A SIMPLE PEDAGOGICAL MODEL LINKING INITIAL-VALUE RELIABILITY WITH TRUSTWORTHINESS IN THE FORCED CLIMATE RESPONSE

T. N. Palmer and A. Weisheimer

Using a simple pedagogical model, it is shown how information about the statistical reliability of initial-value ensemble forecasts can be relevant in assessing the trustworthiness of the climate system's response to forcing.

Understanding and modeling the response of the climate system to external forcings, especially to anthropogenic forcing, are central to climate science. Typically, projections of coupled general circulation models (GCMs) are made that simulate a future world using assumptions, or scenarios, of how such forcings will evolve over time. A related methodology is pursued in attributing extreme weather and climate events to anthropogenic forcing. Here, the response of the current climate system to the ongoing anthropogenic forcing is estimated by comparing model states with and without anthropogenic forcing.

Regional climate projections are increasingly providing the basis for climate-related decision-making in a range of societal sectors. Hence, the degree to which the GCMs' forced responses are trustworthy is becoming an issue of practical as well as theoretical importance. Given the continuing significant biases in regional simulations of key climate variables (IPCC 2013), this trustworthiness cannot be guaranteed.

For example, the frequency of anticyclonic blocking is projected to decrease in most climate models as a result of anthropogenic forcing (Matsueda et al. 2009; Masato et al. 2013). However, the frequency of blocking is severely underestimated in phase 5 of the

AFFILIATIONS: Palmer—Department of Physics, National Centre for Atmospheric Science, Atmospheric, Oceanic, and Planetary Physics, University of Oxford, Oxford, United Kingdom; Weisheimer—Department of Physics, National Centre for Atmospheric Science, Atmospheric, Oceanic, and Planetary Physics, University of Oxford, Oxford, and European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
CORRESPONDING AUTHOR: Tim Palmer, tim.palmer@physics.ox.ac.uk

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-16-0240.1

In final form 3 October 2017

Coupled Model Intercomparison Project (CMIP5) twentieth-century integrations (Anstey et al. 2013) and continues to be underestimated in high-resolution simulations (Schiemann et al. 2017). In the presence of such shortcomings, how trustworthy are projections of blocking frequency? Since anticyclonic blocking is a vital factor in determining the frequency of drought throughout the year, and cold weather in winter, answering the question above is of considerable importance.

Other examples where substantial model biases could lead to untrustworthiness of model-derived regional climate projections are the Asian summer monsoon (e.g., Webster et al. 1998; Turner and Annamalai 2012; Ramesh and Goswami 2014) and the decreasing Artic sea ice and its link to the global atmospheric circulation (e.g., Wettstein and Deser 2014; Francis and Skific 2015; Barnes and Screen 2015).

By definition, we cannot know for certain whether our current climate change projections are untrustworthy until these changes eventually occur in the future and can be verified with observations. However, we build trust in climate models by critically evaluating their performance in present-day or past climate conditions. This is typically achieved by assessing characteristics of the simulated climatological probability distributions, for example, first- and higher-order moments, and patterns of climatological spectral variability.

As will be shown below, these characteristics do not guarantee the trustworthiness of the forced response. Are there other diagnostics that can help determine the trustworthiness of regional climate projections or attributions of observed extreme weather events? One possible diagnostic is the statistical reliability of initial-value ensemble forecasts (Wilks 2011; Weisheimer and Palmer 2014) that relate forecast probability with frequency of occurrence. The possible link between initial-value reliability and the trustworthiness of the forced response was first proposed by Palmer et al. (2008, hereafter P08), who discussed how information from a multimodel ensemble of initialized coupled seasonal forecasts can be used to constrain the trustworthiness of the regional projection of precipitation. However, this notion has proved controversial. For example, Scaife et al (2009) argued that since initial-value predictions (on the seasonal time scale or shorter) merely provide estimates of the future internal variability of the climate system, they are largely irrelevant in assessing the trustworthiness of the long-term forced-response signal. Similarly, in their recent review paper, Stott et al. (2016) discuss, but largely discount, the use of initialized seasonal forecast reliability diagrams for assessing the ability of a particular climate model to be used for extreme event attribution. Stott et al. (2016, p. 32) comment: "A seasonal forecast reliability diagram indicates whether the model is able to capture the predictable features of the event under consideration. Although the use of reliability is well established for forecasting, its meaning for attribution is less clear given that reliable attribution is still possible when there is no inherent real-world predictability."

We will return to this comment below. However, for now it can be noted that Matsueda et al. (2016) have tested the P08 hypothesis based on a series of high- and low-resolution integrations of a comprehensive atmospheric model, where the high resolution is treated as a surrogate of truth. They showed quantitatively that information about the reliability of seasonal forecast ensembles can help improve the skill of regional climate change projections of precipitation.

Although this latter study lends some support to the P08 philosophy, it is clear that a conceptual picture of how initial-value unreliability can undermine trust in the climate-forced response is lacking. In this paper, we present such a picture using a simple to understand nonlinear toy model, qualitatively analogous to the extratropical atmosphere. We show how the failure to produce a trustworthy response to external forcing can be clearly diagnosed from the model's initial-condition forecast unreliability.

