0.54**	0.56***
	(0.21)	(0.20)	(0.21)	(0.21)	(0.21)
Max storm surge	0.99***	0.98***	0.97***	0.97***	0.95***
	(0.24)	(0.24)	(0.25)	(0.25)	(0.24)
12-hour forecast errors	0.34**	0.27*			
	(0.15)	(0.15)			
12-hour error/radius			0.29*		
			(0.16)		
12-hour uncertainty				0.23*	
				(0.14)	
12-hour wRMSE					0.37*
					(0.19)
12-hour radius		2.10**			
		(0.96)			
Outlying storms dummy	-3.44***	-3.19***	-3.49***	-3.50***	-3.43***
	(0.33)	(0.35)	(0.34)	(0.34)	(0.34)
$\hat{\sigma}$	0.812	0.795	0.819	0.822	0.817
R^2	0.927	0.930	0.925	0.925	0.926

Notes: All equations are estimated using 98 observations and include a constant and a dummy variable for Gerda [1969] and Floyd [1987]. Standard errors are in parentheses. R&S(2015): Rossi and Sekhposyan (2015). Weight for RMSE based on AR(1) with a lead coefficient of 0.3. *p< 0.1 **p< 0.05 ***p< 0.01

6.3 Controlling for Potentially Omitted Variables

So far we have assumed that the forecast errors are at least weakly exogenous for hurricane damages. This is a safe assumption if the forecast errors are randomly determined (conditional on the other covariates) and unrelated to other potentially relevant variables that are excluded from the analysis. However, this may not be the case. For example, there are non-random changes in the forecast errors over time. To fully assess the robustness of the results, we need to account for the possibility that omitted variables could bias the relationship between damages and the forecast errors.

The forecast errors may be correlated with storm dynamics. A volatile storm is more difficult to forecast and can cause more damage. For example, the rapid intensification of a hurricane just prior to landfall is difficult to forecast and is also associated with higher damages; see Kaplan et al. (2010). Furthermore, a slow moving storm with high amounts of rainfall can also be forecast poorly. For example, Harvey [2017] experienced rapid intensification in its initial buildup, slowed down as

it made landfall, and dropped record breaking amounts of rain in its wake. In either case, storm dynamics are associated with increased forecast errors and higher damages. Thus, the relationship between forecast errors and damages may be moderated by the storm's dynamics.

Forecast errors may also be correlated with longer-term adaptation and technological change. Adaptation often goes hand-in-hand with efforts to improve hurricane forecasts and could therefore be correlated with improved forecasts. There is evidence of this in column (2) of Table 6.2 where the unrestricted ex-ante risk is positively correlated with hurricane damages. Since ex-ante risk is effectively a five-year moving average of past forecast errors, changes in it are driven less by short-term randomness and more by longer-term trends such as government expenditures and technological advancements. Therefore, the relationship between forecast errors and damages may also be moderated by these longer-term changes.

Given the existence of non-random relationships between the forecast errors and variables excluded from the model, it is important to assess whether controlling for them alters the results. While there is no concise measure of storm dynamics, we construct an instrument based on a measure of forecast skill, which is regularly used to assess the performance of hurricane and weather forecasts; see Cangialosi and Franklin (2016). We measure forecast skill as the ratio of the official forecast errors to naïve forecast errors from a simple climatology and persistence model. Since naïve forecasts suffer from the same natural variability as the official forecast, this measure should remove the effect of storm dynamics from the official forecast errors. Naïve hurricane forecasts are available for all hurricanes starting in 1970.

We also control for hurricane warning lead times to proxy for hurricane evacuations. Hurricane warnings and watches are among the most prominent outputs for hurricane forecasts and are a crucial part of the U.S. government's evacuation system. A hurricane warning indicates that hurricane force winds are expected within the next 36 hours (historically 24). However, since warnings are coordinated with local authorities, they also take into account how much time a given area would need to be evacuated. As a result, earlier warnings are given for stronger storms or when they are expected to

strike more populated locations. Furthermore, earlier warning times can cover larger areas due to the uncertainty associated with the path of the hurricane. We address this issue by looking at the number of hours a warning was issued before a strike divided by the length of coastline under warning.[21]

It is more difficult to control for longer-term adaptation efforts. We include the normalized length of protective levees from the US Army Corps of Engineers' National Levee Database as a measure of how location-specific efforts have changed over time. We also include the radius of uncertainty as a general proxy of longer-term technological improvement. The maximum processing speed of NOAA's supercomputers was also considered but does not seem to matter. Together, these measures should control for longer-term adaptation efforts.

Alternative measures of adaptation include the U.S. Federal Emergency Management Agency's Community Rating System (CRS) as well as its Hazard Mitigation Grant Program (HMGP). The CRS was created in 1990 as a part of the National Flood Insurance Program to incentivize flood damage mitigation through reductions in flood insurance premiums. The HMGP was established in 1988 and includes grant funding for damage mitigation efforts, including improved warning systems; see Davlasheridze et al. (2017). This analysis can be found in Appendix Table A.5.

Table 6.3 shows that we continue to find a significant relationship between errors and damages with an elasticity around 0.5. This suggests that the findings are robust to concerns about omitted variables. However, the results do not hold for longer forecast horizons. Although the direction of the relationship remains consistent with our hypothesis, the estimated coefficient is smaller and not statistically significantly different from zero. This suggests that less attention is paid to longer forecast horizons which is consistent with the survey findings in Milch et al. (2018).

[21] This is a crude measure since some hurricanes affect a large area and so a wider warning area is a good thing. Appendix Figure A.3 shows that despite increasing hurricane warning lead times over time, the ratio of the hurricane warning lead time to the area covered under that initial warning is fairly stable.

Table 6.3: Controlling for Storm Dynamics and Adaptation Efforts

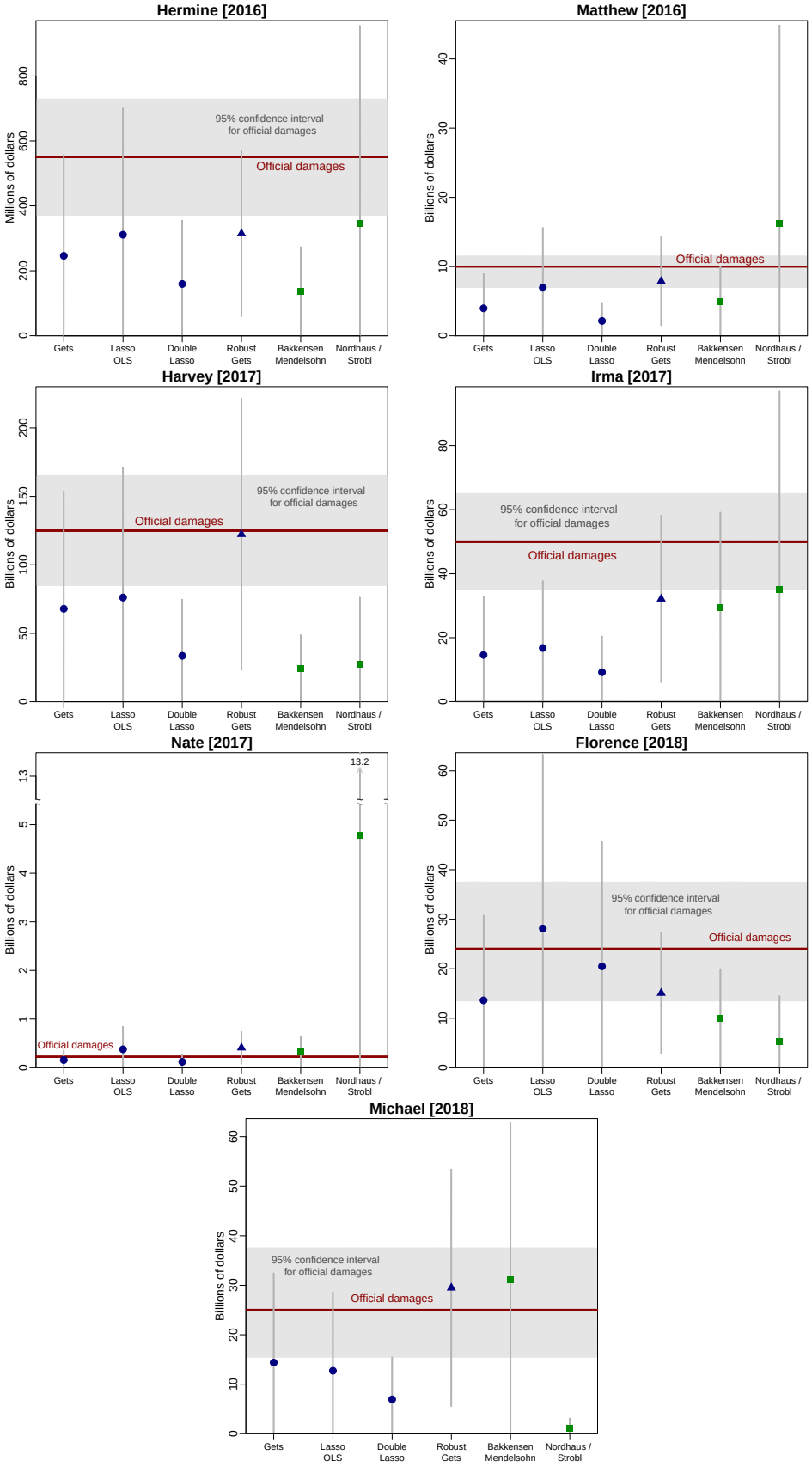
	Dynamics		Dynamics & Adaptation	
	(1) 12-hours	(2) 36-hours	(3) 12-hours	(4) 36-hours
Housing density	0.26**	0.24**	0.18	0.08
	(0.13)	(0.11)	(0.13)	(0.10)
Income per housing unit	1.45***	1.30***	2.04***	2.20***
	(0.25)	(0.20)	(0.32)	(0.30)
Income per housing unit sq.	0.40**	0.53***	0.77***	1.02***
	(0.18)	(0.16)	(0.23)	(0.27)
Min central pressure (-)	55.1***	60.2***	54.5***	58.9***
	(7.11)	(7.46)	(6.69)	(6.76)
Max rainfall	0.45*	0.25	0.47*	0.28
	(0.26)	(0.23)	(0.24)	(0.21)
Max storm surge	1.10***	0.94***	1.10***	0.93***
	(0.29)	(0.28)	(0.27)	(0.31)
Warnings	0.31*	0.23*	0.38**	0.36***
	(0.16)	(0.12)	(0.15)	(0.12)
H-hour error (IV)	0.55**	0.23	0.49*	0.34
	(0.27)	(0.19)	(0.26)	(0.20)
H-hour radius			2.95**	2.44***
			(1.20)	(0.88)
Levee length			-3.07	-5.59***
			(2.34)	(2.34)
Outlying storms dummy	-3.55***	-3.86***	-3.11***	-3.38***
	(0.51)	(0.56)	(0.54)	(0.53)
$\hat{\sigma}$	0.764	0.785	0.721	0.717
R^2	0.928	0.925	0.939	0.939

Notes: All equations are estimated using 2SLS with 65 observations and include a constant and a dummy variable for Floyd [1987]. The standard errors are in parentheses. Heteroskedasticity corrected standard errors are shown in columns (3) and (4). *p< 0.1 **p< 0.05 ***p< 0.01

6.4 Out-of-Sample Fit

Another concern is that we over fit the model in sample. Although *Gets* does not explicitly consider goodness of fit in its selection procedure, this is a common feature of model selection and machine learning techniques that do. We can check for potential over fit by evaluating the out-of-sample performance of the model. Since our estimation sample includes hurricanes that made landfall through 2015, we can evaluate the model performance using hurricanes that made landfall in 2016 through 2018. This provides seven additional observations for the analysis.

We compare the out-of-sample performance across several models. The first is the selected model using *Gets* from column (4) of Table 5.1. The second is the selected model using Lasso from column (5) of Table 5.1. The third is the selected model using Double Lasso from column (6) of Table 5.1. The fourth is the robust model from column (4) of Table 6.1. The fifth is a simple model of damage



Notes: Official damages and confidence intervals are from NOAA's Tropical Cyclone Reports and NOAA's Billion-Dollar Events Database. The forecast standard errors are computed using the delta method under the assumption of normality and are truncated at zero.

Figure 6.1: Out-of-Sample Damage 'Predictions' by Storm and Method

based on the estimated relationship with income and central pressure (Bakkensen and Mendelsohn, 2016). The sixth, and final model, is a simple model of damage using a fixed relationship with income and an estimated relationship with central pressure (Nordhaus, 2010; Strobl, 2011).

