



## RESEARCH ARTICLE

10.1029/2018MS001597

## Key Points:

- Machine learning provides a method of deducing better Earth system models from data
- It is simpler to correct errors in physically-derived models than to replace them and learn from scratch
- Robust improvements in forecast and climate diagnostics are shown for the chaotic Lorenz '96 system

## Correspondence to:

P. A. G. Watson,  
peter.watson@physics.ox.ac.uk

## Citation:

Watson, P. A. G. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, 11. <https://doi.org/10.1029/2018MS001597>

Received 18 DEC 2018

Accepted 20 APR 2019

Accepted article online 25 APR 2019

# Applying Machine Learning to Improve Simulations of a Chaotic Dynamical System Using Empirical Error Correction

Peter A. G. Watson<sup>1</sup>
<sup>1</sup>Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK

**Abstract** Dynamical weather and climate prediction models underpin many studies of the Earth system and hold the promise of being able to make robust projections of future climate change based on physical laws. However, simulations from these models still show many differences compared with observations. Machine learning has been applied to solve certain prediction problems with great success, and recently, it has been proposed that this could replace the role of physically-derived dynamical weather and climate models to give better quality simulations. Here, instead, a framework using machine learning together with physically-derived models is tested, in which it is learnt how to correct the errors of the latter from time step to time step. This maintains the physical understanding built into the models, while allowing performance improvements, and also requires much simpler algorithms and less training data. This is tested in the context of simulating the chaotic Lorenz '96 system, and it is shown that the approach yields models that are stable and that give both improved skill in initialized predictions and better long-term climate statistics. Improvements in long-term statistics are smaller than for single time step tendencies, however, indicating that it would be valuable to develop methods that target improvements on longer time scales. Future strategies for the development of this approach and possible applications to making progress on important scientific problems are discussed.

## 1. Introduction

Numerical weather prediction and climate models attempt to predict and simulate components of the Earth system, including the atmosphere and perhaps also the oceans, land surface, and biosphere. While the fundamental physical equations governing the system are known, they cannot be solved accurately with available computational resources. Instead, approximations are made in the models' equations, and this gives rise to errors in their output. Methods to reduce these errors are highly valuable for giving better warning of major meteorological and climatic events.

Recently, great advances in machine learning have taken place, for example, in the domains of image recognition and game playing (e.g., He et al., 2016; Silver et al., 2017). The algorithms developed have been found to excel at certain problems that involve predicting an unknown value given values of predictor variables (e.g., predicting what objects a photograph contains given its pixel values)—this is similar to the problem of predicting future behavior of the Earth system given knowledge of its past and present state, and so there has been high interest in applying machine learning to improve such predictions. This has included predicting future weather events directly from observations and postprocessing dynamical models' output (e.g., Krasnopolsky & Lin, 2012; McGovern et al., 2017; Rasp & Lerch, 2018; Scher & Messori, 2018).

Another emerging application is applying machine learning to improve components of the dynamical Earth system models themselves, particularly the parameterizations of unresolved small-scale processes such as radiative interactions and cloud processes. This could allow larger improvements in prediction skill than is achieved by postprocessing models' output, by better representing the physical interactions between variables.

Previous work has primarily used artificial neural networks (ANNs), which are functions constructed from “neurons.” Neurons are simply functions that linearly combine their inputs and then apply a given (generally nonlinear) transformation to produce the output value. ANNs pass input data into neurons, whose output may then be used as inputs to more neurons, and so on until a final output value (or vector of values) is produced. ANNs can relatively efficiently encode complex functional relationships. Indeed, an ANN

©2019. The Authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

with neurons arranged in layers, with the outputs of neurons in one layer being inputs to neurons in the next layer, can represent any real continuous function to an arbitrarily small level of accuracy, given a sufficient number of neurons (Cybenko, 1989). Nielsen (2015) and Goodfellow et al. (2016) provide general introductions to the theory and applications of ANNs.