The pedagogical model is configured in two different ways. One configuration, the more complex, is defined as REALITY, and the other, simpler, configuration defines our weather/climate MODEL. An external forcing, representing, for example, anthropogenic forcing, is then applied to both configurations. The MODEL is deliberately constructed in such a way that its response to the forcing is completely incorrect compared with REALITY. We use this setup to investigate whether and how it is possible to know if the MODEL response to an external forcing is untrustworthy without directly comparing the forced MODEL with its corresponding forced REALITY (which corresponds to the real-world situation where the forced reality will be known only in the future). It is demonstrated in the conceptual model that the untrustworthiness of the forced response is not only diagnosable from, but it is also dynamically linked with the unreliability of the initial-value forecasts. By contrast, the untrustworthiness of the forced response could not be determined by simple diagnostics of biases in the MODEL's unforced climatology.

The structure of the paper is as follows: The conceptual toy model is introduced in the next section. In

"Forecast reliability," the concept of forecast reliability is discussed in the context of the conceptual model. A brief analysis of the (un)reliability of operational subseasonal and seasonal precipitation forecasts in the European Centre for Medium-Range Weather Forecasts (ECMWF) model is presented in the section "Reliability on seasonal and subseasonal time scales." Some discussion and conclusions are given in the final section.

THE TOY MODEL. Our climate can be considered a nonlinear dynamical system represented schematically by the equation

$$\frac{dX}{dt} = F(X),$$

where F is at least quadratic in the state variable X. Nonlinearity can manifest in two different ways. First, if we linearize these equations to describe how small perturbations δX evolve in time, the linearized equations have the form

$$\frac{d\delta X}{dt} = \frac{dF}{dX}\delta X. \quad (1)$$

Notice that if F is at least quadratic in X, then the Jacobian operator dF/dX will be at least linear in X. That is to say, in a nonlinear system, the growth of small perturbations (which characterize the predictability of the system) will vary with the underlying state X of the system. This dependence has been illustrated, for example, by Buizza and Palmer (1995) when studying the fastest-growing linear perturbations in the atmosphere. The existence of relatively stable regions of phase space, which characterize the phenomena of circulation regimes (Legras and Ghil 1985), is another manifestation of nonlinearity.

Figure 1 shows two configurations of an idealized system that incorporate these manifestations of nonlinearity (Palmer 1999). A ball is dropped onto the ridge separating two channels. If the ball is dropped slightly to the right of the ridge, the ball falls into the right-hand cup for the first configuration and into the left-hand cup for the more complex second configuration. The position of the ball can be considered as defining the variable X. Near the ridge, small perturbations to X will grow; this can be considered an unstable part of the system. By contrast, the cups can be considered very stable parts of the system; essentially, the ball will remain stationary in the cups until it is lifted out and dropped again into the funnel. The cups can be considered to represent atmospheric circulation regimes.

A fan is shown in Fig. 1, which, in analogy to the anthropogenic climate change situation, demonstrates the effects of a simple external forcing on the system. In both configurations, when the fan is off and the ball is repeatedly dropped into the funnel, the probability of the ball dropping in either of the cups is equal (i.e., 50%). However, when the fan is switched on, the probability that the ball will drop into the right-hand cup in the top configuration, or the left-hand cup in the bottom configuration, decreases.

The second configuration is clearly more complicated than the first. So let us suppose that the second configuration corresponds to REALITY and the first is a simplified MODEL of REALITY. Hence, compared with REALITY, the MODEL responds incorrectly to the applied forcing. For example, if the right-hand cup defines what we shall call an "anticyclonic blocking regime" and the left-hand cup defines a "zonal-flow regime" and the fan defines "anthropogenic forcing," then while the MODEL predicts an increase in the zonal regime with anthropogenic forcing, REALITY predicts an increase in the blocked regime.

Most current generation CMIP models predict that an increase in greenhouse gas forcing will lead to a decrease in the frequency of northern European blocking (Matsueda et al. 2009; Anstey et al. 2013). However, in the absence of a clear theory explaining this, most modelers would not feel especially confident in such a result, especially as the CMIP models do not simulate long-lived blocking anticyclones with any degree of realism (Masato et al 2013), and even models with higher atmospheric resolution still suffer large biases in Euro-Atlantic blocking (Schiemann et al. 2017). In this regard, the missing "twist" in the simplified MODEL of REALITY could be associated with an occasional phase error in the Rossby wave response to tropical forcing or, perhaps, an inadequate stratosphere.

The key question we wish to consider in this paper is the following: Given the incorrect response to an external forcing in our toy MODEL, compared to our toy REALITY, how could we determine a priori that the described idealized MODEL response to forcing was wholly incorrect before the fan was actually switched on? This clearly relates to the more practical question of how to know whether the CMIP models are correctly simulating the response to forcing at a time before the response to this forcing is known in reality.

Note that we cannot readily answer our question by studying the unforced MODEL climatological probability distribution. The idealized MODEL has regimes (i.e., cups) in exactly the same location as REALITY. Not only that, the unforced frequency of occurrence of each of the regimes is also exactly correct (i.e., 50%). That is to say, our idealized MODEL has none of the climatological errors in blocking and yet it responds