Table 6.4 presents the results. Overall, the robust model performs considerably better than any of the other models. This is in part driven by Harvey [2017] where the robust model almost perfectly predicts official damages. However, as Figure 6.1 shows, the robust model generally performs well for each hurricane in the sample. In fact, it never does worse than most models and official damages always fall within 1 standard deviation of the forecast. This reinforces the conclusion that the selected model is robust both in-sample and out.

Table 6.4: Relative Out-of-Sample Model Performance

	(1) Bias	(2) RMSE	(3) MAPE
Gets	0.89	0.66	0.94
Lasso OLS	0.69	0.58	0.84
Double Lasso	1.20	0.98	1.18
Robust Gets	**0.20**	**0.20**	**0.64**
Nordhaus (2010)	1.07	0.99	6.47

RMSE: Root mean square error. MAPE: Mean absolute percent error. Metrics are computed relative to the performance of the Bakkensen and Mendelsohn (2016) model. Values smaller than 1 indicate better performance.

7 The Value of Improved Hurricane Forecast Accuracy

Given the consistency of our results, we can conduct a counterfactual experiment to assess the short-term impact of improvements in forecast accuracy on hurricane damages. In order for this experiment to be valid we require that the underlying model is correctly specified and that the parameters of the model are invariant to changes in the forecast errors (i.e. they are super exogenous; Engle et al., 1983). Our results suggest that the robust model in column (4) of Table 6.1 is well specified in that it satisfies all of the diagnostic checks and has the best in-sample and out-of-sample performance. However, we still need to test whether the model is invariant to changes in the forecast errors.

There are several tests of super exogeneity by Engle and Hendry (1993), Hendry and Santos

Table 7.1: A Test of Damage Model Invariance

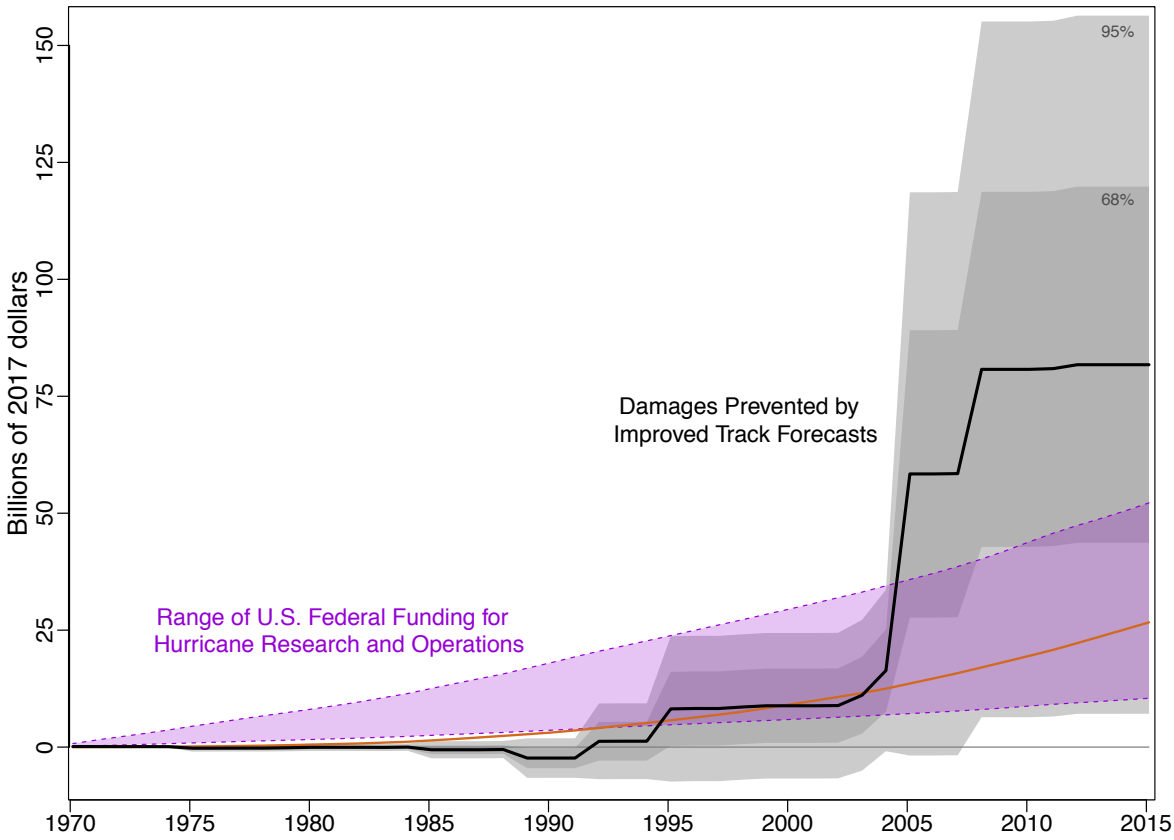
	(1) Marginal + IIS	(2) Augmented Conditional
Dependent variable:	12-hour errors	Damages
Connie [1955]	1.29 (0.57)**	-0.96 (0.96)
Cleo [1964]	-1.21 (0.51)**	1.38 (0.87)
Opal [1995]	-1.48 (0.52)***	0.24 (0.89)
Isabel [2003]	-1.19 (0.50)**	0.98 (0.85)
Arthur [2014]	-1.73 (0.54)***	-0.53 (0.94)
F-statistic	.	1.05
p-value	.	[0.394]

Notes: Both equations are estimated with 98 observations and all variables from the Robust model. IIS is run with a 1% target gauge. Standard errors are in parentheses. The tail probability associated with the null hypothesis of the test statistic is in square brackets. *p< 0.1 **p< 0.05 ***p< 0.01

(2010), and Castle et al. (2017). We use the automatic test for super exogeneity proposed by Hendry and Santos (2010). Their procedure is described in three steps. First, we run IIS on a marginal model of the forecast errors regressed on all other explanatory variables. We retain any impulses detected in the marginal model. Next, we estimate the conditional model of damages regressed on the forecast errors, all other explanatory variables, and the retained impulses from the marginal model. Finally, we test the joint significance of these impulses in the conditional model. The null hypothesis is that the model is invariant so that any large changes in the marginal model equation will pass directly through the forecast errors and not otherwise alter the damage model.

Table 7.1 shows that using this automatic test we find five impulses in the marginal equation: Connie [1955], Cleo [1964], Opal [1995], Isabel [2003] and Arthur [2014]. However, after adding them to the conditional model and testing for their joint significance, we obtain a test statistic of 1.05 which when compared against an F-distribution with (5,83) degrees of freedom indicates that the null hypothesis of super exogeneity is not rejected with a p-value of 39.4 percent.

Since super exogeneity is not rejected and the model is well specified, we can proceed with the counterfactual exercise. We use the model to predict what damages would have been if the average landfall-forecast error from 1955-1969 had occurred for all strikes since 1970. From that, we subtract the predicted damages based on the actual forecast error for each strike to get a prediction of the damages prevented since 1970. We cumulate this difference to see how it has evolved over time. Technical Appendix B.2 describes how we calculate uncertainty around the cumulative difference.



Notes: Damages prevented is calculated as the difference in cumulative damages estimated using the actual forecast error vs. damages estimated using the average forecast error from 1955-1969. The model used for estimation is column (4) of Table 6.1. The 5% - 95% confidence interval around this estimate is computed using the delta method assuming independence and normality. See Technical Appendix B.2 for details. Cumulative federal funding for hurricane research and operations is taken from historical editions of the Office of the Federal Coordinator for Meteorology's Federal Plan for Meteorological Services and Supporting Research. The range is computed as being between 7% and 33% of total funding for meteorological operations and research costs following National Science Board (2007, see footnote 46). The single line is willingness to pay and is calculated as the cumulative sum of the population in every coastal county struck by a hurricane since 1970 times the real value of $14.73 over time.

Figure 7.1: The Costs and Some Benefits of Improving Forecast Accuracy since 1970

Figure 7.1 illustrates that the prediction of the cumulative damages prevented between 1970-2015 due to forecast improvements is around $82 billion. To put this into context, this value is 14 percent of the total predicted damages over this period so just shy of the damage caused by Hurricane Maria [2018] alone. We can conduct a cost benefit analysis by comparing the total benefit against the cumulative cost of producing the forecasts and their improvements since 1970, which were obtained from historical reports of the Office of the Federal Coordinator for Meteorology. This comparison illustrates that after accounting for the costs, the predicted net savings are around $\$30-71$ billion.

We can also compare our prediction of damages prevented against a measure of the cumulative private willingness to pay for forecast improvements since 1970 as extrapolated from the findings of Katz and Lazo (2011). Figure 7.1 illustrates that damages prevented due to forecast improvements

are greater than both public and private willingness to pay. This suggests that both individuals and the federal government have severely underestimated the value of improving hurricane forecasts. This is robust to alternative model specifications but should be considered a lower bound of the total net benefit from hurricane forecast improvements since we do not account for fatalities prevented (Willoughby et al., 2007) or reduced evacuation and damage mitigation costs (Regnier, 2008).

8 Conclusions

Forecasts of natural disasters can alter the destructiveness of an event when operating through individuals' beliefs about the costs of damage mitigation. This is particularly true if individuals have a false sense of security about the accuracy of these imperfect forecasts. In this chapter we test for and quantify this relationship using an empirical model of damages for all hurricanes to strike the continental United States in the past 60 years. We start by estimating the full empirical model using OLS. Next, we simplify the model using model selection methods and show that a small subset of drivers, including the 12-hour-ahead forecast errors, explain most of the variation in hurricane damages.

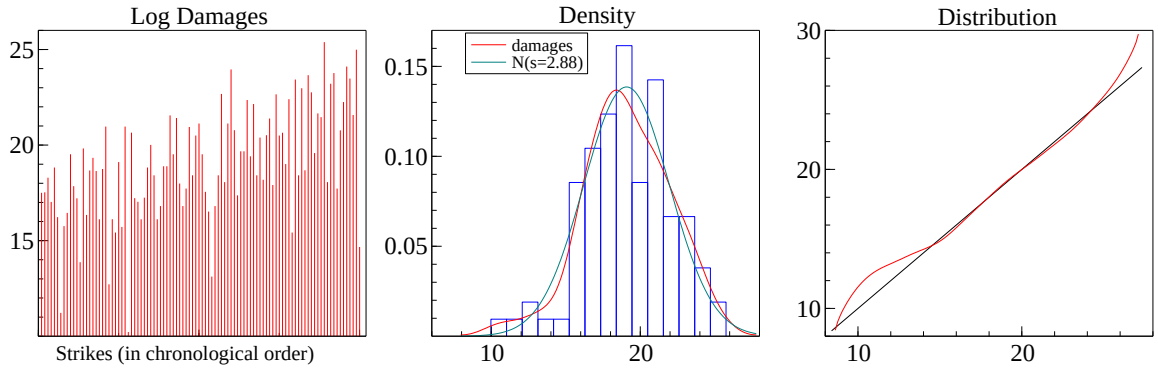
There is a positive and statistically significant relationship between the 12-hour-ahead landfall-forecast errors and damages. This relationship is consistent with the predictions of our theory. It is robust to outliers, alternative measures of uncertainty, model specifications, out-of-sample storms, and controls for storm dynamics and technological change. Overall, we find that a one standard deviation increase in the distance between the storm's predicted and actual landfall location leads to \$3,000 additional dollars in damages per household. Interpreting this result through the lens of the theoretical framework, it indicates that the forecasts play a critical role in guiding individuals' beliefs about the value of short-term damage mitigation efforts. It also illustrates that, in aggregate, individual decisions to protect and relocate property in the face of a disaster can have a significant impact on the overall cost of a natural disaster.

Focusing on the specific implications of the results, we find that improvements in the forecasts since 1970 have resulted in total damages being approximately \$82 billion less than they otherwise

would have been. Although damages increased due to changes in vulnerabilities and natural hazards, improvements in forecast accuracy along with other longer-term adaptation efforts have kept damages from rising faster than they otherwise would have. Comparing the cumulative damages prevented against the cost of producing the forecasts, we find that there is a net benefit of around $\$30 - 71$ billion. This illustrates that improvements in hurricane forecasts over the past few decades produced benefits beyond the well-documented reduction in fatalities and have outweighed the associated costs.

This is particularly important since hurricanes are expected to become even more difficult to forecast in the future. Knutson et al. (2010) argue that climate change will increase hurricane intensity. As a result, in the future we are more likely to see hurricanes akin to Harvey [2017] whose storm dynamics are hard to predict (Emanuel, 2017a,b). In light of this, our findings support maintained investment in, and continued measures to improve, hurricane forecasting capabilities along with other longer-term adaptation efforts so that any future loss of life and property can be minimized.

A Appendix: Additional Figures and Tables



Notes: Hurricane damages are the logs of the nominal values of hurricane damages for each hurricane strike.