One promising approach has been to use algorithms such as ANNs to reproduce the behavior of atmospheric parameterization schemes at a reduced computational cost. For example, Chevallier et al. (1998, 2000) found that ANNs could cheaply reproduce the behavior of the radiative transfer scheme in the European Centre for Medium-Range Weather Forecasts model, although Morcrette et al. (2008) note that this approach did not work sufficiently well after the model's vertical resolution was increased. Krasnopolsky et al. (2005, 2010) also found that an ANN could be used to cheaply reproduce the output of the radiation scheme in the National Center for Atmospheric Research Community Atmosphere Model. More recent work has focused on replacing atmospheric models' convection schemes with ANNs, in order not just to reduce the cost of presently used schemes but to allow more expensive, higher-quality schemes to be used, such as superparameterization (Gentine et al., 2018; Rasp et al., 2018) or emulations of convection-resolving models (Brenowitz & Bretherton, 2018). O'Gorman and Dwyer (2018) also showed that a model's convection scheme could be replaced by a random forest algorithm, and the model could run stably and reasonably reproduce precipitation extremes.

Machine learning also holds promise of being able to reduce errors in dynamical models' predictions. Schneider et al. (2017) describe how better values of parameters of models could be learnt by algorithms being fed data from observations and high-resolution models. Dueben and Bauer (2018) examine whether ANNs could be trained to simulate atmospheric dynamics and provide prediction skill exceeding that of existing models, using a time stepping scheme where ANNs predict the tendency of the system in a similar approach to that used in existing dynamical models, and they conclude that it is possible. However, their simulations became unstable after about 2 weeks. Additionally, Bolton and Zanna (2019) showed that a convolutional ANN could skillfully predict small-scale momentum forcing in a quasi-geostrophic ocean model given spatially smoothed output from a simulation, although this was not implemented into a freely evolving model. Significant advances have also been made in learning symbolic equations and invariant quantities from data (Rudy et al., 2017; Schmidt & Lipson, 2009; Wu & Tegmark, 2018; Zhang & Lin, 2018), which could be applied in Earth system modeling (Gaitan et al., 2016).

The above-mentioned property of ANNs that they can represent any continuous function means that, in principle, given sufficient data of high enough quality to learn from and adequate computational resources for training, an ANN's representation of the equations of motion of the Earth system could reach the maximum skill possible for given inputs. So it seems that ANNs could potentially learn to reduce systematic model errors without the need for human ingenuity, greatly speeding-up model development and helping us to address challenges like predicting extreme weather and the impacts of climate change.

However, the use of ANNs as discussed in Schneider et al. (2017) and Dueben and Bauer (2018) has been presented as being in competition with improving the conventional physically derived aspects of Earth system models. Schneider et al. (2017) argue that improving physically derived parameterization schemes is preferable to using ANNs because they will obey conservation laws and symmetries. Dueben and Bauer (2018) ask whether models based entirely on ANNs can compete with physically derived models.

One purpose of the work presented here is to explore whether it is actually possible to use such algorithms to complement physically-derived model components, thereby preserving the benefits of using the latter, such as having better physical interpretability of the model behavior and better trust that the model will perform reasonably well in an unseen physical situation. Karpatne et al. (2017a) and Reichstein et al. (2019) provide overviews of the ways in which statistical algorithms can be combined with physical modeling and the associated challenges. The specific proposal tested here is to use algorithms to perform empirical error correction in dynamical models. Rather than predict the whole tendency of a system, as in the models considered by Dueben and Bauer (2018), the algorithms would predict the difference between the measured and the observed tendencies. Then the total tendency would be  $\mathcal{M}(x) + \epsilon(x)$ , where  $x$  is the system state at, and potentially before, the start of the time step,  $\mathcal{M}(x)$  is the tendency predicted by the physically-derived model, and  $\epsilon(x)$  is the correction output by the algorithm. If  $\mathcal{M}(x)$  is close to the optimum tendency,  $\epsilon(x)$  should be small, and so concerns about  $\epsilon(x)$  not obeying conservation laws and symmetries are consequently less important than in the case where whole model components are replaced by algorithms (note, though,

that it may also be possible to constrain  $\epsilon(x)$  to obey these physical principles more strictly; e.g., Jia et al., 2018).  $\epsilon(x)$  should only have a large effect on the simulations when the prediction by the physically-derived model is poor, in which case the value of improving the total simulated tendency is larger compared to concerns about whether physical principles are strictly abided by. The physically-derived model maintains a key role, and it is desirable to continue improving it to strengthen the link between the simulation results and our physical understanding. This study focuses on the use of ANNs as model error correctors, but other algorithms could also be applied in a similar way.