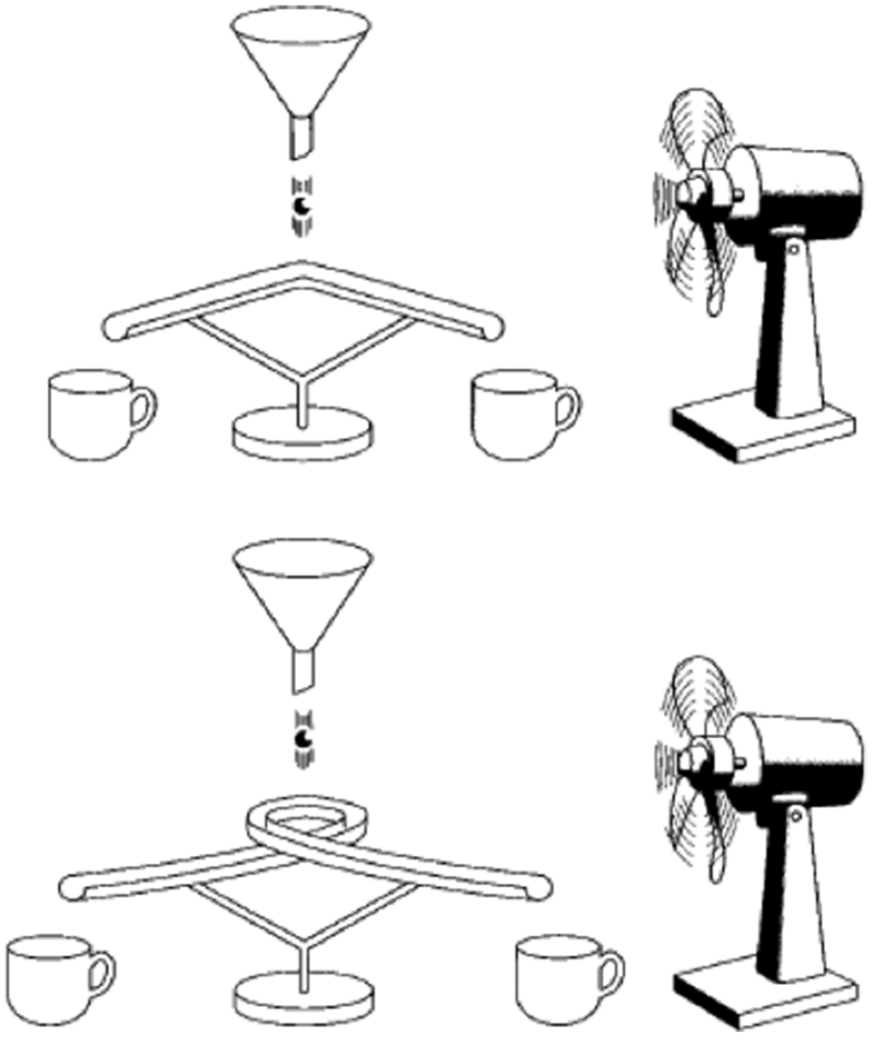

Fig. 1. Two configurations of a nonlinear toy system that respond in an opposite manner to external forcing, as represented by the fan. (bottom) The more complex system, denoted as REALITY. (top) The simpler system, considered a MODEL of REALITY.

incorrectly to forcing. Is there a way that we could know this before the fan was switched on?

FORECAST RELIABILITY. What specific information from a set of initial-value forecasts is relevant here? We argue that it is the probabilistic forecast reliability. Consider an (initial value) ensemble forecast system run over a large number of independent initial states. Focus on some binary event: rainfall exceeding the lower climatological tercile, for example. For a given grid point and forecast lead time, we can partition forecasts from our ensemble system into subsamples, or bins, where the probability of the event lay in the ranges, say, 0%–10%, 10%–20%, 20%–30%, …, and 90%–100%. For each forecast, the binary event either did occur or did not occur. Therefore, we can associate a frequency of occurrence for each of the subsamples of ensemble forecasts. For a reliable ensemble prediction system and a large-enough sample of events, the frequency of occurrence of events in the 0%–10% bin should lie between 0% and 10% and so on for all other bins. Put another way, on a 2D plot, where the frequency of occurrence defines distance along the *y* axis and the different forecast probability bins are plotted along the *x* axis, the data points should lie on or close to the diagonal. Examples of actual reliability diagrams from a state-of-the-art numerical weather prediction model are given below.

With this as background let us return to the idealized system shown in Fig. 1. The points A, B, C, D, E, and F in Fig. 2 correspond to sets of initial conditions for ensemble forecasts using observations of REALITY, assimilated (with some small random observation error) into the faulty MODEL during a period before the fan has been switched on. That is to say, we will study the probabilistic reliability of the unforced MODEL.

For initial conditions corresponding to A and B, the MODEL reliably predicts, respectively, a 100% or 0% probability for the left-hand regime. That is to say, in the subset of situations where the initial conditions belong to A, the forecast probability of occurrence of the left-hand regime is always 100%, and the frequency of occurrence of this regime is 100%. Similarly, in the subset of situations where the initial conditions belong to B, the forecast probability of occurrence of the left-hand regime is always 0%, and the frequency of occurrence of this regime is 0%.

For a set of initial conditions belonging to category C, that is, close to the unstable ridge, half of the time the ball falls into the left-hand regime, and half of the time the ball falls into the right-hand regime.

Hence, the frequency of occurrence of either regime is 50%. The forecast ensembles reliably predict that this is a situation with no predictability; here, in any ensemble the forecast probability of the left-hand regime is 50%. This illustrates an important point relevant to the Stott et al. (2016) comment in the introduction: reliability diagrams do not only test the ability of models to capture the predictable features of the event under consideration, they also test the ability of the model to predict reliably the situations where there is no predictable signal, that is, by producing an ensemble probability equal (within sampling uncertainty) to the climatological frequency of the event.