Figure A.1: Hurricane Damages

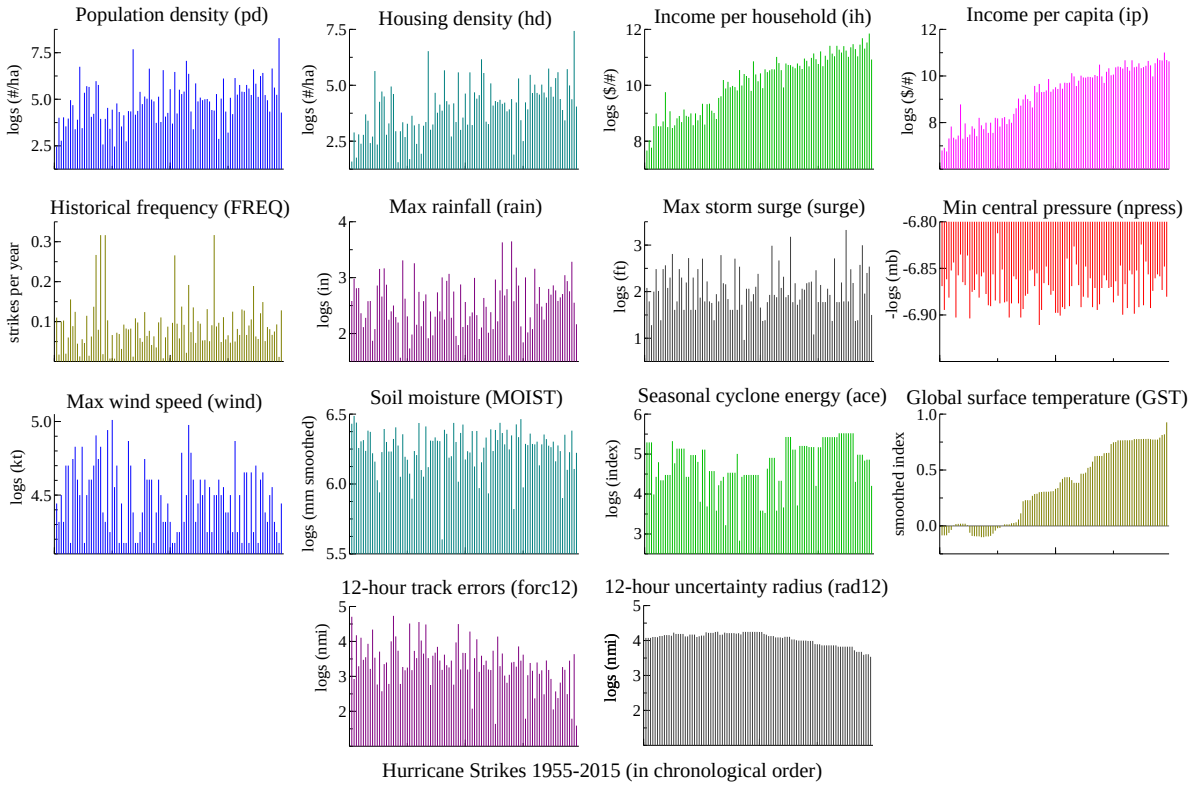


Figure A.2: Data Plots

Table A.1: Bivariate Correlations

	pd	hd	ih	ip	FREQ	rain	surge	npress	wind	MOIST	ace	GST	forc12	cone12	Trend
dam	**0.51**	**0.51**	**0.59**	**0.54**	-0.14	0.32	**0.54**	**0.60**	0.34	0.09	0.30	**0.52**	-0.25	-0.43	**0.53**
pd		**0.98**	**0.55**	0.49	-0.02	0.14	0.12	0.11	-0.01	-0.02	0.05	0.39	-0.15	-0.29	0.41
hd			**0.64**	**0.61**	0.01	0.11	0.08	0.09	-0.05	-0.01	0.08	0.50	-0.19	-0.36	**0.53**
ih				**0.99**	0.02	0.11	0.07	0.05	-0.15	0.00	0.21	**0.93**	-0.42	**-0.63**	**0.95**
ip					0.04	0.07	0.03	0.02	-0.18	0.01	0.21	**0.94**	-0.42	**-0.64**	**0.97**
FREQ						-0.17	-0.10	-0.04	0.12	0.05	0.03	0.01	-0.04	-0.03	0.02
rain							0.09	0.11	0.01	0.22	0.19	0.14	-0.24	-0.20	0.12
surge								**0.65**	**0.52**	0.11	0.08	0.01	-0.10	-0.09	0.01
npress									**0.81**	-0.02	0.18	0.03	-0.01	-0.14	0.01
wind										0.03	0.12	-0.20	0.08	0.08	-0.19
MOIST											0.08	0.01	0.01	-0.02	-0.02
ace												0.35	-0.26	-0.43	0.28
GST													-0.46	**-0.79**	**0.98**
forc12														0.43	-0.44
cone12															**0.74**

Table A.2: Estimated Terminal Models for a Target Gauge of 3%

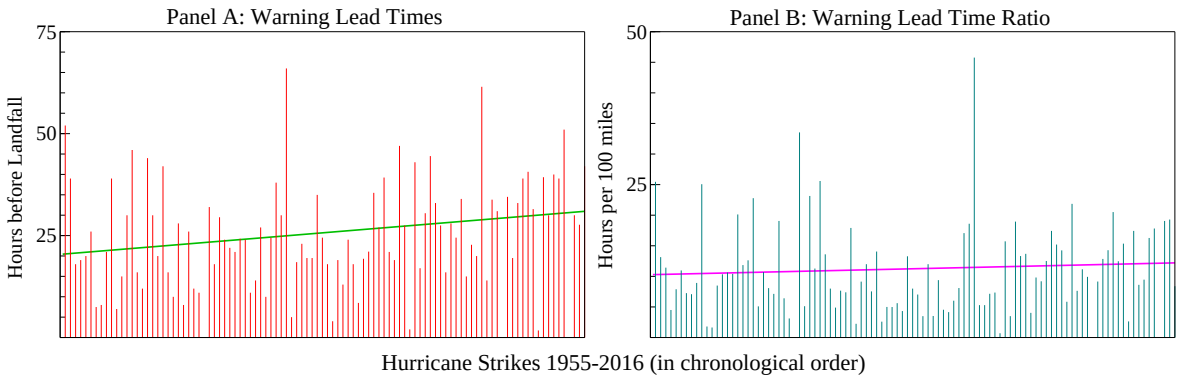
Terminal Model:	Final GUM	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Housing density	0.60***	0.58***	0.64***	0.61***	0.59***	0.40***	0.64***	0.56***	0.44***	0.55***
	(0.18)	(0.14)	(0.19)	(0.14)	(0.14)	(0.15)	(0.18)	(0.14)	(0.15)	(0.19)
Income per household	0.47		1.01***			1.28***	1.01***		1.22***	1.12***
	(0.58)		(0.19)			(0.18)	(0.17)		(0.20)	(0.21)
Historical fequency	-4.46*		-6.35**	-4.85**	-5.18**		-6.54***	-4.21*		
	(2.48)		(2.53)	(2.24)	(2.26)		(2.31)	(2.25)		
Max rainfall	0.59*	1.12***				1.14***		0.72**	0.97***	0.92***
	(0.36)	(0.32)				(0.31)		(0.32)	(0.31)	(0.34)
Max storm surge	1.16***	1.44***		1.41***	1.45***	1.34***	1.30***	1.24***	1.30***	1.27***
	(0.41)	(0.38)		(0.38)	(0.39)	(0.37)	(0.38)	(0.39)	(0.38)	(0.38)
Min central pressure (-)	63.0***	53.9***	87.7***	53.6***	53.2***	52.0***	52.8***	51.2***	52.5***	68.3***
	(14.0)	(8.66)	(13.0)	(8.73)	(8.90)	(8.43)	(8.58)	(8.62)	(8.57)	(13.9)
Max wind speed	-1.03		-1.59							-1.53
	(1.16)		(1.23)							(1.17)
Seasonal cyclone energy	0.53**			0.57**	0.61**		0.47*	0.42*	0.48*	
	(0.25)			(0.25)	(0.26)		(0.24)	(0.24)	(0.25)	
Soil Moisture	0.85		1.69*							
	(0.97)		(0.95)							
12-hour forecast error	0.52**	0.44*			0.33	0.48**		0.53**		
	(0.24)	(0.24)			(0.24)	(0.23)		(0.24)		
12-hour radius	1.95	2.35*		2.96**	2.52**				1.23	0.88
	(1.44)	(1.19)		(1.23)	(1.25)				(1.01)	(0.99)
Strike trend	0.03	0.05***		0.05***	0.05***			0.04***		
	(0.02)	(0.01)		(0.01)	(0.01)			(0.01)		
AUG	0.20		0.16							0.15
	(0.29)		(0.32)							(0.301)
NY	-1.05		-1.78**				-1.54**			-0.85
	(0.85)		(0.81)				(0.74)			(0.80)
VA	1.73**			1.55*				1.79**		
	(0.82)			(0.78)				(0.81)		
NC	-0.71*		-0.74**					-0.74**		
	(0.42)		(0.35)					(0.36)		
k	16	7	9	8	8	6	7	10	7	9
Log-likelihood (-)	147.9	157.6	161.3	157.6	158.8	157.1	157.5	153.5	157.1	157.8
AIC	3.406	3.420	3.537	3.441	3.464	**3.389**	3.417	3.397	3.411	3.464
HQ	3.609	3.527	3.666	3.558	3.582	**3.485**	3.524	3.536	3.517	3.592
BIC	3.908	3.684	3.854	3.731	3.754	**3.626**	3.681	3.740	3.675	3.780
$\hat{\sigma}$	1.219	1.275	1.340	1.282	1.298	1.261	1.273	1.244	1.269	1.291
R^2	0.851	0.819	0.804	0.819	0.814	0.821	0.819	0.833	0.821	0.818

Notes: Estimated using 98 observations and include a constant and dummy variables for Gerda [1969] and Floyd [1987]. Each terminal model is selected from the Final GUM using a target gauge (target value) of 3%. Bold values indicate terminal models with the lowest information criteria. The standard errors are in parentheses. k is the number of selected regressors in the model. *p< 0.1 **p< 0.05 ***p< 0.01

Table A.3: Final Estimated GUMs Across Different Targets

Target gauge:	(Original GUM)	(All)	1%	2%	3%	4%	5%
Housing density	0.53**	0.50**	0.46**	0.43**	0.60***	0.44**	0.46**
	(0.24)	(0.21)	(0.20)	(0.20)	(0.18)	(0.20)	(0.20)
Income per household	0.70	0.74	0.75	1.08	0.47	0.79	0.87
	(0.79)	(0.70)	(0.68)	(0.67)	(0.58)	(0.68)	(0.61)
Historical fequency	-6.63**	-6.81**	-7.09***	-6.14**	-4.46*	-6.30**	-6.10**
	(2.94)	(2.70)	(2.58)	(2.68)	(2.48)	(2.63)	(2.57)
Max rainfall	0.51	0.59	0.59	0.67*	0.59*	0.61	0.65*
	(0.43)	(0.38)	(0.37)	(0.38)	(0.36)	(0.37)	(0.36)
Max storm surge	1.21**	1.13**	1.15***	1.18***	1.16***	1.26***	1.25***
	(0.48)	(0.43)	(0.42)	(0.43)	(0.41)	(0.41)	(0.40)
Min central pressure (-)	52.6***	56.3***	51.8***	59.5***	63.0***	53.6***	53.4***
	(18.2)	(15.0)	(9.25)	(14.8)	(14.0)	(14.4)	(13.9)
Max wind speed	-0.27	-0.49		-0.98	-1.03	-0.49	-0.65
	(1.54)	(1.25)		(1.24)	(1.16)	(1.23)	(1.17)
Seasonal cyclone energy	0.42	0.32	0.31	0.45	0.53**	0.38	0.38
	(0.32)	(0.28)	(0.28)	(0.27)	(0.25)	(0.27)	(0.25)
Soil moisture	0.77	0.57	0.65	0.11	0.85		
	(1.42)	(1.09)	(1.03)	(1.05)	(0.97)		
Sea surface temperature	0.94	0.13	0.23	-1.79		0.18	
	(3.21)	(2.51)	(2.39)	(2.28)		(2.47)	
12-hour forecast error	0.54*	0.62**	0.62**	0.59**	0.52**	0.64**	0.69***
	(0.29)	(0.26)	(0.25)	(0.26)	(0.24)	(0.25)	(0.24)
12-hour radius	3.17	2.48	2.55	0.51	1.95	2.43	2.11
	(2.60)	(2.06)	(2.01)	(1.65)	(1.44)	(1.95)	(1.62)
Trends:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects:	Yes	Yes	Yes	Yes	No	Yes	No
Month fixed effects:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
U.S. State fixed effects:	Yes	Yes	Yes	Yes	Yes	Yes	Yes
k	37	27	24	22	16	24	21
$\hat{\sigma}$	1.300	1.234	1.210	1.247	1.219	1.219	1.203
R^2	0.876	0.869	0.868	0.856	0.851	0.866	0.864

Notes: All equations are estimated using 98 observations and include a constant and dummy variables for Gerda [1969] and Floyd [1987]. The standard errors are in parentheses. 'All' combines the retained variables from each Final GUM. k is the number of selected regressors in the model. *p< 0.1 **p< 0.05 ***p< 0.01



Note: Hurricane warning lead times are computed as the number of hours prior to making landfall the first warning that was issued for the actual landfall location. The warning lead time ratio is computed as the warning lead time divided by the (straight line) distance in miles which the initial warning covered to proxy the uncertainty surrounding the warning area.