A further advantage of using algorithms to correct models' errors rather than replace physically-derived models entirely is that it greatly simplifies the process of incorporating ANNs into dynamical models. Dueben and Bauer (2018) detail the numerous challenges in replacing physically derived models with ANNs (or other algorithms), such as obtaining the required data for training a full-complexity model and learning to use algorithms with the required complexity. A lot of development effort would be required before a model with better performance than current models would be produced. By contrast, development of error-correcting algorithms can start just by improving a small number of outputs as much as possible given a small number of inputs, which is achievable with a smaller research program, and progress can build from there. A disadvantage of this approach is that the computational cost of the models cannot easily be reduced this way if the resolution and parameterization schemes are kept the same—the focus is on improving the simulation quality. However, it may turn out to be more cost effective than using more expensive parameterization schemes or increasing the model resolution, and it could reduce costs if it allows the same or greater skill to be obtained using cheaper physically-derived parameterizations. It would also be very informative about the problems that would need to be overcome to get dynamical models based entirely on algorithms like ANNs to perform well.

An empirical error-correcting approach using an ANN was applied by Forssell and Lindsog (1997) to predict the water level in a tank. Previous environmental applications include the prediction of groundwater flow by Xu and Valocchi (2015) and the prediction of lake temperatures by Karpatne et al. (2017b) and Jia et al. (2018). These systems exhibit variability that is strongly influenced by external drivers, while Earth's atmosphere and oceans have a large component of unforced variability due to chaotic dynamics. A demonstration that using an algorithm to correct model errors could improve short-range forecasts of simple chaotic systems was given by Pathak et al. (2018), who used reservoir computing, but it is unclear if this would also produce stable simulations with improved long-term statistics. Cooper and Zanna (2015) applied this approach to improve a chaotic shallow water model but did so using a linear system of equations, which seems unlikely to be able to represent errors in complex dynamical systems as well as a more flexible algorithm such as an ANN in general—for example, atmospheric convection is a highly nonlinear process (Arakawa, 2004), and errors in its representation seem unlikely to be captured well by linear equations. Their work also focused on improving long-term statistics of the simulation, and it is not clear whether there was a simultaneous improvement in short-range forecast skill that would indicate that the dynamics were being better represented.

It is a key problem in Earth system prediction to improve our models in such a way that they are stable and give both improved skill in initialized predictions and better long-term climate statistics. In the remainder of this paper, it is tested whether this can be achieved by using an error-correcting ANN to better simulate a chaotic system, namely, the Lorenz '96 dynamical system (Lorenz, 1996; sometimes also referred to as the Lorenz '95 system; e.g., by Dueben & Bauer, 2018). This system, or variants of it, has been used in many previous studies to test concepts for how to improve dynamical Earth system models (e.g., Arnold et al., 2013; Dueben & Bauer, 2018; Schneider et al., 2017; Wilks, 2005). The results are informative about the potential for machine learning approaches to improve skill at simulating dynamical systems such as the Earth's climate, albeit in a much simpler setting.

## 2. Experiments With the Lorenz '96 System

### 2.1. The Lorenz '96 Equations and Coarse-Resolution Models

The Lorenz '96 dynamical equations describe the evolution of variables arranged in a ring, intended to be analogous to a latitude circle. The variables are divided into two types: slowly varying  $X_k$  and quickly varying  $Y_{j,k}$ , defined for  $k = 1, \dots, K$ , and  $j = 0, \dots, J + 2$ . Lorenz (1996) suggested that the  $Y_{j,k}$  be considered analogous to a convective-scale quantity in the real atmosphere and  $X_k$  analogous to an environmental vari-

able that favors convective activity. Here one of the systems used by Arnold et al. (2013) is simulated as the “Truth” system—there is no particular strong reason to choose this system over other variants, but its prior use in the parameterization development work of Arnold et al. (2013) makes it seem like a good choice for exploring how the design of models can be further advanced. This has  $K = 8$  and  $J = 32$ , in which

$$\begin{aligned}\frac{dX_k}{dt} &= -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - (hc/b) \sum_{j=1}^J Y_{j,k}, \\ \frac{dY_{j,k}}{dt} &= -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + (hc/b)X_k,\end{aligned}\quad (1)$$

with cyclic boundary conditions  $X_k = X_{k+K}$  and  $Y_{j,k} = Y_{j,k+K}$  and parameter values  $h = 1$ ,  $F = 20$ ,  $b = 10$ , and  $c = 4$ . The  $Y$  variables are connected in a ring, such that  $Y_{0,k} = Y_{J,k-1}$ ,  $Y_{J+1,k} = Y_{1,k+1}$ , and  $Y_{J+2,k} = Y_{2,k+1}$ , so there are  $J$  unique  $Y_{j,k}$  variables associated with each  $X_k$  variable. The time units are arbitrary and denoted as model time units (MTUs). These equations were integrated in time with a time step of 0.001 MTU using a fourth-order Runge-Kutta time stepping scheme.