However, consider sets of initial conditions denoted by D in Fig. 2b. Some of the time the initial conditions will lie on the top channel; on other occasions, the initial conditions will lie on the bottom channel. Our imperfect MODEL, on the other hand, is unable to discriminate between these situations and in all circumstances the MODEL initial conditions (cf. Fig. 2a) lie on the left-hand channel (this is an example where initial-condition error is predominantly due to model error rather than observation error). As such, the MODEL always predicts a 100% probability of occurrence of the left-hand regime, when the observed frequency of occurrence of REALITY is actually 50%. Probabilistic forecasts from initial conditions D are therefore unreliable.

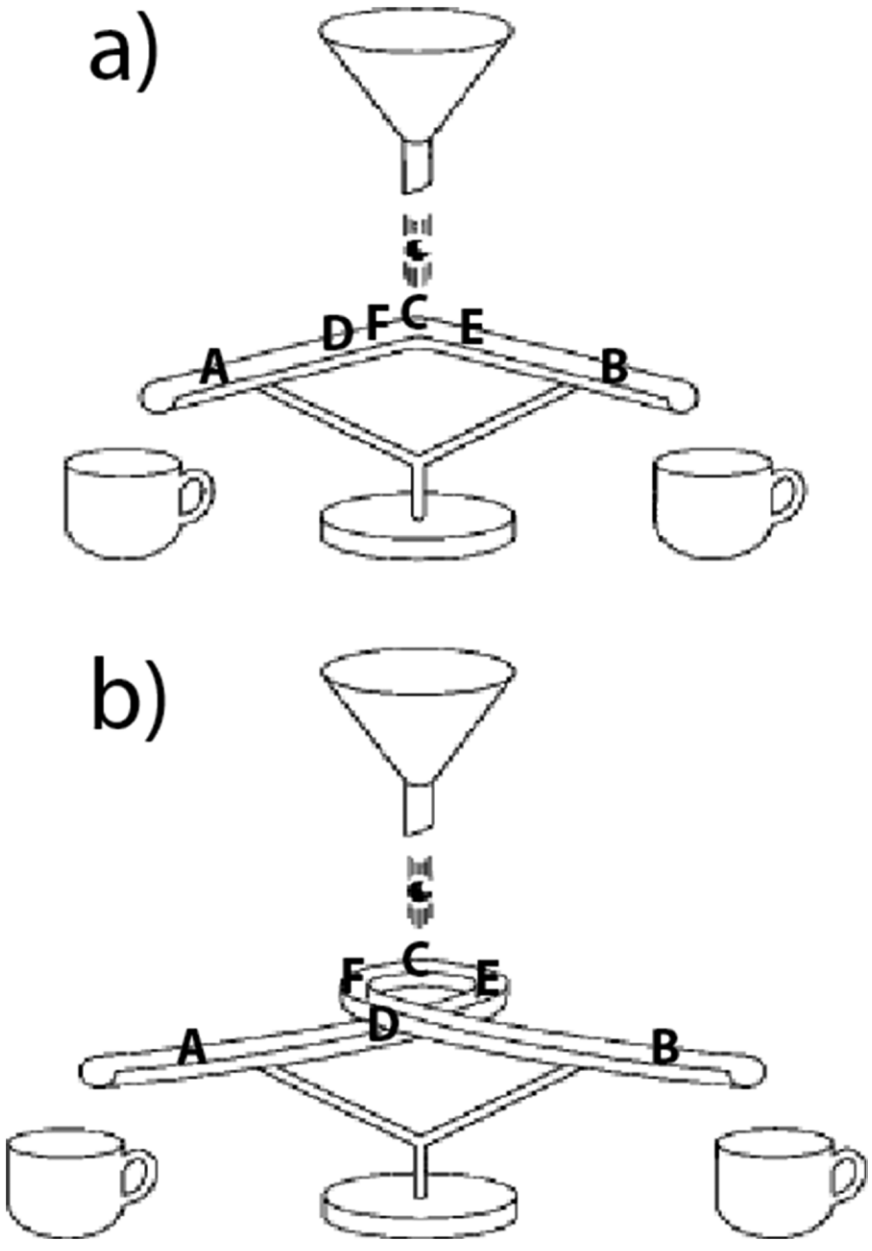

FIG. 2. Based on the toy systems shown in Fig. 1, we imagine a set of ensemble forecasts of the MODEL of REALITY from the different initial conditions A–F. We study the probabilistic reliability of these forecasts.

Worse still, for initial conditions E the MODEL ensembles predict 0% probability of the left-hand regime, when the observed frequency of occurrence in REALITY is 100%, and for initial conditions F the MODEL predicts a probability of 100% when the observed frequency in REALITY is 0%.

Binning the data according to prescribed probability bins and regressing over all the data points, the reliability curve for the MODEL would be flat, as illustrated in Fig. 3.

The key problem with the MODEL is that it does not have the twist in its channel, compared with REALITY. In principle, this error should show up in studies of the climatology of the model. However, the twist itself is not an energetically dominant aspect of the system.