Figure A.3: Hurricane Warnings

Table A.4: Alternative Measures of Normalized Damages

	(1) Nominal	(2) Real	(3) Norm-1	(4) Norm-2	(5) Norm-3
Housing density	0.27***	0.35***	0.43***	0.38***	-0.78***
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)
Income per housing unit	1.43***	0.74***	-0.00	-0.01	0.27**
	(0.12)	(0.19)	(0.13)	(0.14)	(0.13)
Income per housing unit sq.	0.44***	0.42***	0.38***	0.42***	0.38***
	(0.09)	(0.09)	(0.10)	(0.10)	(0.09)
Min central pressure (-)	56.6***	57.2***	61.2***	61.2***	46.1***
	(5.49)	(5.53)	(6.06)	(6.11)	(5.82)
Max rainfall	0.57***	0.56***	0.66***	0.64***	0.35
	(0.21)	(0.21)	(0.23)	(0.23)	(0.22)
Max storm surge	0.99***	1.06***	0.96***	0.95***	0.98***
	(0.24)	(0.25)	(0.27)	(0.27)	(0.26)
12-hour forecast errors	0.34**	0.33**	0.24	0.29*	0.40**
	(0.15)	(0.15)	(0.17)	(0.17)	(0.16)
Outlying storms dummy	-3.44***	-3.47***	-3.63***	-3.56***	-3.20***
	(0.33)	(0.34)	(0.37)	(0.37)	(0.35)
$\hat{\sigma}$	0.812	0.817	0.896	0.902	0.860
R^2	0.927	0.908	0.882	0.879	0.853

Notes: All equations are estimated using 98 observations and include a constant and a dummy variable for Gerda [1969] and Floyd [1987]. Standard errors are in parentheses. The different normalizations are: (2) CPI inflation; (3) Pielke Jr and Landsea (1998); (4) Pielke Jr et al. (2008); (5) Neumayer and Barthel (2011). *p< 0.1 **p< 0.05 ***p< 0.01

Table A.5: Controlling for Storm Dynamics and Adaptation Efforts

	(1) Final	(2) (1)+ RAD+LEV	(3) (1)+ CRS+HMG	(4) (1)+ (2)+(3)	(5) (1)+ Naïve	(6) (5)+ RAD+LEV	(7) (5)+ CRS+HMG	(8) (5)+ (6)+(7)
Housing density	0.27	0.16	0.17	0.10	0.25	0.15	0.15	0.09
	(0.18)	(0.19)	(0.21)	(0.22)	(0.18)	(0.19)	(0.21)	(0.22)
Income per housing unit	4.12***	4.39***	4.17***	4.24***	4.18***	4.41***	4.23***	4.23***
	(1.16)	(1.16)	(1.23)	(1.23)	(1.17)	(1.18)	(1.24)	(1.24)
Income per housing unit sq.	-1.07	-0.64	-1.01	-0.51	-1.14*	-0.70	-1.07	-0.57
	(0.62)	(0.66)	(0.64)	(0.70)	(0.63)	(0.67)	(0.65)	(0.71)
Min central pressure (-)	65.6***	64.5***	63.8***	63.2***	64.7***	63.7***	62.9***	62.4***
	(9.43)	(9.32)	(9.77)	(9.67)	(9.55)	(9.47)	(9.91)	(9.82)
Max rainfall	0.20	0.16	0.26	0.19	0.21	0.17	0.27	0.19
	(0.31)	(0.31)	(0.32)	(0.32)	(0.31)	(0.31)	(0.32)	(0.33)
Max storm surge	1.09**	1.03**	1.15**	1.12**	1.09**	1.03**	1.16**	1.12**
	(0.40)	(0.40)	(0.42)	(0.42)	(0.40)	(0.40)	(0.42)	(0.42)
12-hour official error	0.39	0.29	0.41	0.36	0.37	0.28	0.39	0.35
	(0.23)	(0.24)	(0.25)	(0.25)	(0.24)	(0.24)	(0.26)	(0.26)
12-hour naïve error					0.19	0.17	0.19	0.19
					(0.24)	(0.26)	(0.25)	(0.262)
12-hour radius		2.58		2.76		2.53		2.77
		(1.52)		(1.72)		(1.54)		(1.74)
Levee length		-1.61		-1.74		-0.79		-0.85
		(3.17)		(3.25)		(3.42)		(3.51)
Community rating system			-0.17	-0.19			-0.18	-0.20
			(0.22)	(0.21)			(0.22)	(0.22)
HMG spending per capita			-0.02	0.00			-0.02	0.00
			(0.03)	(0.03)			(0.03)	(0.04)
Outlying storms dummy	-3.56***	-3.22***	-3.57***	-3.15***	-3.58***	-3.23***	-3.59***	-3.14***
	(0.55)	(0.59)	(0.57)	(0.64)	(0.56)	(0.60)	(0.58)	(0.65)
$\hat{\sigma}$	0.783	0.772	0.797	0.788	0.788	0.780	0.802	0.795
R^2	0.923	0.930	0.925	0.932	0.924	0.931	0.927	0.933

Notes: All equations are estimated using 41 observations and include a constant. The standard errors are in parentheses. *p< 0.1 **p< 0.05 ***p< 0.01

B Technical Appendix

B.1 Approximation of the Normal Distribution

Given the assumption that the past forecast errors have a normal density then

$$g_i(F_{1i}|\hat{F}_{1i}) = \frac{1}{\sqrt{2\pi}\sigma_{i,(F_{1_},\hat{F}_{1_})}} e^{-(F_{1i}-\hat{F}_{1i})^2/2\sigma_i^2}, \tag{B.1}$$

which we can rewrite in terms of a standard normal density as

$$g(z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i{}^2}{2}}, \tag{B.2}$$

where $z_i = \frac{F_{1i}-\hat{F}_{1i}}{\sigma_{i,(F_{1_},\hat{F}_{1_})}}$ so that the density is centered around the prediction. The integral of the standard normal density becomes the standard normal distribution function so that

$$\Phi\left(F_{1i},\hat{F}_{1i}\right) = \int_{-\infty}^{(F_i-\hat{F}_i)} g(z_i)\, dz_i. \tag{B.3}$$

This can be rewritten as

$$\Phi\left(F_{1i},\hat{F}_{1i}\right) = \frac{1}{2} erf\left\{\frac{F_{1i}-\hat{F}_{1i}}{\sqrt{2}\sigma_{i,(F_{1_},\hat{F}_{1_})}}\right\}, \tag{B.4}$$

where $erf(\cdot)$ is the Gaussian error function, which can be expanded as

$$\begin{aligned} erf(x_i) &= \frac{2x_i}{\sqrt{\pi}} {}_1F_1\left(\frac{1}{2},\frac{3}{2},-x_i^2\right) \\ &= \frac{2x_i}{\sqrt{\pi}}\left[\sum_{k=0}^{\infty}\frac{(-1)^k x_i^{2k}}{(2k+1)\,k!}\right], \end{aligned} \tag{B.5}$$

where ${}_1F_1(a,b,c)$ is a confluent hypergeometric function of the first kind. Plugging (B.5) into (B.4):

$$\Phi\left(F_{1i},\hat{F}_{1i}\right) = \frac{F_{1i}-\hat{F}_{1i}}{\sigma_{i,(F_{1_},\hat{F}_{1_})}}\left[\frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{\left(-\frac{1}{2}\right)^k}{(2k+1)\,k!}\left(\frac{F_{1i}-\hat{F}_{1i}}{\sigma_{i,(F_{1_},\hat{F}_{1_})}}\right)^{2k}\right]. \tag{B.6}$$

(B.6) illustrates that the normal distribution is a rescaling of the forecast errors relative to their standard deviation. When $\frac{|F_{1i}-\hat{F}_{1i}|}{\sigma_{i,(F,\hat{F})}} < 1$ then $\frac{|F_{1i}-\hat{F}_{1i}|}{\sigma_{i,(F,\hat{F})}}$ provides a close approximation of (B.6) since the terms inside the brackets collapse to zero for large k. However, when $\frac{|F_{1i}-\hat{F}_{1i}|}{\sigma_{i,(F,\hat{F})}} \geq 1$, then the terms inside the brackets downscale (B.6) so that $\frac{|F_{1i}-\hat{F}_{1i}|}{\sigma_{i,(F,\hat{F})}}$ will be more sensitive to larger forecast errors.

B.2 Counterfactual Policy Analysis

Consider a simple representation of damages as:

$$ln(p_i y_i) = X_i'\beta + \epsilon_i, \tag{B.7}$$

where X_i is a vector of explanatory variables, ϵ_i is assumed to be iid normal and $\mathbb{E}\left(X_i \epsilon_j\right) = \mathbf{0}\, \forall i, j$. We estimate this model, apply the delta method, and re-scale by prices to get a prediction of real damages

$$\widehat{y}_i \sim \mathrm{IN}\left(y_i, \left(y_i \sigma_{i,\widehat{y}}\right)^2\right). \tag{B.8}$$

The difference between this prediction and some counterfactual, $\widetilde{y}_i = k_i y_i$, gives

$$\begin{aligned}(\widetilde{y}_i - \widehat{y}_i) &\sim \mathrm{IN}\left((k_i - 1)\, y_i, \mathbb{V}(\widetilde{y}_i - \widehat{y}_i)\right), \\ \mathbb{V}(\widetilde{y}_i - \widehat{y}_i) &= \left(k_i y_i \sigma_{i,\widetilde{y}}\right)^2 + \left(y_i \sigma_{i,\widehat{y}}\right)^2 - 2 * Cov(\widetilde{y}_i, \widehat{y}_i),\end{aligned} \tag{B.9}$$

From the independence, normality and exogeneity assumptions, (B.9) can be cumulated as:

$$\sum_{i=1}^{n} (\widetilde{y}_i - \widehat{y}_i) \sim \mathrm{N}\left(\sum_{i=1}^{n} (k_i - 1)\, y_i, \sum_{i=1}^{n} \mathbb{V}(\widetilde{y}_i - \widehat{y}_i)\right). \tag{B.10}$$

The variance is the residual error variance plus parameter estimation uncertainty

$$\sigma^2_{i,\hat{y}} = \left(\beta - \widehat{\beta}\right)' \mathbb{V}\left(X_i\right)\left(\beta - \widehat{\beta}\right) + \sigma^2_{i,\epsilon} \tag{B.11}$$

$$\sigma^2_{i,\tilde{y}} = \left(\beta - \widehat{\beta}\right)' \mathbb{V}\left(\tilde{X}_i\right)\left(\beta - \widehat{\beta}\right) + \sigma^2_{i,\epsilon}, \tag{B.12}$$

which Doornik and Hendry (2013b) show can be approximately estimated as

$$\hat{\sigma}^2_{i,\hat{y}} = X_i' \left[\widehat{\mathbb{V}\left(\widehat{\beta}\right)}\right] X_i + \hat{\sigma}^2_{i,\epsilon} \tag{B.13}$$

$$\hat{\sigma}^2_{i,\tilde{y}} = \tilde{X}_i' \left[\widehat{\mathbb{V}\left(\widehat{\beta}\right)}\right] \tilde{X}_i + \hat{\sigma}^2_{i,\epsilon}, \tag{B.14}$$

where $\mathbb{V}\left(\widehat{\beta}\right)$ is the covariance matrix of the parameter estimates. The covariance is estimated as

$$\widehat{Cov}\left(\widetilde{y}_i, \widehat{y}_i\right) = k_i y_i^2 \left\{\tilde{X}_i' \left[\widehat{\mathbb{V}\left(\widehat{\beta}\right)}\right] X_i + \hat{\sigma}^2_{i,\epsilon}\right\}. \tag{B.15}$$

Bringing these pieces together and simplifying, the total estimate of the variance is

$$\widehat{\mathbb{V}\left(\widetilde{y}_i - \widehat{y}_i\right)} = y_i^2 \left\{\left(k_i \tilde{X}_i - X_i\right)' \left[\widehat{\mathbb{V}\left(\widehat{\beta}\right)}\right] \left(k_i \tilde{X}_i - X_i\right) + \left(k_i - 1\right)^2 \hat{\sigma}^2_{i,\epsilon}\right\}. \tag{B.16}$$

When $\tilde{y}_i = \hat{y}_i$ then $k_i = 1$ and $X_i = \tilde{X}_i$ so (B.16) collapses to zero. Using (B.16), we construct a confidence interval around the difference between the predicted and counterfactual damages.