A “training” simulation of this system of length 3,000 MTUs (not including 10 MTUs discarded as “spin up” at the beginning) was produced to provide a sample of “true” statistics to use in constructing coarse-resolution models below. The simulation was then extended for 10 MTUs so that memory of the training data set was effectively lost, and then a further 3,000 MTUs of data were generated to use as a validation data set, which is used only to evaluate and not to develop the coarse-resolution models below. (Note that evaluation against a separate “test” data set is not done because the aim is not to select a single best-performing ANN structure and estimate the model’s true skill. This means that an ANN that is selected on the basis of having an especially good performance on the validation data set would not be expected to perform as well relative to other ANNs on separate test data, since sampling variability would be expected to have contributed to its diagnosed skill in the validation data. It is shown below that performance improvements on the validation data are obtained for a wide range of ANN structures, and so the conclusion that ANNs can improve prediction skill for this system is robust to sampling variability.)

### 2.1.1. Coarse-Resolution Model

Suppose that a much computationally cheaper model of equation (1) is desired for making short-term forecasts and simulating the long-run statistics of this system. Following Wilks (2005) and Arnold et al. (2013), inspection of equation (1) suggests that it may be reasonable to forego simulating the  $Y$  variables explicitly and parameterize their effect on the  $X$  variables with a function  $U(X)$ , analogous to how the effect of unresolvable physical processes on resolved scales is parameterized in Earth system models. This yields the coarse-resolution system

$$\frac{dX_k^*}{dt} = -X_{k-1}^*(X_{k-2}^* - X_{k+1}^*) - X_k^* + F - U(X_k^*) \quad (2)$$

with  $X_k = X_{k+K}$ . The time step is also increased to 0.005 MTU, so that the system has a coarsened time resolution as well, analogous to how Earth system models cannot resolve real Earth processes that happen on very fast time scales.

The function  $U(X_k^*)$  is derived using the same method as Arnold et al. (2013), using essentially a coarse-graining approach. It is defined as a cubic function,

$$U(X) = \sum_{n=0}^3 a_n X^n,$$

and its parameters are chosen using tendencies of the  $X$  variables over intervals of length 0.005 MTU, derived from the run of the truth system sampled every 0.005 MTU. Its parameters were fit to minimize the root-mean-square error (RMSE) of predictions of these tendencies made using equation (2), taking values  $a_0 = -0.207$ ,  $a_1 = 0.577$ ,  $a_2 = -0.00553$ , and  $a_3 = -0.000220$ . This is a statistical procedure, but note that here  $U(X)$  is not considered part of the machine learning algorithm used to correct the model errors. Following Wilks (2005) and Arnold et al. (2013),  $U(X)$  is thought of as being analogous to parameterizations of unresolved processes in an Earth system model—note that development of physically-derived Earth system model parameterizations can also involve fitting parameters to data (e.g., Hourdin et al., 2016). Including  $U(X)$  in equation (2) helps to test whether complex algorithms such as ANNs are able to give skill improve-

ments after much of the possible progress has been made with physical reasoning and simpler statistical methods, using the deterministic model of Arnold et al. (2013) as a benchmark. It would also be possible to do a similar study with  $U(X) = 0$ , which would probably increase the potential improvement made by using error-correcting algorithms as they learnt to represent the improvements made by including  $U(X)$ .

Hereafter, the model given by equation (2) is referred to as “No-ANN.”

### 2.1.2. Models With ANNs

To produce coarse-resolution models with error-correcting ANNs, ANNs with a multilayer perceptron architecture (Goodfellow et al., 2016; Nielsen, 2015) were trained to predict the difference between the true system tendency and that predicted by the coarse-resolution model for one  $X$  variable at a time:

$$\epsilon_k = \frac{dX_k}{dt} - \frac{dX_k^*}{dt}.$$

The inputs to the ANNs are  $X$  variables up to two points away from the location where the prediction is being made, so that five  $X$  values in total are used as input. This is one more than used by the No-ANN model and takes advantage of the ability of ANNs to use inputs that are difficult to know how to include by physical reasoning—this may be helpful in Earth system modeling to account for subgrid phenomena that propagate between grid boxes but are difficult to include in parameterization schemes, such as horizontally propagating gravity waves (Alexander et al., 2010). Results for ANNs using the same inputs as the No-ANN model are discussed at the end of section 2.2 to quantify the impact made by using  $X_{k+2}$  as an additional input—it does not qualitatively affect the findings. Not all of the  $X$  variables were used as input in order that the ANNs are only using information from nearby grid points. This is desirable in Earth system models so they can be run much more quickly in parallel computing environments (Dueben & Bauer, 2018). Also, since on Earth phenomena at one location could not be meaningfully influenced by phenomena on the other side of the world within one model time step, this structure constrains the ANNs to be more faithful to the true equations, so they are more likely to work well in novel situations. The results presented here are not expected to depend qualitatively on the number of  $X$  variables used, and they were found to not be sensitive to using  $X$  variables up to three points away instead.

The  $X$  variables are transformed by subtracting their mean and dividing by their standard deviation before being used as ANN inputs, since this tends to speed up training of ANNs (LeCun et al., 2012). The values of the mean and standard deviation are derived during training and are not changed when the model is tested on the validation data.

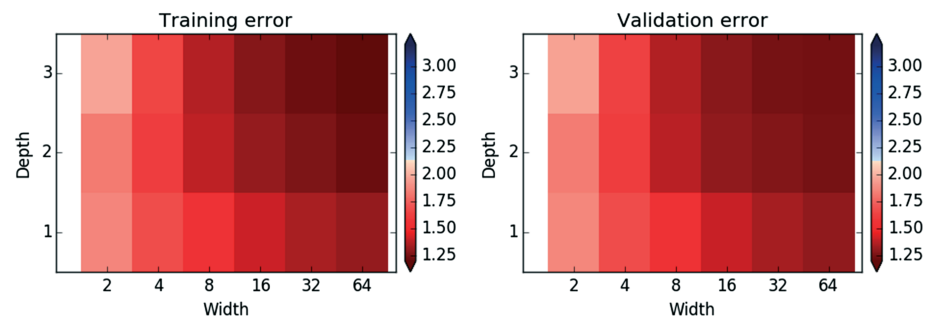
It is also interesting to compare using ANNs in this way against using ANNs that are trained to replace dynamical models or components of them, as done by Dueben and Bauer (2018). Therefore, multilayer perceptron ANNs were also trained to predict the full tendency of the Truth system  $dX_k/dt$ , given the same inputs as the ANN error correctors. Again, these predict the tendency for one  $X$  variable at a time. In an Earth system model, individual parameterization schemes could also be replaced by ANNs. However, the Lorenz '96 system lacks the complexity to do an interesting experiment where something closely analogous to replacing a parameterization scheme is carried out, given the simplicity of  $U(X)$ .

ANNs with different arrangements of neurons were tested, with one or more hidden layers (the “depth”) and with an equal number of neurons in each hidden layer (the “width”). As a shorthand notation, an ANN with depth  $D$  and width  $W$  will be referred to as “dDwW” (e.g., d2w32 refers to an ANN with depth-2 and width-32). The ANNs all use a linear output activation function and rectified linear unit activation functions on the hidden layers, which were found to work more robustly than hyperbolic tangent functions for the case of training ANNs to predict the full system tendencies. For this case, the outputs of ANNs with hyperbolic tangent activation functions were found to be prone to saturating, so that the largest tendencies could not be simulated. This may have happened because the magnitude of the output is limited to be the sum of the magnitudes of the weights connecting the final hidden layer to the output, which were not made large enough in training. This suggests that for predicting values that can take any size, using activation functions whose output values are not typically limited in magnitude is more likely to give good results.

### 2.1.3. Training ANNs

Results are presented for models using ANNs trained on 1,000 MTUs of truth model data, using the tendency over every 0.005 MTU interval, in order to test their potential skill when data availability is not a limitation. Using all 3,000 MTUs of the training, data were not found to increase the skill of ANNs substantially when





**Figure 1.** The root-mean-square error of the tendency predictions over one coarse time step (0.005 model time units) of coarse-resolution models with error-correcting artificial neural networks (ANNs) with different widths and depths evaluated on the training data (left) and validation data (right). Red indicates better performance relative to the No-ANN model and blue worse performance (with the value for the No-ANN model being the value at which the color changes). The ANNs give robust outperformance of the No-ANN model.

tested on a few chosen ANN structures. It is also shown in section 2.2.1 that the performance of the ANNs is similar at predicting tendencies in the training and validation truth data sets, indicating that the ANNs are not substantially overfitting the training data, so increasing the amount of training data would not be expected to improve the ANNs' performances much.