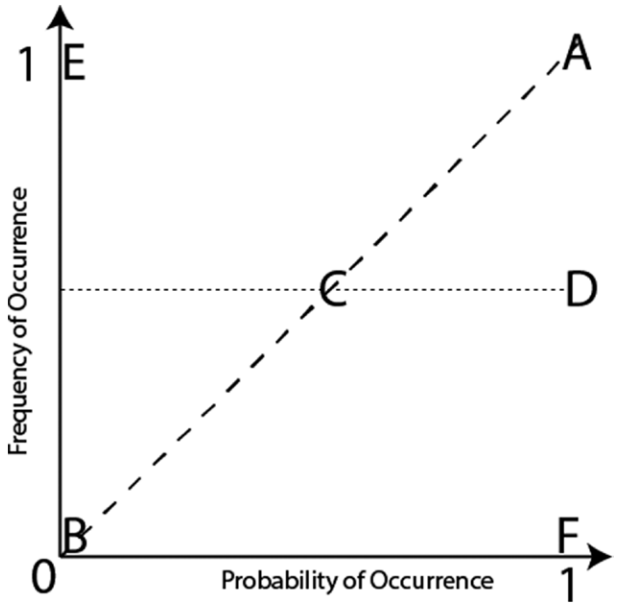

Fig. 3. The reliability diagram associated with the idealized model of reality for the six sets of initial conditions A–F. By binning over probability, and regressing over all data points, the overall reliability curve (dotted) is flat and quite different than the ideal diagonal (dashed).

That is to say, although the model error could in principle be revealed by some empirical orthogonal function (EOF) analysis, one may have to go to some very high-order EOF before the problem becomes apparent. This problem is revealed by focusing on forecast reliability. We therefore assert that forecast reliability should be considered a necessary (but certainly not sufficient) condition for having confidence that the model responds correctly to external forcing.

RELIABILITY ON SEASONAL AND SUBSEASONAL TIME SCALES. In the previous sections, the relevance of initial-value reliability for the forced climate response was demonstrated in a pedagogical nonlinear system. In this section, we discuss and present some reliability diagrams on seasonal and subseasonal time scales from the ECMWF operational system to assess the reliability of initialized ensemble forecasts performed with complex state-of-the-art coupled ocean–atmosphere models.

Seasonal predictions provide estimates of seasonal-mean statistics of weather, typically up to 3 months ahead of the season in question. As such, a seasonal forecast can provide information on how likely it is that the coming season will be wetter, drier, warmer, or colder than normal. Seasonal climate forecasts are increasingly being used across a range of application sectors, and reliable inputs are essential for any forecast-based decision-making. Weisheimer and Palmer (2014) characterized the reliability of regional temperature and precipitation forecasts from ECMWF's operational seasonal forecast system 4 in terms of usefulness and found a wide range of rankings, depending on region and variable. Most of the temperature forecasts over land were found to be at least marginally useful in terms of reliability. Overall, the reliability performance for precipitation was poorer than for temperature with more regions classified with lower reliability scores. For example, over northern Europe, the reliability of precipitation forecasts for winters [December–February (DJF)] is classified as not useful for dry events and marginally useful for wet events. The reliability for dry summers over Europe is notably poor with southern Europe classified in reliability categories as not useful and northern Europe as "dangerously useless."

The most frequent reliability category for precipitation was the marginally useful category. Such forecasts are not very reliable but might be marginally useful for some applications; see the calibration technique discussed in the next section. The category with the second-highest number of regions is the one of perfect reliability, which is an optimistic result for the usefulness of seasonal forecasts of precipitation. However, there are substantially more cases of areas that have a poorer precipitation forecast reliability than there are for temperature. It is exactly those areas that our analysis above suggests that the response to forcing is untrustworthy.

Weisheimer et al. (2017) have proposed the use of reliability information from a 110-yr-long dataset of seasonal retrospective forecasts of extreme events on the seasonal time scale to contrast the reliability of forecast probabilities at the beginning of the twentieth century, indicative of a "pre–climate change" period, with the reliability of forecast probabilities for more recent decades, indicative of a "post–climate change" period. Consistent with the discussion here, they conclude that such information can be important for increasing the confidence in attribution statements of extreme weather and climate events.

The problem of seasonal forecasts becoming seriously unreliable for certain events, regions, and seasons, as highlighted above, starts to manifest itself already much earlier in the forecast range. Here, we analyze data from a subseasonal retrospective forecast experiment using the ECMWF monthly forecasting system of model cycle 41R1. The 32-day-long hindcasts were run for 80 start dates over the period 1989–2008. We estimate forecast reliability for four weekly periods from day 4 to 32 based on weekly precipitation anomalies over Europe. The results are