Chapter 5

Conclusions and Future Work

This chapter summarizes the findings from each of the individual chapters and provides a potential path forward for future research avenues that extend the multi-step system forecast evaluation framework set forth in this thesis. There are many ways to evaluate imperfect forecasts. We show the importance of considering the joint distributions over variables and horizons. Once a multi-step system context is envisaged, potential singularities need to be addressed, as do tests of significant differences between the forecasts, and assessments of the value of improved forecasts.

Chapter 2 produces design-free and model-free estimates of forecast-error second-moment matrices in high dimensional settings where there are more variables times forecast horizons than forecast-error observations. Simulations show that the "standard" approach is biased downwards and provides poor rankings of forecast accuracy when there are relatively few observations. This has important implications for analyses of second-moment matrices with few observations. The simulations also illustrate that "design-free" and model-free methods substantially reduce the bias relative to the standard approach and other more restricted methods across a variety of forecast models and data generation processes. The methods also produce the best forecast rankings relative to other procedures even when there is a relatively small number of observations.

The simulations and empirical results suggest that design-free estimators that split subsamples evenly for the eigenvalues and eigenvectors perform best in most situations. While restricted methods

perform well when the forecast errors are transformed, the application illustrates that performance deteriorates markedly when they are not transformed and the imposed restrictions are invalid. The application also illustrates the importance of accounting for outliers and time-varying bias when applying these methods. Overall, this illustrates the benefits of extending design-free and model-free estimates to second-moment matrices in high dimensional cases while allowing for time-varying bias.

The analysis in Chapter 2 can be extended in a variety of important ways. First, design-free and model-free methods could be used to estimate high dimensional multivariate autocovariance matrices and compared with other approaches in the literature to assess their strengths and weaknesses; see Proietti and Giovannelli (2018). Second, the model-free decomposition of multi-horizon forecast errors with time-varying bias could be a used to produce alternative measures of multi-horizon time-varying forecast-error uncertainty; see Clark et al. (2019). Finally, it may be possible to extend estimates of forecast-error second-moment matrices to produce generalized loss functions that are robust to other types of measurement concerns; see Patton (2011) and Laurent et al. (2012, 2013).

Chapter 3 proposes general tests for differences in path forecast accuracy. The focus on point forecasts ignores forecast dynamics which are crucial to understanding the future trajectory of a series. Additional focus should be placed on the path, which accounts for these dynamics. We construct a general test for differences in path forecast accuracy by exploiting the link between the joint forecast error density and a general path loss function. We show how this joint density can be decomposed to allow for alternative weights across horizons. We demonstrate how for specific assumptions about the density / loss function, path tests are closely related to other joint multi-horizon tests. We also derive and examine the properties of these tests in an important special case.

Monte Carlo simulations illustrate the trade-offs associated with using path tests vis-a-vis alternative joint loss functions. Although path forecast accuracy tests are less able to detect differences in biases, which are offset by higher covariances, path tests are more likely to capture differences in error variances, covariances and dynamics across forecast models. This is particularly true for differences in forecast-error dynamics, which other joint multi-horizon loss functions cannot detect.

We also show how path tests can be extended to higher dimensional settings.

We apply the path tests in two different settings. In the first application we compare the Federal Reserve Board's Greenbook path forecasts against four DSGE model path forecasts of GDP, inflation and interest rates. We find that while the Greenbook path inflation forecasts are dominant, the DSGE models ability to track the long-run mean does well for GDP. Decomposing differences along the path, we find that forecast dynamics play an important role in these findings and that there are common differences across models. In the second application we evaluate the path forecast accuracy of models for Hurricane Irma in 2017. We reduce the roughly 100 different methods down to a handful of best performing models. While official forecasts tend to dominate all methods, path tests show that alternative models can match its accuracy depending on how the path is weighted.

Overall, the analysis indicates that forecast dynamics play an important role in path forecast accuracy. The findings from the simulation exercises and applications reinforce the view that important aspects of forecast accuracy are overlooked when focusing exclusively on the point forecasts at individual horizons. There is additional value to using path forecast measures which capture the dependencies across horizons when examining multi-horizon forecasting systems.

The findings in Chapter 3 suggest several important possible future research avenues. First, it would be interesting to explore the links between the general path tests proposed in this chapter and the multi-horizon forecast encompassing tests proposed in Hungnes (2018). Second, the forecast density decomposition illustrated in this chapter could be used to think about multi-horizon forecast combinations and how they could account for forecast dynamics across horizons by updating the conditional densities; see Geweke and Amisano (2011), Ganics (2017), and McAlinn et al. (2019).

Chapter 4 proposes a framework to identify, test for, and quantify the economic impact of forecast accuracy in the context of damages from natural disasters. It shows how, in a theoretical framework, imperfect forecasts can lead to suboptimal levels of adaptation. Then in a log-linear version of this framework it illustrates that it is possible to incorporate forecast uncertainty and disentangle its effects from forecast accuracy. This allows us to identify the effects of forecast accuracy.

We apply this framework to an application of damages from all hurricanes to strike the continental United States since 1955. We start by estimating the full empirical model with many possible drivers and then simplify it using general-to-specific model selection procedures. The results from the model selection exercise show that a small but stable subset of drivers, including the 12-hour-ahead forecast errors, explain most of the variation in hurricane damages over the last six decades.

We find a positive and statistically significant relationship between the 12-hour-ahead landfall-forecast errors and damages. This relationship is consistent with the predictions from our theoretical framework. It is robust to outliers, alternative measures of uncertainty, model specifications, out-of-sample storms, and controls for storm dynamics and technological change. When interpreted through the lens of our framework, this result indicates that short-term hurricane forecasts play an important role in guiding individuals' beliefs about the value of short-term adaptation efforts.

Quantifying the overall impact of changes in forecast accuracy, we find that improvements in the forecasts since 1970 have resulted in total damages being approximately $82 billion less than they otherwise would have been. Although hurricane damages have increased over time due to changes in vulnerabilities and natural hazards, improvements in forecast accuracy along with other longer-term adaptation efforts have kept damages from rising faster than they otherwise would have. Comparing the cumulative damages prevented against the cost of producing the forecasts, we find that there is a net benefit of around $\$30 - 71$ billion. This illustrates that improvements in hurricane forecasts over the last several decades have produced large benefits.

The analysis in Chapter 4 suggests that there are several areas of potential future work. First, it may be possible to modify measures of forecast uncertainty so that they account for forecaster disagreement and multi-horizon forecast revisions; see Patton and Timmermann (2012). Another important question is how the framework discussed in this chapter might be adapted to be used in general and how it might be possible to test for or account for feedback between multi-horizon forecasts, economic decisions, and their consequences; see Beaudry and Willems (2018).

Overall the chapters in this thesis provide the outlines of a general framework for evaluating

multi-horizon system forecasts. They produce results addressing several important concerns and demonstrate the benefits of general-to-specific approaches to evaluate forecasts. They also provoke many unanswered questions that can guide a potential future research agenda for years to come.

References

Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-Free Estimation of Variance Matrices. *Journal of Econometrics*, 181(2):165 – 180.

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*, volume 1 of *Econometric Exercises*. New York, NY: Cambridge University Press.

Adolfson, M., Lindé, J., and Villani, M. (2007). Forecasting Performance of an Open Economy DSGE Model. *Econometric Reviews*, 26(2-4):289–328.

Amisano, G. and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics*, 25(2):177–190.

Anderson, R. G., Hoffman, D. L., and Rasche, R. H. (2002). A Vector Error-Correction Forecasting Model of the U.S. Economy. *Journal of Macroeconomics*, 24(4):569 – 598.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. New York: Wiley, third edition.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858.

Athanasopoulos, G., De Carvalho Guillén, O. T., Issler, J. V., and Vahid, F. (2011). Model Selection, Estimation and Forecasting in VAR Models with Short-Run and Long-Run Restrictions. *Journal of Econometrics*, 164(1):116 – 129.

Auerbach, A. J. (1995). Tax Projections and the Budget: Lessons From the 1980's. *American Economic Review: Papers and Proceedings*, 85(2):165–169.

Auerbach, A. J. (1999). On the Performance and Use of Government Revenue Forecasts. *National Tax Journal*, 52(4):767–782.

Baba, Y., Hendry, D. F., and Starr, R. M. (1992). The Demand for M1 in the USA, 1960–1988. *Review of Economic Studies*, 59(1):25–61.

Baillie, R. T. and Morana, C. (2009). Modelling Long Memory and Structural Breaks in Conditional Variances: An Adaptive FIGARCH Approach. *Journal of Economic Dynamics and Control*, 33(8):1577 – 1592.

Baker, S. R. and Bloom, N. (2013). Does Uncertainty Reduce Growth? Using Disasters as Natural Experiments. NBER Working Paper No. 19475, National Bureau of Economic Research, Cambridge, MA.

Bakkensen, L. A. and Mendelsohn, R. O. (2016). Risk and Adaptation: Evidence From Global Hurricane Damages and Fatalities. *Journal of the Association of Environmental and Resource Economists*, 3(3):555–587.

Bao, Y., Lee, T.-H., and Saltoğlu, B. (2007). Comparing Density Forecast Models. *Journal of Forecasting*, 26(3):203–225.

Beatty, T. K., Shimshack, J. P., and Volpe, R. J. (2019). Disaster Preparedness and Disaster Response: Evidence From Sales of Emergency Supplies Before and After Hurricanes. *Journal of the Association of Environmental and Resource Economists*, forthcoming.

Beaudry, P. and Willems, T. (2018). On the Macroeconomic Consequences of Over-Optimism. NBER Working Paper No. 24685, National Bureau of Economic Research, Cambridge, MA.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects After Selection Among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650.

Belongia, M. T. (1988). Are Economic Forecasts by Government Agencies Biased? *Federal Reserve Bank of St. Louis Review*, 70(6):15–23.

Berenguer Rico, V. and Nielsen, B. (2017). Marked and Weighted Empirical Processes of Residuals with Applications to Robust Regrssions. Department of Economics Disscusion Paper No. 841, University of Oxford, Oxford, United Kingdom.

Berg, T. O. (2016). Multivariate Forecasting with BVARs and DSGE Models. *Journal of Forecasting*, 35(8):718–740.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., and Others (2013). Valid Post-Selection Inference. *Annals of Statistics*, 41(2):802–837.

Bjørnstad, J. F. (1990). Predictive Likelihood: A Review. *Statistical Science*, 5(2):242–254.

Blackley, P. R. and DeBoer, L. (1993). Bias in OMB's Economic Forecasts and Budget Proposals. *Public Choice*, 76(3):215–232.

Blackwell, C. (2014). Power Law or Lognormal? Distribution of Normalized Hurricane Damages in the United States, 1900–2005. *Natural Hazards Review*, 16(3):04014024.

Blake, E. S., Rappaport, E. N., and Landsea, C. W. (2011). The Deadliest, Costliest, and Most Intense United States Tropical Cyclones From 1851 to 2010 (and Other Frequently Requested Hurricane Facts). NOAA Technical Memorandum NWS NHC-6, NOAA, National Weather Service, National Hurricane Center, Miami, Florida.

Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica*, 77(3):623–685.

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives*, 28(2):153–175.

Bowsher, C. G. and Meeks, R. (2008). The Dynamics of Economic Functions: Modeling and Forecasting the Yield Curve. *Journal of the American Statistical Association*, 103(484):1419–1437.

Bårdsen, G., den Reijer, A., Jonasson, P., and Nymoen, R. (2012). MOSES: Model for Studying the Economy of Sweden. *Economic Modelling*, 29(6):2566 – 2582.

Brayton, F., Laubach, T., and Reifschneider, D. (2014). Optimal-Control Monetary Policy in the FRB/US Model Board of Governors of the Federal Reserve System. FEDS Notes, Board of Governors of the Federal Reserve System, Washington, D.C.