ANNs are trained to minimize the sum of the squared prediction error and an  $L_2$  regularisation term for the weights with coefficient  $10^{-4}$ . This was done using stochastic gradient descent with the Adam algorithm (Kingma & Ba, 2014). Minibatches of size 200 sets of input and output were used together with a learning rate of 0.001. Training stopped when the squared prediction error failed to decrease by at least  $10^{-4}$  twice consecutively after iterating over the whole training data set.

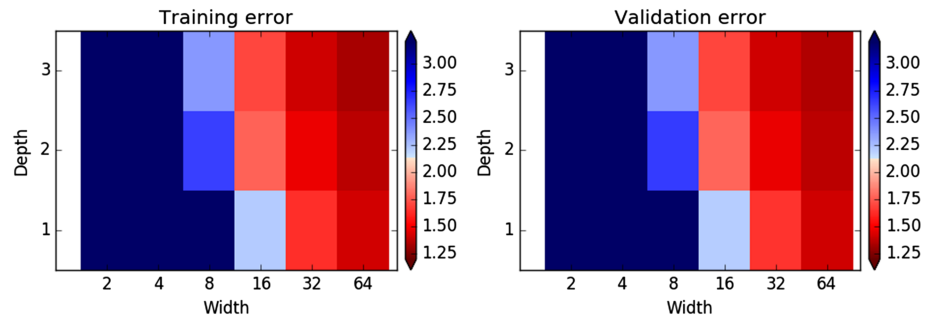
## 2.2. Results

Diagnostics comparing simulations from the Truth, No-ANN model, and models using ANNs are shown below. All results are very robust to sampling variability, as determined by checking that they are very similar when only half of the data is used, except for the difference in the mean biases (section 2.2.2) for which the use of ANNs was not found to give a statistically significant difference in most cases.

### 2.2.1. One Time Step Tendency Forecast Errors

Figure 1 shows the RMSE of tendency predictions over a single coarse time step (0.005 MTUs) for coarse-resolution models with error-correcting ANNs. Results are shown for models with ANNs with depths up to 3 and widths that are integer powers of 2 between 2 and 64 (width-1 ANNs did not generally perform well, as would be expected). The RMSE is calculated for 10,000 randomly chosen time steps in each of the training and validation data sets, using the same time steps for each ANN. The error is reduced compared to that for the No-ANN model for every ANN structure, showing that even very simple ANNs (e.g., with two neurons in a single layer) can improve the skill of predicting the tendencies. The errors do generally decrease as the ANN width and depth each increase, indicating that the optimal function relating the coarse-resolution model's tendency errors to the  $X$  variables may be quite complex. The maximum error reduction on the validation data set is 42% for the largest (d3w64) ANN, showing that ANNs can greatly reduce the error. The RMSEs on the validation data set are not more than 3% above those on the training data set, indicating that no substantial overfitting is occurring. The errors decrease as ANN size increases, and it seems likely that further error reductions are possible. The aim here is to explore how it might be made easier to get some improvement using ANNs, rather than to find the optimum performance, so results for larger ANNs and varied training hyperparameters are not shown.