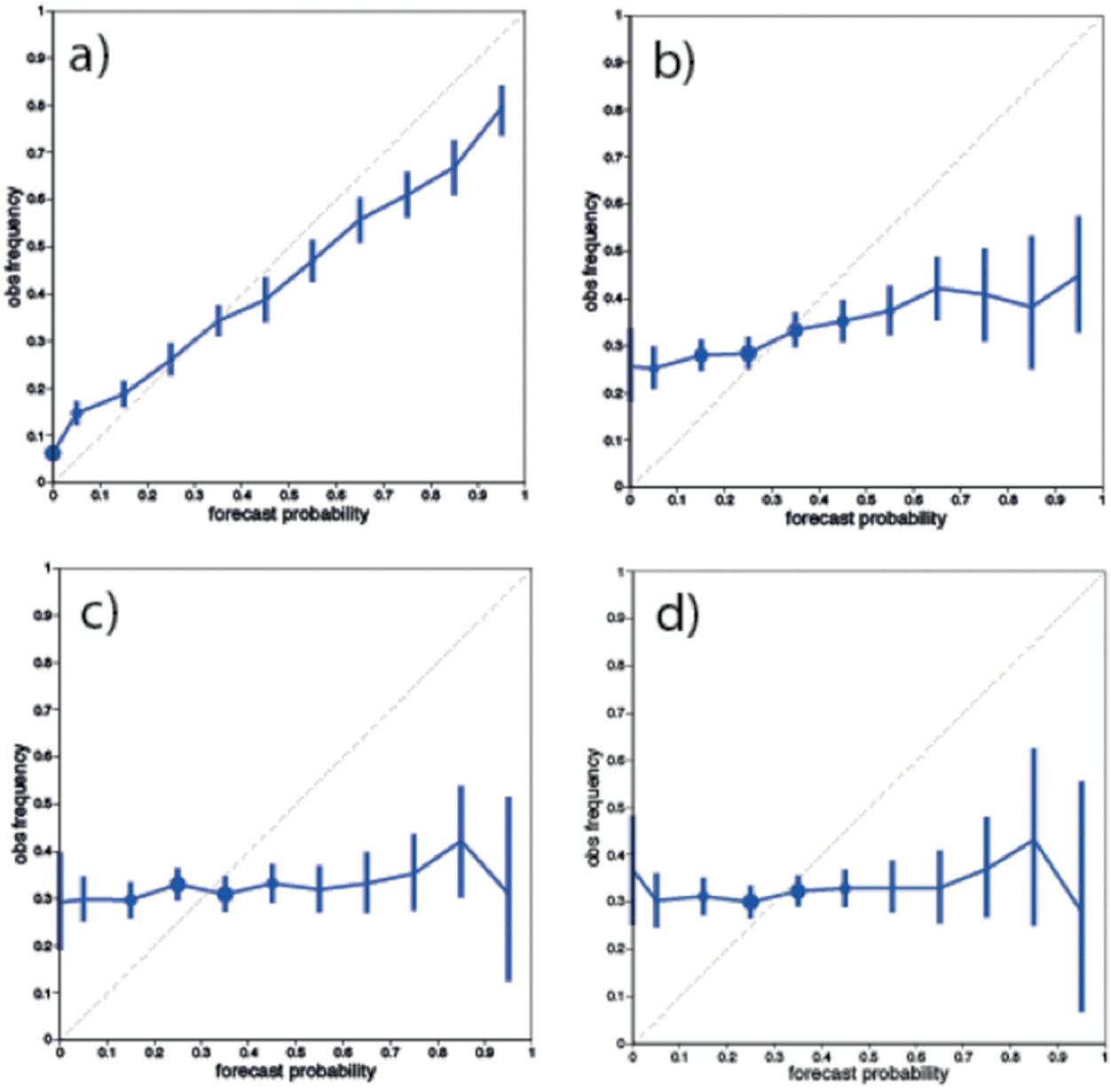

Fig. 4. Reliability of lower-tercile precipitation from ECMWF monthly forecasts for all European grid points: (a) days 4–11, (b) days 12–18, (c) days 19–25, and (d) days 26–32, from start dates of 1 Feb, 1 May, 1 Aug, and 1 Nov over the period 1989–2008.

presented in Fig. 4 and show that while forecast reliability is good for days 4–11, it quickly drops to a flat line in the reliability diagram for days 12–18 onward, indicating no useful forecast reliability on these time scales. The decrease of reliability only after the end of the conventional medium range of weather forecasts points to the likelihood that such unreliability is due to the relatively slowly growing model error rather than the more rapidly growing initial-condition uncertainty.

Although multimodel ensembles typically have greater initial-value reliability than single-model ensembles, reliability is still far from perfect. For example, the component models of a contemporary multimodel ensemble all suffer from the same systematic deficiencies in regime statistics (such as undersimulation of long-lived blocks). Hence, the results here are also relevant to multimodel ensembles.

DISCUSSION AND CONCLUSIONS. In this paper, we have used a simple nonlinear toy model of the atmospheric circulation to clarify the notion that the unreliability of initial-value ensemble predictions can directly reveal the untrustworthiness of

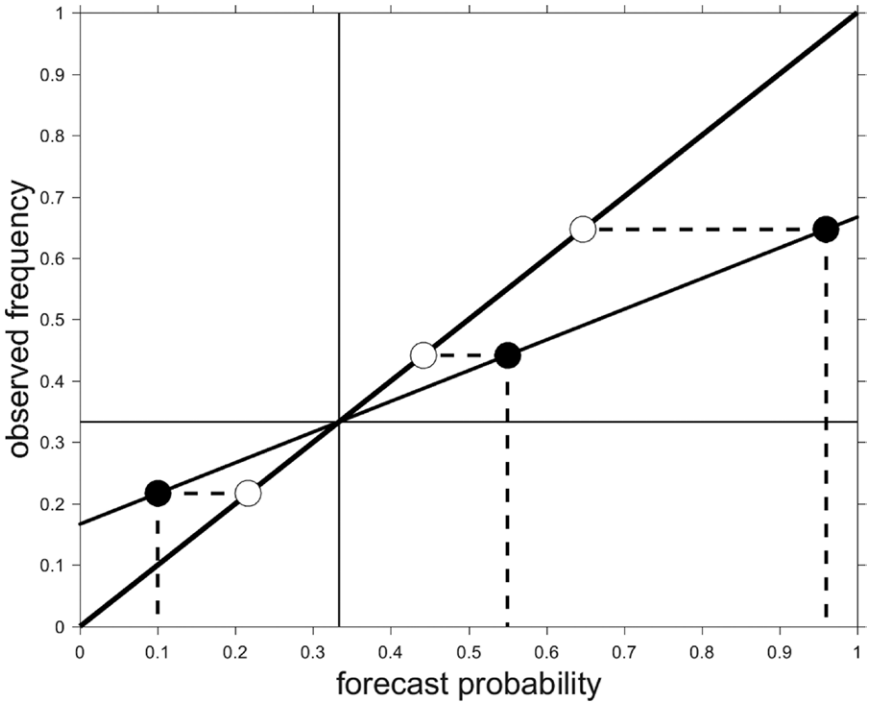

Fig. 5. An idealized situation showing how a posteriori calibration can make unreliable probability forecasts (solid circles) reliable (open circles).