Broad, K., Leiserowitz, A., Weinkle, J., and Steketee, M. (2007). Misinterpretations of the "Cone of Uncertainty" in Florida During the 2004 Hurricane Season. *Bulletin of the American Meteorological Society*, 88(5):651–667.

Burgess, S., Fernandez Corugedo, E., Groth, C., Harrison, R., Monti, F., Theodoridis, K., and Waldron, M. (2013). The Bank of England's Forecasting Platform: COMPASS, MAPS, EASE and the Suite of Models. Working Paper Series No. 471, The Bank of England, London, United Kingdom.

Cai, T. T., Liang, T., and Zhou, H. H. (2015). Law of Log Determinant of Sample Covariance Matrix and Optimal Estimation of Differential Entropy for High-Dimensional Gaussian Distributions. *Journal of Multivariate Analysis*, 137:161–172.

Cai, T. T., Liu, W., and Xia, Y. (2013). Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings. *Journal of the American Statistical Association*, 108(501):265–277.

Campbell, B. and Ghysels, E. (1995). Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency. *Review of Economics and Statistics*, 77(1):17–31.

Campos, J., Ericsson, N. R., and Hendry, D. F. (2005). *General-to-Specific Modelling*. The International Library of Critical Writings in Econometrics Series. Cheltenham: Edward Elgar.

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent Model Selection by an Automatic Gets Approach. *Oxford Bulletin of Economics and Statistics*, 65:803–819.

Cangialosi, J. P. and Franklin, J. L. (2016). 2015 Hurricane Season. NHC Forecast Verification Report, NOAA, National Hurricane Center, Miami, Florida.

Capistrán, C. (2006). On Comparing Multi-horizon Forecasts. *Economics Letters*, 93(2):176–181.

Carsell, K. M., Pingel, N. D., and Ford, D. T. (2004). Quantifying the Benefit of a Flood Warning System. *Natural Hazards Review*, 5(3):131–140.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Advanced Series. Pacific Grove, CA: Duxbury, second edition.

Castle, J. L. (2017). Sir Clive WJ Granger Model Selection. *European Journal of Pure and Applied Mathematics*, 10(1):133–156.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2011). Evaluating Automatic Model Selection. *Journal of Time Series Econometrics*, 3(1).

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2012). Model Selection When There Are Multiple Breaks. *Journal of Econometrics*, 169(2):239–246.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2018). Robust Model Selection. *mimeo*.

Castle, J. L., Doornik, J. A., Hendry, D. F., and Pretis, F. (2015). Detecting Location Shifts During Model Selection by Step-Indicator Saturation. *Econometrics*, 3(2):240–264.

Castle, J. L. and Hendry, D. F. (2008). Forecasting UK Inflation: The Roles of Structural Breaks and Time Disaggregation. In Rapach, D. E. and Wohar, M. E., editors, *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, volume 3 of *Frontiers of Economics and Globalization*, chapter 2, pages 41–92. Bingley, UK: Emerald Group.

Castle, J. L., Hendry, D. F., and Martinez, A. B. (2017). Evaluating Forecasts, Narratives and Policy Using a Test of Invariance. *Econometrics*, 5(3):39–65.

CBO (2002). CBO's Economic Forecasting Record: A Supplement to the Budget and Economic Oulook: Fiscal Years 2003-2012. Technical report, Congressional Budget Office, Washington, D.C.

CBO (2004). CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO Made from January 1976 through January 2002. Technical report, Congressional Budget Office, Washington, D.C.

CBO (2005). CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO Made from January 1976 through January 2003. Technical report, Congressional Budget Office, Washington, D.C.

CBO (2006). CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO Made from January 1976 through January 2004. Technical report, Congressional Budget Office, Washington, D.C.

CBO (2007). CBO's Economic Forecasting Record: 2007 Update. Publication No. 3042, Congressional Budget Office, Washington, D.C.

CBO (2009). CBO's Economic Forecasting Record: 2009 Update. Publication No. 3255, Congressional Budget Office, Washington, D.C.

CBO (2010). CBO's Economic Forecasting Record: 2010 Update. Publication No. 4138, Congressional Budget Office, Washington, D.C.

CBO (2013). CBO's Economic Forecasting Record: 2013 Update. Publication No. 4431, Congressional Budget Office, Washington, D.C.

CBO (2015a). Cbo's Economic Forecasting Record: 2015 Update. Publication No. 49891, Congressional Budget Office, Washington, D.C.

CBO (2015b). CBO's Revenue Forecasting Record. Publication No. 50831, Congressional Budget Office, Washington, D.C.

CBO (2017a). CBO's Economic Forecasting Record: 2017 Update. Publication No. 53090, Congressional Budget Office, Washington, D.C.

CBO (2017b). An Evaluation of CBO's Past Outlay Projections. Publication No. 53328, Congressional Budget Office, Washington, D.C.

Chavas, D. R., Reed, K. A., and Knaff, J. A. (2017). Physical Understanding of the Tropical Cyclone Wind-Pressure Relationship. *Nature Communications*, 8(1360).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics*, 7:649–88.

Cho, J. S. and Phillips, P. C. (2018). Pythagorean Generalization of Testing the Equality of Two Symmetric Positive Definite Matrices. *Journal of Econometrics*, 202(1):45–56.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric Evaluation of Linear Macro-Economic Models. *The Review of Economic Studies*, 53(4):671–690.

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3):591–605.

Christoffersen, P. F. and Diebold, F. X. (1998). Cointegration and Long-Horizon Forecasting. *Journal of Business & Economic Statistics*, 16(4):450–456.

Cichocki, A., Cruces, S., and Amari, S.-i. (2015). Log-Determinant Divergences Revisited: Alpha-Beta and Gamma Log-Det Divergences. *Entropy*, 17(5):2988–3034.

Clark, T. E. and McCracken, M. W. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics*, 105(1):85–110.

Clark, T. E. and McCracken, M. W. (2013a). Advances in Forecast Evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2B, chapter 20, pages 1107–1201. Amsterdam: Elsevier.

Clark, T. E. and McCracken, M. W. (2013b). Evaluating the Accuracy of Forecasts From Vector Autoregressions. In Fomby, T. B., Lutz Kilian, and Anthony Murphy, editors, *VAR Models in Macroeconomics-New Developments and Applications: Essays in Honor of Christopher A. Sims*, volume 32 of *Advances in Econometrics*, chapter 4, pages 117–168. Bingley, UK: Emerald Group.

Clark, T. E., McCracken, M. W., and Mertens, E. (2019). Modeling Time-Varying Uncertainty of Multiple-Horizon Forecast Errors. *Review of Economics and Statistics*, forthcoming.

Clements, M. P. (2014). Forecast Uncertainty–Ex Ante and Ex Post: U.S. Inflation and Output Growth. *Journal of Business & Economic Statistics*, 32(2):206–216.

Clements, M. P. and Galvão, A. B. (2004). A Comparison of Tests of Nonlinear Cointegration with Application to the Predictability of U.S. Interest Rates Using the Term Structure. *International Journal of Forecasting*, 20(2):219 – 236.

Clements, M. P. and Hendry, D. F. (1993). On the Limitations of Comparing Mean Square Forecast Errors. *Journal of Forecasting*, 12(8):617–637.

Clements, M. P. and Hendry, D. F. (1995). Forecasting in Cointegrated Systems. *Journal of Applied Econometrics*, 10(2):127–146.

Clements, M. P. and Hendry, D. F. (1997). An Empirical Study of Seasonal Unit Roots in Forecasting. *International Journal of Forecasting*, 13(3):341 – 355.

Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M. P. and Hendry, D. F. (2002). An Overview of Economic Forecasting. In Clements, M. P. and Hendry, D. F., editors, *A Companion to Economic Forecasting*, Blackwell Companions to Contemporary Economics, chapter 1, pages 1–18. Oxford: Blackwell Publishing Ltd.

Cohen, D. and Follette, G. (2003). Forecasting Exogenous Fiscal Variables in the United States. FEDS Working Paper No. 2003-59, Board of Governors of the Federal Reserve System, Washington, D.C.

Corder, J. K. (2005). Managing Uncertainty: The Bias and Efficiency of Federal Macroeconomic Forecasts. *Journal of Public Administration Research and Theory*, 15(1):55–70.

Croushore, D. and Van Norden, S. (2017). Fiscal Surprises At the FOMC. CIRANO Working Paper No. 2017s-09, CIRANO.

Croushore, D. and Van Norden, S. (2018). Fiscal Forecasts At the FOMC: Evidence From the Greenbooks. *Review of Economics and Statistics*, 100(5):933–945.

Davlasheridze, M., Fisher Vanden, K., and Klaiber, H. A. (2017). The Effects of Adaptation Measures on Hurricane Induced Property Losses: Which FEMA Investments Have the Highest Returns? *Journal of Environmental Economics and Management*, 81:93–114.

De Gooijer, J. G. and Ray, B. K. (2003). Modeling Vector Nonlinear Time Series Using POLYMARS. *Computational Statistics & Data Analysis*, 42(1–2):73 – 90.

De Gooijer, J. G. and Vidiella-i-Anguera, A. (2004). Forecasting Threshold Cointegrated Systems. *International Journal of Forecasting*, 20(2):237 – 253.

Dehring, C. A. and Halek, M. (2013). Coastal Building Codes and Hurricane Damage. *Land Economics*, 89(4):597–613.

Del Negro, M. and Schorfheide, F. (2004). Priors From General Equilibrium Models for VARs. *International Economic Review*, 45(2):643–673.

Del Negro, M. and Schorfheide, F. (2013). DSGE Model-based Forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2A, chapter 2, pages 57–140. Amsterdam: Elsevier.

Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007). On the Fit of New Keynesian Models. *Journal of Business & Economic Statistics*, 25(2):123–143.

DeMaria, M., Knaff, J., Knabb, R., Lauer, C., Sampson, C., and DeMaria, R. (2009). A New Method for Estimating Tropical Cyclone Wind Speed Probabilities. *Weather and Forecasting*, 24(6):1573–1591.

Deryugina, T. (2017). The Fiscal Cost of Hurricanes: Disaster Aid Versus Social Insurance. *American Economic Journal: Economic Policy*, 9(3):168–198.

Deryugina, T., Kawano, L., and Levitt, S. (2018). The Economic Impact of Hurricane Katrina on Its Victims: Evidence From Individual Tax Returns. *American Economic Journal: Applied Economics*, 10(2):202–233.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and Conditional Projection Using Realistic Prior Distributions. *Econometric Reviews*, 3(1):1–100.

Doornik, J. A. (2009). Autometrics. In Shephard, N. and Castle, J. L., editors, *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, chapter 4, pages 88–121. Oxford: Oxford University Press.

Doornik, J. A. (2013). *Ox 7: An Object-Oriented Matrix Programming Language*. London: Timberlake Consultants Ltd.

Doornik, J. A. and Hansen, H. (2008). An Omnibus Test for Univariate and Multivariate Normality. *Oxford Bulletin of Economics and Statistics*, 70(s1):927–939.

Doornik, J. A. and Hendry, D. F. (2013a). *Empirical Econometric Modelling*, volume I of *PcGive 14*. London: Timberlake Consultants Ltd.

Doornik, J. A. and Hendry, D. F. (2013b). *Modelling Dynamic Systems*, volume II of *PcGive 14*. London: Timberlake Consultants Ltd.

Edge, R. M., Kiley, M. T., and Laforte, J.-P. (2008). Natural Rate Measures in an Estimated DSGE Model of the U.S. Economy. *Journal of Economic Dynamics and Control*, 32(8):2512–2535.

Emanuel, K. (2001). Contribution of Tropical Cyclones to Meridional Heat Transport by the Oceans. *Journal of Geophysical Research: Atmospheres*, 106(D14):14771–14781.

Emanuel, K. (2005). Increasing Destructiveness of Tropical Cyclones Over the Past 30 Years. *Nature*, 436(7051):686–688.

Emanuel, K. (2017a). Assessing the Present and Future Probability of Hurricane Harvey's Rainfall. *Proceedings of the National Academy of Sciences*, 114(48):12681–12684.

Emanuel, K. (2017b). Will Global Warming Make Hurricane Forecasting More Difficult? *Bulletin of the American Meteorological Society*, 98(3):495–501.

Engle, R. F. (1984). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Griliches, Z. and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume 2, chapter 13, pages 775–826. New York, NY: North Holland.

Engle, R. F. (1993). On the Limitations of Comparing Mean Square Forecast Errors: Comment. *Journal of Forecasting*, 12(8):642–644.

Engle, R. F. and Hendry, D. F. (1993). Testing Superexogeneity and Invariance in Regression Models. *Journal of Econometrics*, 56(1-2):119–139.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2):277–304.

Ericsson, N. R. (1992). Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration. *Journal of Policy Modeling*, 14(4):465–495.