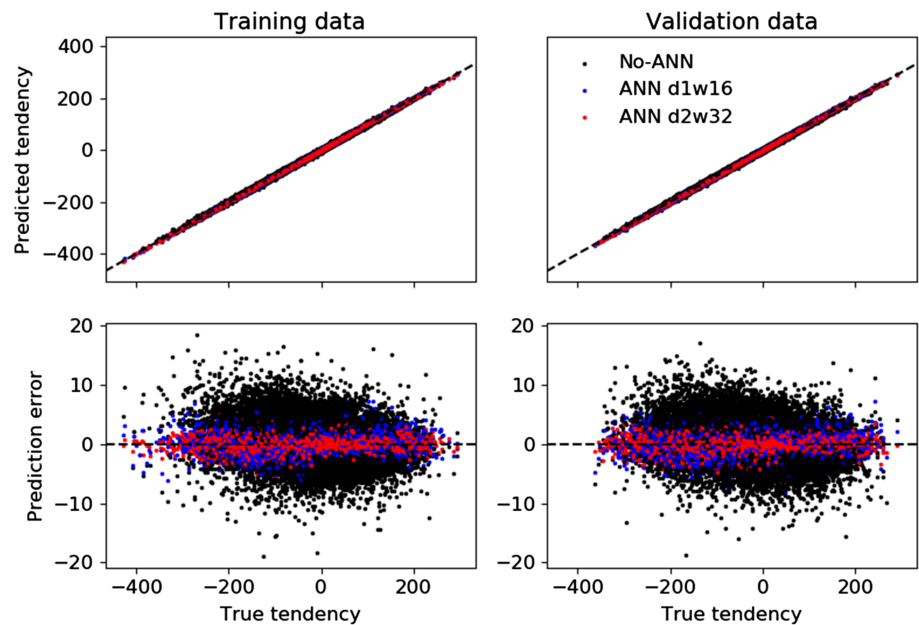
For comparison, Figure 2 shows RMSEs of tendency predictions of coarse-resolution models using ANNs to predict the full tendency. Errors are generally higher than for the models using error-correcting ANNs for a given width and depth and are only better than the No-ANN model once the ANNs become sufficiently large (with width at least 32 or width at least 16 with a depth of 2 or more). This illustrates how more complex ANNs are generally required to replace the model components rather than just to correct their errors, making it harder to achieve better performance using this approach. More parameters are also needed, increasing the risk of overfitting the data. Note that it could still be the case that the optimum attainable performance, using ANNs larger than those tested here, is better than in models with error-correcting ANNs.



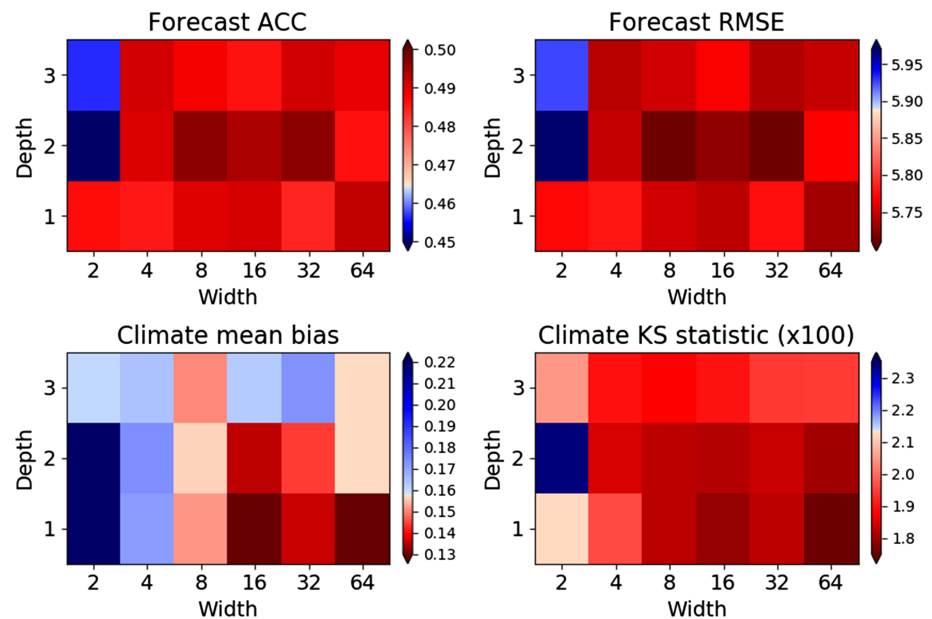
**Figure 2.** Root-mean-square errors of tendency predictions, as in Figure 1, but for artificial neural networks (ANNs) predicting the full  $X$  tendencies (section 2.1.2). Note that the color scale is saturated for the smallest ANNs. More complex ANNs are required to outperform the No-ANN model than for models using error-correcting ANNs.

Improvements upon the No-ANN model can also be obtained with much smaller amounts of training data using the error-correcting approach. Models with ANN correctors trained on just 2 MTUs of training data predicted tendencies for the validation data set with RMSEs that were robustly less than those predicted from the No-ANN model in most cases (not shown). Pure ANN models require at least about 20 MTUs to achieve this. For comparison with the typical time scales of the true system, the autocorrelation of the  $X$  variables falls to about 0.05 after a lag of 0.4 MTUs.

In order to determine if there are any situations in which the errors of tendencies predicted by the models using error-correcting ANNs are large, Figure 3 shows scatter plots of predicted tendencies and their errors versus the true tendencies. The sets of tendencies shown are from the No-ANN model and two example coarse-resolution models with error-correcting ANNs (with d1w16 and d2w32 structures, the latter performing particularly well at improving short-term forecast and climate skill scores [section 2.2.2]). One hundred thousand scatter points are shown for tendencies predicted in each of the training and validation data sets.