the model's forced response. The system consists of a REALITY with two circulation regimes and a simpler MODEL of REALITY. While, by construction, the MODEL simulates these two regimes with the correct climatological structure and frequency, the MODEL is not able to correctly simulate the response to an external forcing. We have shown that it is possible to know a priori (i.e., without performing the MODEL simulation under forcing) that the MODEL response to forcing is incorrect. The key that allows such knowledge is data from initial-value (unforced) predictions of REALITY using the simplified MODEL. Specifically, the MODEL's unreliability indicates that its response to an external forcing is also not trustworthy.

With this in mind, it is worth returning to the Stott et al. (2016) quote in the introduction. As discussed in the previous section "Forecast reliability," an initial-value reliability diagram does not only indicate whether the model is able to capture the predictable features of the event under consideration, it also indicates whether the model can predict climatological probabilities in situations where the event is unpredictable. That is to say, initial-value reliability is still an important diagnostic to study even in situations where there is no inherent real-world predictability. We therefore do not agree with the conclusion of Stott et al. (2016, p. 32) that "although the use of reliability is well established for forecasting, its meaning for attribution is less clear given that reliable attribution is still possible when there is no inherent real-world predictability."

How can we use initial-condition reliability for improving quantitatively the trustworthiness of a climate model's response to forcing? The reliability diagram provides a simple framework for calibrating probabilities to make them (in sample) reliable. Suppose the raw forecast probabilities are not reliable because the ensemble system is overconfident and creates, on average, insufficient spread to account for the forecast errors (i.e., the ensemble is underdispersive). In Fig. 5 such a case is schematically illustrated by the solid circles in the reliability diagram. Here, the diagram shows an example for overconfident predictions of a binary tercile event. Where forecast probabilities are larger than 1/3, the events occur less frequently than predicted by the ensemble. Where forecast probabilities are smaller than 1/3, the events occur more frequently. By moving the uncalibrated forecast probabilities (solid circles) horizontally toward the perfect reliability diagonal (open circles), the forecast probabilities can be corrected to be more reliable, keeping the observed frequency of occurrence unchanged. This is illustrated in Fig. 5 by shifting the solid circles to the open circles.

Here, we propose that such forecast calibration could, for example, be used to downweight probabilistic attribution statements. Exactly how this should be done is a matter for further research, but a plausible methodology would be to downweight the probabilities based on the reliability of initial-value forecasts with shorter lead times that can be verified. The lead time of the forecasts could be linked to the time scale of the attributed event of interest. For example, probabilistic attribution statements for seasonal-mean extreme events could be downweighted on the basis of seasonal forecast reliability for percentile categories in which the event occurs. Similarly, statements about extremes on shorter time scales could be calibrated with forecast reliability from extended-range or subseasonal predictions.

How could one test such a proposal? Here, we note the study of Matsueda et al. (2016) mentioned above. Here, both high- and low-resolution versions of the model were run first with prescribed late-twentieth-century SSTs and second in climate change time-slice mode for the late twenty-first century with prescribed SSTs based on CMIP climate change integrations. A series of low-resolution seasonal ensemble hindcasts were also made, again with prescribed SSTs from the late twentieth century. These were verified against the high-resolution twentieth-century integration, that is, treating the high-resolution model output as a "surrogate of truth." Probabilities from the low-resolution time-slice integration were then downweighted using the calibration statistics from the seasonal reliability diagrams and found to be closer to those from the "surrogate truth," high-resolution time-slice integration than were the uncalibrated probabilities from the low-resolution time-slice run.

Ultimately, a key way to improve the reliability of initial-value forecasts is to reduce model error. A recent study by Bellprat and Doblas-Reyes (2016), based on a simple statistical climate model, concluded that increased model error can lead to reduced reliability that results in a systematic overestimation of the fraction of attributable risk (FAR), one of the most common measures in climate event attribution. Along similar lines, Lott and Stott (2016) analyzed noninitialized climate GCM simulations of CMIP5 and, using a perfect model approach, were able to relate reduced reliability of the climate model simulations to increased estimates of the error in FAR. Reducing model error means solving the relevant laws of physics more accurately. Schiemann et al. (2017) have noted an improvement in the ability to simulate blocking behavior with increasing atmospheric resolution but find that even with 25-km horizontal resolution, blocking frequencies are still substantially undersimulated. It is plausible that further increases in resolution, coupled with more sophisticated stochastic parameterization schemes, may be necessary to reduce initial-value unreliability to small values.

The toy model discussed in this paper provides support for the notion of seamless prediction: the attempt to unify the tools needed to predict weather and climate. The essential philosophy of seamless prediction is that through such unification, scientific insights gained by running the model in the initial-value mode can be transferred to understanding the forced climate problem and vice versa (Palmer and Webster 1999).

ACKNOWLEDGMENTS. T. N. Palmer was supported by the European Research Council Grant 291406: "Towards the Prototype Probabilistic Earth-System Model" (PESM). A. Weisheimer was supported by the project EUCLEIA (Grant Agreement 607085) funded by the European Commission Seventh Framework Research Programme. We thank Dr. David MacLeod, the reviewers, and the editor for their helpful comments on an earlier version of this paper.