Ericsson, N. R. (2001). Forecast Uncertainty in Economic Modeling. In Hendry, D. F. and Neil R. Ericsson, editors, *Understanding Economic Forecasts*, chapter 5, pages 68–92. MIT Press.

Ericsson, N. R. (2017). How Biased Are U.S. Government Forecasts of the Federal Debt? *International Journal of Forecasting*, 33(2):543–559.

Estrada, F., Botzen, W. W., and Tol, R. S. (2015). Economic Losses From U.S. Hurricanes Consistent with an Influence From Climate Change. *Nature Geoscience*, 8:880–884.

Faust, J. and Wright, J. H. (2009). Comparing Greenbook and Reduced Form Forecasts Using a Large Realtime Dataset. *Journal of Business & Economic Statistics*, 27(4):468–479.

Faust, J. and Wright, J. H. (2013). Forecasting Inflation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2A, chapter 1, pages 3–56. Amsterdam: Elsevier.

Frendreis, J. and Tatalovich, R. (2000). Accuracy and Bias in Macroeconomic Forecasting by the Administration, the CBO, and the Federal Reserve Board. *Polity*, 32(4):623–632.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models Via Coordinate Descent. *Journal of Satistical Software*, 33(1):1–22.

Fuhrer, J. C. (1997). Inflation/output Variance Trade-offs and Optimal Monetary Policy. *Journal of Money, Credit, and Banking*, pages 214–234.

Fujikoshi, Y. (1968). Asymptotic Expansion of the Distribution of the Generalized Variance in the Non-central Case. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 32(2):293–299.

Gallagher, J. and Hartley, D. (2017). Household Finance After a Natural Disaster: The Case of Hurricane Katrina. *American Economic Journal: Economic Policy*, 9(3):199–228.

Ganics, G. A. (2017). Optimal Density Forecast Combinations. Documentos De Trabajo No. 1751, Banco De España, Madrid, Spain.

Geiger, T., Frieler, K., and Levermann, A. (2016). High-income Does Not Protect Against Hurricane Losses. *Environmental Research Letters*, 11(8):084012.

Gelper, S. and Croux, C. (2007). Multivariate Out-of-sample Tests for Granger Causality. *Computational Statistics & Data Analysis*, 51(7):3319–3329.

Geweke, J. and Amisano, G. (2011). Optimal Prediction Pools. *Journal of Econometrics*, 164(1):130–141.

Ghodsi, M., Alharbi, N., and Hassani, H. (2015). The Empirical Distribution of the Singular Values of a Random Hankel Matrix. *Fluctuation and Noise Letters*, 14(03):1550027.

Giacomini, R. and Rossi, B. (2010). Forecast Comparisons in Unstable Environments. *Journal of Applied Econometrics*, 25(4):595–620.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Granger, C. W. J. (1993). On the Limitations of Comparing Mean Square Forecast Errors: Comment. *Journal of Forecasting*, 12(8):651–652.

Granger, C. W. J. (1999). Outline of Forecast Theory Using Generalized Cost Functions. *Spanish Economic Review*, 1(2):161–173.

Granger, C. W. J. and Jeon, Y. (2004). Thick Modeling. *Economic Modelling*, 21(2):323–343.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2):453–497.

Hansen, P. R. and Timmermann, A. (2015). Equivalence Between Out-of-sample Forecast Comparisons and Wald Statistics. *Econometrica*, 83(6):2485–2505.

Hendry, D. F. (2006). Robustifying Forecasts From Equilibrium-Correction Systems. *Journal of Econometrics*, 135(1-2):399–426.

Hendry, D. F. (2018). First-in, First-out: Modelling the UK's CO_2 Emissions, 1960-2016. *mimeo*.

Hendry, D. F. and Doornik, J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. Arne Ryde Memorial Lectures Series. Cambridge, MA: MIT Press.

Hendry, D. F. and Johansen, S. (2015). Model Discovery and Trygve Haavelmo's Legacy. *Econometric Theory*, 31(01):93–114.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic Selection of Indicators in a Fully Saturated Regression. *Computational Statistics*, 23(2):317–335.

Hendry, D. F. and Martinez, A. B. (2017). Evaluating Multi-step System Forecasts with Relatively Few Forecast-error Observations. *International Journal of Forecasting*, 33(2):359 – 372.

Hendry, D. F. and Mizon, G. E. (2011). Econometric Modelling of Time Series with Outlying Observations. *Journal of Time Series Econometrics*, 3(1).

Hendry, D. F. and Mizon, G. E. (2014). Unpredictability in Economic Analysis, Econometric Modeling and Forecasting. *Journal of Econometrics*, 182(1):186–195.

Hendry, D. F. and Muellbauer, J. N. J. (2018). The Future of Macroeconomics: Macro Theory and Models At the Bank of England. *Oxford Review of Economic Policy*, 34(1-2):287–328.

Hendry, D. F. and Santos, C. (2005). Regression Models with Data-based Indicator Variables. *Oxford Bulletin of Economics and Statistics*, 67(5):571–595.

Hendry, D. F. and Santos, C. (2010). An Automatic Test of Super Exogeneity. In Watson, M. W., Bollerslev, T., and Russell, J., editors, *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Advanced Texts in Econometrics, chapter 12, pages 164–193. Oxford: Oxford University Press.

Hinkley, D. (1979). Predictive Likelihood. *Annals of Statistics*, 7(4):718–728.

Hoffman, D. L. and Rasche, R. H. (1996). Assessing Forecast Performance in a Cointegrated System. *Journal of Applied Econometrics*, 11(5):495–517.

Howard, J. A. (1987). Government Economic Projections: A Comparison Between CBO and OMB Forecasts. *Public Budgeting & Finance*, 7(3):14–25.

Hsiang, S. M. and Narita, D. (2012). Adaptation to Cyclone Risk: Evidence From the Global Cross-Section. *Climate Change Economics*, 3(02):1250011.

Hungnes, H. (2018). Encompassing Tests for Evaluating Multi-Step System Forecasts Invariant to Linear Transformations. Discussion Paper No. 871, Statistics Norway, Research Department, Oslo, Norway.

Huntley, J. and Miller, E. (2009). An Evaluation of CBO Forecasts. Working Paper Series No. 2009-02, Congressional Budget Office, Washington, D.C.

Iliopoulos, G. (2008). UMVU Estimation of the Ratio of Powers of Normal Generalized Variances Under Correlation. *Journal of Multivariate Analysis*, 99(6):1051–1069.

Jafari, A. A. (2012). Inferences on the Ratio of Two Generalized Variances: Independent and Correlated Cases. *Statistical Methods & Applications*, 21(3):297–314.

Jarrell, J. D., Hebert, P. J., and Mayfield, M. (1992). Hurricane Experience Levels of Coastal County Populations From Texas to Maine. NOAA Technical Memorandum NWS NHC-46, NOAA, National Weather Service, National Hurricane Center, Miami, Florida.

Jentsch, C. and Politis, D. N. (2015). Covariance Matrix Estimation and Linear Process Bootstrap for Multivariate Time Series of Possibly Increasing Dimension. *Aannals of Statistics*, 43(3):1117–1140.

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59(6):1551–1580.

Johansen, S. and Juselius, K. (1990). Maximum Likelihood Estimation and Inference on Cointegration — With Applications to the Demand for Money. *Oxford Bulletin of Economics and Statistics*, 52(2):169–210.

Johansen, S. and Nielsen, B. (2009). An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator. In Castle, J. L. and Shephard, N., editors, *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, chapter 1, pages 1–36. Oxford: Oxford University Press.

Johansen, S. and Nielsen, B. (2016). Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models. *Scandinavian Journal of Statistics*, 43(2):321–348.

Jordà, Ò., Knüppel, M., and Marcellino, M. (2013). Empirical Simultaneous Prediction Regions for Path-forecasts. *International Journal of Forecasting*, 29(3):456–468.

Jordà, Ò. and Marcellino, M. (2010). Path Forecast Evaluation. *Journal of Applied Econometrics*, 25(4):635–662.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring Uncertainty. *American Economic Review*, 105(3):1177–1216.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2):263–292.

Kamlet, M. S., Mowery, D. C., and Su, T.-T. (1987). Whom Do You Trust? an Analysis of Executive and Congressional Economic Forecasts. *Journal of Policy Analysis and Management*, 6(3):365–384.

Kaplan, J., DeMaria, M., and Knaff, J. A. (2010). A Revised Tropical Cyclone Rapid Intensification Index for the Atlantic and Eastern North Pacific Basins. *Weather and Forecasting*, 25(1):220–241.

Katz, R. W. and Lazo, J. K. (2011). Economic Value of Weather and Climate Forecasts. In Clements, M. P. and David F. Hendry, editors, *Oxford Handbook on Economic Forecasting*.

Kellenberg, D. K. and Mobarak, A. M. (2008). Does Rising Income Increase Or Decrease Damage Risk From Natural Disasters? *Journal of Urban Economics*, 63(3):788–802.

Kidder, S. Q., Knaff, J. A., Kusselson, S. J., Turk, M., Ferraro, R. R., and Kuligowski, R. J. (2005). The Tropical Rainfall Potential (TRaP) Technique. Part I: Description and Examples. *Weather and Forecasting*, 20(4):456–464.

Kitchen, J. (2003). Observed Relationships Between Economic and Technical Receipts Revisions in Federal Budget Projections. *National Tax Journal*, 56(2):337–353.

Kliesen, K. L. and Thornton, D. L. (2001). The Expected Federal Budget Surplus: How Much Confidence Should the Public and Policymakers Place in the Projections? *Federal Reserve Bank of St. Louis Review*, 83(2):11–24.

Kliesen, K. L. and Thornton, D. L. (2012). How Good Are the Government's Deficit and Debt Projections and Should We Care? *Federal Reserve Bank of St. Louis Review*, 94(1):21–39.

Knight, F. H. (1921). *Risk, Uncertainty and Profit*. Boston, MA: Houghton Mifflin.

Knüppel, M. (2018). Forecast-error-based Estimation of Forecast Uncertainty When the Horizon is Increased. *International Journal of Forecasting*, 34(1):105–116.

Knutson, T. R., McBride, J. L., Chan, J., Emanuel, K., Holland, G., Landsea, C., Held, I., Kossin, J. P., Srivastava, A., and Sugi, M. (2010). Tropical Cyclones and Climate Change. *Nature Geoscience*, 3(3):157–163.

Komunjer, I. and Owyang, M. T. (2012). Multivariate Forecast Evaluation and Rationality Testing. *Review of Economics and Statistics*, 94(4):1066–1080.

Krause, G. A. and Douglas, J. W. (2005). Institutional Design Versus Reputational Effects on Bureaucratic Performance: Evidence From Us Government Macroeconomic and Fiscal Projections. *Journal of Public Administration Research and Theory*, 15(2):281–306.

Krause, G. A. and Douglas, J. W. (2006). Does Agency Competition Improve the quality of Policy Analysis? Evidence From OMB and CBO Fiscal Projections. *Journal of Policy Analysis and Management*, 25(1):53–74.

Krol, R. (2014). Forecast Bias of Government Agencies. *Cato Journal*, 34(1):99–112.

Kruttli, M., Roth Tran, B., and Watugala, S. W. (2019). Pricing Poseidon: Extreme Weather Uncertainty and Firm Return Dynamics. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3284517.

Krzysztofowicz, R. and Davis, D. R. (1983). A Methodology for Evaluation of Flood Forecast-Response Systems. *Water Resources Research*, 19(6):1423–1454.

Kunst, R. M. and Franses, P. H. (1998). The Impact of Seasonal Constants on Forecasting Seasonally Cointegrated Time Series. *Journal of Forecasting*, 17(2):109–124.

Laurent, S., Rombouts, J. V., and Violante, F. (2013). On Loss Functions and Ranking Forecasting Performances of Multivariate Volatility Models. *Journal of Econometrics*, 173(1):1–10.

Laurent, S., Rombouts, J. V. K., and Violante, F. (2012). On the Forecasting Accuracy of Multivariate GARCH Models. *Journal of Applied Econometrics*, 27(6):934–955.

Lazarus, E., Lewis, D. J., and Stock, J. H. (2017). The Size-Power Tradeoff in HAR Inference. Working Paper No. 1-60, Harvard University, Cambridge, MA.

Letson, D., Sutter, D. S., and Lazo, J. K. (2007). Economic Value of Hurricane Forecasts: An Overview and Research Needs. *Natural Hazards Review*, 8(3):78–86.

Li, J. and Chen, S. X. (2012). Two Sample Tests for High-Dimensional Covariance Matrices. *The Annals of Statistics*, 40(2):908–940.

Lipford, J. W. (2001). How Transparent is the U.S. Budget? *The Independent Review*, 5(4):575–591.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Martinez, A. B. (2011). Comparing Government Forecasts of the United States' Gross Federal Debt. Research Program on Forecasting Working Paper No. 2011-002, The George Washington University, Washington, D.C.

Martinez, A. B. (2015). How Good Are U.S. Government Forecasts of the Federal Debt? *International Journal of Forecasting*, 31(2):312–324.

Mathiasen, P. E. (1979). Prediction Functions. *Scandinavian Journal of Statistics*, 6(1):1–21.

McAlinn, K., Aastveit, K., Nakajima, J., and West, M. (2019). Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting. Working Paper No. 2/2019, Norges Bank, Oslo, Norway.

McMurry, T. L. and Politis, D. N. (2010). Banded and Tapered Estimates for Autocovariance Matrices and the Linear Process Bootstrap. *Journal of Time Series Analysis*, 31(6):471–482.

McNees, S. K. (1995). An Assessment of the "Official" Economic Forecasts. *New England Economic Review*, (July/August):13–24.

Milch, K., Broad, K., Orlove, B., and Meyer, R. (2018). Decision Science Perspectives on Hurricane Vulnerability: Evidence From the 2010–2012 Atlantic Hurricane Seasons. *Atmosphere*, 9(1):32.

Miller, S. M. (1991). Forecasting Federal Budget Deficits: How Reliable Are U.S. Congressional Budget Office Projections? *Applied Economics*, 23(12):1789–1799.

Mitchell, J. and Hall, S. G. (2005). Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation. *Oxford Bulletin of Economics and Statistics*, 67(s1):995–1033.

Mitchell, J. and Wallis, K. F. (2011). Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness. *Journal of Applied Econometrics*, 26(6):1023–1040.

Mizon, G. E. and Richard, J.-F. (1986). The Encompassing Principle and its Application to Testing Non-Nested Hypotheses. *Econometrica*, 54(3):657–678.

Montiel Olea, J. L. and Plagborg Møller, M. (2017). Simultaneous Confidence Bands: Theoretical Comparisons and Suggestions for Practice. Working Paper, Harvard University, Cambridge, MA.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Müller, U. K. (2014). HAC Corrections for Strongly Autocorrelated Time Series. *Journal of Business & Economic Statistics*, 32(3):311–322.

MunichRe (2018). NatCatSERVICE. `https://natcatservice.munichre.com/`. [last Accessed 27-September-2018].

Murnane, R. J. and Elsner, J. B. (2012). Maximum Wind Speeds and U.S. Hurricane Losses. *Geophysical Research Letters*, 39(16).

National Science Board (2007). Hurricane Warning: The Critical Need for a National Hurricane Research Initiative. Technical Report NSB-06-115, National Science Foundation, Arlington, Virginia.

Neumayer, E. and Barthel, F. (2011). Normalizing Economic Loss From Natural Disasters: A Global Analysis. *Global Environmental Change*, 21(1):13–24.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Mmatrix. *Econometrica*, 55(3):703–708.

NOAA NCEI (2018). U.S. Billion-Dollar Weather and Climate Disasters. `https://www.ncdc.noaa.gov/billions/`. [last Accessed 27-September-2018].

Nordhaus, W. D. (2010). The Economics of Hurricanes and Implications of Global Warming. *Climate Change Economics*, 1(1):1–20.

Paap, R., Franses, P. H., and Hoek, H. (1997). Mean Shifts, Unit Roots and Forecasting Seasonal Time Series. *International Journal of Forecasting*, 13(3):357 – 368.

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J. (2015). The Monetary Benefit of Early Flood Warnings in Europe. *Environmental Science & Policy*, 51:278–291.

Patton, A. J. (2011). Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. J. and Timmermann, A. (2012). Forecast Rationality Tests Based on Multi-Horizon Bounds. *Journal of Business & Economic Statistics*, 30(1):1–17.

Penner, R. G. (2001). Errors in Budget Forecasting. Technical report, The Urban Institute, Washington, D.C.

Penner, R. G. (2002). Dealing with Uncertain Budget Forecasts. *Public Budgeting & Finance*, 22(1):1–18.

Penner, R. G. (2008). Federal Revenue Forecasting. In Sun, J. and Lynch, T. D., editors, *Government Budget Forecasting, Theory and Practice*, chapter 2, pages 31–46. CRC Press.

Pielke Jr, R. A., Gratz, J., Landsea, C. W., Collins, D., Saunders, M. A., and Musulin, R. (2008). Normalized Hurricane Damage in the United States: 1900–2005. *Natural Hazards Review*, 9(1):29–42.

Pielke Jr, R. A. and Landsea, C. W. (1998). Normalized Hurricane Damages in the United States: 1925-95. *Weather and Forecasting*, 13(3):621–631.

Plesko, G. A. (1988). The Accuracy of Government Forecasts and Budget Projections. *National Tax Journal*, 41(4):483–501.

Pretis, F., Reade, J., and Sucarrat, G. (2018). Automated General-to-Specific (GETS) Regression Modeling and Indicator Saturation for Outliers and Structural Breaks. *Journal of Statistical Software*, 86(1):1–44.

Pretis, F. and Roser, M. (2017). Carbon Dioxide Emission-Intensity in Climate Projections: Comparing the Observational Record to Socio-Economic Scenarios. *Energy*, 135:718–725.

Pretis, F., Schneider, L., Smerdon, J. E., and Hendry, D. F. (2016). Detecting Volcanic Eruptions in Temperature Reconstructions by Designed Break-Indicator Saturation. *Journal of Economic Surveys*, 30(3):403–429.

Proietti, T. and Giovannelli, A. (2018). A Durbin–Levinson Regularized Estimator of High-Dimensional Autocovariance Matrices. *Biometrika*, 105(4):783–795.

Quaedvlieg, R. (2019). Multi-Horizon Forecast Comparison. *Journal of Business & Economic Statistics*, forthcoming.

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 350–371.

Rappaport, E. N., Franklin, J. L., Avila, L. A., Baig, S. R., Beven, J. L., Blake, E. S., Burr, C. A., Jiing, J.-G., Juckins, C. A., and Knabb, R. D. (2009). Advances and Challenges At the National Hurricane Center. *Weather and Forecasting*, 24(2):395–419.

Rasche, R. H. (1987). M1 — Velocity and Money-Demand Functions: Do Stable Relationships Exist? *Carnegie-Rochester Conference Series on Public Policy*, 27(0):9 – 88.

Ravn, M. O. and Uhlig, H. (2002). On Adjusting the Hodrick-Prescott Filter for the Frequency of Observations. *The Review of Economics and Statistics*, 84(2):371–376.

Regnier, E. (2008). Public Evacuation Decisions and Hurricane Track Uncertainty. *Management Science*, 54(1):16–28.

Resio, D. T., Powell, N. J., Cialone, M. A., Das, H. S., and Westerink, J. J. (2017). Quantifying Impacts of Forecast Uncertainties on Predicted Storm Surges. *Natural Hazards*, 88(3):1423–1449.

Rossi, B. and Sekhposyan, T. (2015). Macroeconomic Uncertainty Indices Based on Nowcast and Forecast Error Distributions. *American Economic Review: Papers and Proceedings*, 105(5):650–655.

Rossi, B., Sekhposyan, T., and Soupre, M. (2018). Understanding the Sources of Macroeconomic Uncertainty. Working Paper No. 920, Barcelona GSE, Barcelona, Spain.

Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*, 8:283–297.

Sadowski, N. C. and Sutter, D. (2005). Hurricane Fatalities and Hurricane Damages: Are Safer Hurricanes More Damaging? *Southern Economic Journal*, 72(2):422–432.

Schmidt, P. (1974). The Asymptotic Distribution of Forecasts in the Dynamic Simulation of an Econometric Model. *Econometrica*, 42(2):303–309.

Schmidt, P. (1993). On the Limitations of Comparing Mean Square Forecast Errors: Comment. *Journal of Forecasting*, 12(8):660–662.

Schorfheide, F. and Song, D. (2015). Real-Time Forecasting with a Mixed-Frequency VAR. *Journal of Business & Economic Statistics*, 33(3):366–380.

Sheets, R. C. (1990). The National Hurricane Center–Past, Present, and Future. *Weather and Forecasting*, 5(2):185–232.

Shuman, F. G. (1989). History of Numerical Weather Prediction At the National Meteorological Center. *Weather and Forecasting*, 4(3):286–296.

Silvester, J. R. (2000). Determinants of Block Matrices. *Mathematical Gazette*, 84(501):460–467.

Simionescu, M. (2013). The Use of VARMA Models in Forecasting Macroeconomic Indicators. *Economics & Sociology*, 6(2):94–102.

Sinclair, T. M., Stekler, H. O., and Warren Carnow (2012). A New Approach for Evaluating Economic Forecasts. *Economics Bulletin*, 32(3):2332–2342.

Sinclair, T. M., Stekler, H. O., and Warren Carnow (2015). Evaluating a Vector of the Fed's Forecasts. *International Journal of Forecasting*, 31(1):157 – 164.

Smets, F. and Wouters, R. (2007). Shocks and Frictions in U.S. Business Cycles: A Bayesian DSGE Approach. *American Economic Review*, 97(3):586–606.

Smith, A. B. and Katz, R. W. (2013). US Billion-dollar Weather and Climate Disasters: Data Sources, Trends, Accuracy and Biases. *Natural Hazards*, 67(2):387–410.

Staszewska-Bystrova, A. (2011). Bootstrap Prediction Bands for Forecast Paths From Vector Autoregressive Models. *Journal of Forecasting*, 30(8):721–735.

Stewart, K. G. (1995). The Functional Equivalence of the W, LR, and LM Statistics. *Economics Letters*, 49(2):109–112.

Strobl, E. (2011). The Economic Growth Impact of Hurricanes: Evidence From U.S. Coastal Counties. *Review of Economics and Statistics*, 93(2):575–589.

Tsuchiya, Y. (2016). Directional Analysis of Fiscal Sustainability: Revisiting Domar's Debt Sustainability Condition. *International Review of Economics & Finance*, 41:189–201.

U.S. Army Corps of Engineers (2004). Hurricane Assessment Concerns and Recommendations. Available Online At: `https://web.archive.org/web/20090726033748/http://chps.sam.usace.army.mil/USHESdata/Assessments/2004Storms/2004-{R}ecommendations.htm` (last accessed December 22, 2017).

Vahid, F. and Issler, J. V. (2002). The Importance of Ccommon Cyclical Features in VAR Analysis: A Monte-Carlo Study. *Journal of Econometrics*, 109(2):341 – 363.

Van De Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., and Others (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *Annals of Statistics*, 42(3):1166–1202.

van Den Dool, H., Huang, J., and Fan, Y. (2003). Performance and Analysis of the Constructed Analogue Method Applied to U.S. Soil Moisture Over 1981–2001. *Journal of Geophysical Research: Atmospheres*, 108(D16):8617.

Veloce, W. (2004). Forecasting Inbound Canadian Tourism: An Evaluation of Error Corrections Model Forecasts. *Tourism Economics*, 10(3):263–280.

Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, 23(176):88–93.

Von Neumann, J. and Morgenstern, O. (1945). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333.

Wallis, K. F. (1993). On the Limitations of Comparing Mean Square Forecast Errors: Comment. *Journal of Forecasting*, 12(8):663–666.

West, K. D. (1996). Asymptotic Inference About Predictive Ability. *Econometrica*, 64(5):1067–1084.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838.

Willoughby, H. E. (2012). Distributions and Trends of Death and Destruction From Hurricanes in the United States, 1900–2008. *Natural Hazards Review*, 13(1):57–64.

Willoughby, H. E., Rappaport, E., and Marks, F. (2007). Hurricane Forecasting: The State of the Art. *Natural Hazards Review*, 8(3):45–49.

Wolf, M. and Wunderli, D. (2015). Bootstrap Joint Prediction Regions. *Journal of Time Series Analysis*, 36(3):352–376.

Wolters, M. H. (2015). Evaluating Point and Density Forecasts of DSGE Models. *Journal of Applied Econometrics*, 30(1):74–96.

Yellen, J. L. (2012). Perspectives on Monetary Policy. Remarks by Vice Chair Janet L. Yellen At the Boston Economic Club Dinner, Boston, Massachussetts. Available Online At: `https://www.federalreserve.gov/newsevents/speech/yellen20120606a.pdf` [last Accessed: 16th May, 2018], Board of Governors of the Federal Reserve System, Washington, D.C.

Zeng, T. and Swanson, N. R. (1998). Predictive Evaluation of Econometric Forecasting Models in Commodity Futures Markets. *Studies in Nonlinear Dynamics & Econometrics*, 2(4):159–177.