REFERENCES

Anstey, J. A., and Coauthors, 2013: Multi-model analysis of Northern Hemisphere winter blocking: Model biases and the role of resolution. *J. Geophys. Res. Atmos.*, **118**, 3956–3971, https://doi.org/10.1002/jgrd.50231.

Barnes, E. A., and J. Screen, 2015: The impact of Arctic warming on the midlatitude jetstream: Can it? Has it? Will it? *Wiley Interdiscip. Rev.: Climate Change*, **6**, 277–286, https://doi.org/10.1002/wcc.337.

Bellprat, O., and F. Doblas-Reyes, 2016: Attribution of extreme weather and climate events overestimated by unreliable climate simulations. *Geophys. Res. Lett.*, **43**, 2158–2164, https://doi.org/10.1002/2015GL067189.

Buizza, R., and T. N. Palmer, 1995: The singular vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2.

Francis, J. A., and N. Skific, 2015: Evidence linking rapid Arctic warming to mid-latitude weather patterns. *Philos. Trans. Roy. Soc.*, **A373**, 20140170, https://doi.org/10.1098/rsta.2014.0170.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp., https://doi.org/10.1017/CBO9781107415324.

Legras, B., and M. Ghil, 1985: Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.*, **42**, 433–466, https://doi.org/10.1175/1520-0469(1985)042<0433:PABAVI>2.0.CO;2.

Lott, F. C., and P. A. Stott, 2016: Evaluating simulated fraction of attributable risk using climate observations. *J. Climate*, **29**, 4565–4575, https://doi.org/10.1175/JCLI-D-15-0566.1.

Masato, G., B. J. Hoskins, and T. Woollings, 2013: Winter and summer Northern Hemisphere blocking in CMIP5 models. *J. Climate*, **26**, 7044–7059, https://doi.org/10.1175/JCLI-D-12-00466.1.

Matsueda, M., R. Mizuta, and S. Kusunoki, 2009: Future changes in wintertime atmospheric blocking simulated using a 20-km-mesh atmospheric global circulation model. *J. Geophys. Res.*, **114**, D12114, https://doi.org/10.1029/2009JD011919.

——, A. Weisheimer, and T. N. Palmer, 2016: Calibrating climate change predictions with estimates of seasonal forecast reliability. *J. Climate*, **29**, 3831–3840, https://doi.org/10.1175/JCLI-D-15-0087.1.

Palmer, T. N., 1999: A nonlinear dynamical perspective on climate prediction. *J. Climate*, **12**, 575–591, https://doi.org/10.1175/1520-0442(1999)012<0575:ANDPOC>2.0.CO;2.

——, and P. J. Webster, 1995: Towards a unified approach to climate and weather prediction. *Global Change: Proceedings of the First Demetra Meeting on Climate Change*, A. Speranza, S. Tibaldi, and R. Fantechi, Eds., European Community Press, 265–280..

——, F. Doblas-Reyes, A. Weisheimer, and M. Rodwell, 2008: Reliability of climate change projections of precipitation: Towards "seamless" climate predictions. *Bull. Amer. Meteor. Soc.*, **89**, 459–470, https://doi.org/10.1175/BAMS-89-4-459.

Ramesh, K. V., and P. Goswami, 2014: Assessing reliability of regional climate projections: The case of Indian monsoon. *Sci. Rep.*, **4**, 4071, https://doi.org/10.1038/srep04071.

Scaife, A., C. Buontempo, M. Ringer, M. Sanderson, C. Gordon, and J. Mitchell, 2009: Comment on "Toward seamless prediction: Calibration of climate change projections using seasonal forecasts." *Bull. Amer. Meteor. Soc.*, **90**, 1549–1551, https://doi.org/10.1175/2009BAMS2753.1.

Schiemann, R., and Coauthors, 2017: The resolution sensitivity of Northern Hemisphere blocking in four 25-km atmospheric global circulation models. *J. Climate*, **30**, 337–358, https://doi.org/10.1175/JCLI-D-16-0100.1.

Stott, P. A., and Coauthors, 2016: Attribution of extreme weather and climate-related events. *Wiley Interdiscip. Rev. Climate Change*, **7**, 23–41, https://doi.org/10.1002/wcc.380.

Turner, A. G., and H. Annamalai, 2012: Climate change and the South Asian summer monsoon. *Nat. Climate Change*, **2**, 587–595, https://doi.org/10.1038/nclimate1495.

Webster, P. J., V. O. Magaña, T. N. Palmer, J. Shukla, R. A. Tomas, M. Yanai, and T. Yasunari, 1998: Monsoons: Processes, predictability, and the prospects for prediction. *J. Geophys. Res.*, **103**, 14 451–14 510, https://doi.org/10.1029/97JC02719.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

——, N. Schaller, C. O'Reilly, D. MacLeod, and T. Palmer, 2017: Atmospheric seasonal forecasts of the twentieth century: Multi-decadal variability in predictive skill of the North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quart. J. Roy. Meteor. Soc.*, **143**, 917–926, https://doi.org/10.1002/qj.2976.

Wettstein, J. J., and C. Deser, 2014: Internal variability in projections of twenty-first century Arctic sea ice loss: Role of the large-scale atmospheric circulation. *J. Climate*, **27**, 527–550, https://doi.org/10.1175/JCLI-D-12-00839.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.