**Figure 3.** The top panels show the tendencies predicted by several coarse-resolution models plotted against the true tendencies in the training data set (top left) and validation data set (top right). The models are the No-ANN model and the models with depth-1 width-16 and depth-2 width-32 error-correcting artificial neural networks (ANNs). The lower panels show the prediction errors plotted against the true tendencies. The dashed lines indicate the values for perfect predictions. The tendency predictions for the models using ANNs are mostly closer to the truth than the predictions from the No-ANN model, including for rare extreme values in the validation data set, which is evidence that the ANNs have learnt to improve the representation of the dynamics.



**Figure 4.** Forecast skill and climate simulation diagnostics for models with error-correcting ANNs with different widths and depths, evaluated using the validation data set as a reference: the forecast ACC (top left) and RMSE (top right), both at lead time 1 MTU, and the climate time-mean bias (bottom left) and KS statistic (bottom right) of the  $X$  variables. Red indicates better performance relative to the No-ANN model and blue worse performance. The models with ANNs generally have a better ACC, RMSE, and KS statistic than the No-ANN model, but most differences in the mean bias were not found to be statistically significant (see text). ACC = anomaly correlation coefficient; KS = Kolmogorov-Smirnov; RMSE = root-mean-square error.

The models with error-correcting ANNs predict tendencies that are close to the true tendencies in both the training and validation data sets, including for extreme positive and negative tendencies. The predicted tendencies are generally closer to the true tendencies than those made by the No-ANN model throughout the whole range of true tendency values, including for extreme cases, although the No-ANN model also does not make any particularly large errors. This is evidence that the ANNs have learnt how to actually improve the representation of the dynamics, so that they can improve most predictions and not degrade predictions of extreme values in the validation data set even when there are few examples of the latter in the training data. This is generally the case for all of the different ANN structures, even for the smallest ANN that was tested (d1w2; not shown). It is important to show that ANNs do not simply fit the training data and perform poorly at extrapolating to make predictions for rare, extreme situations, since it is essential in Earth system modeling applications that models' performance does not severely degrade in these cases.

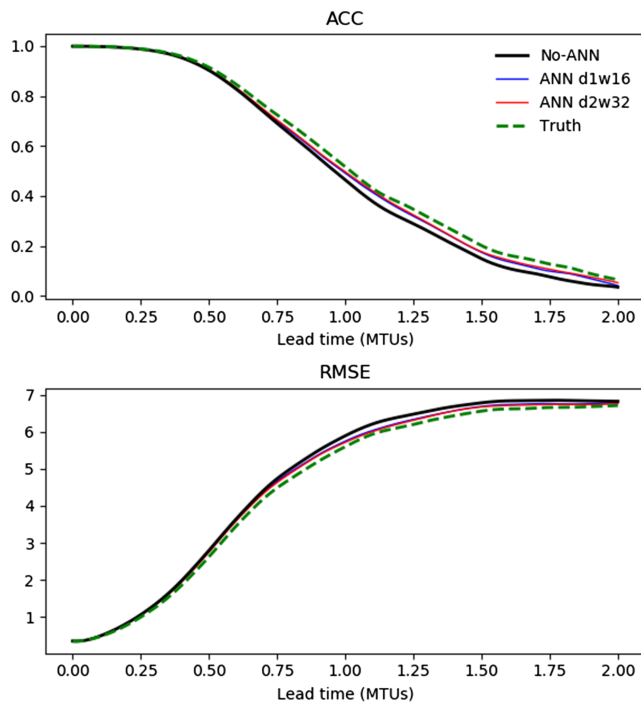
### 2.2.2. Forecast and Climate Simulation Skill

Metrics of forecast skill and the quality of the simulated climate of the  $X$  variables are shown in Figure 4 for coarse-resolution models with error-correcting ANNs of different depths and widths, evaluated using the validation data set only (this is the case for all model quality metrics shown from now on).

Forecast diagnostics were computed from 10-member ensembles of simulations initialized from each of 3,000 states of the  $X$  variables sampled from the Truth validation run, each separated by 1 MTU, giving effectively independent initial conditions. To form the initial conditions for each ensemble member for each Truth initial condition, random perturbations were sampled for each  $X$  variable independently (noting that correlations between  $X$  variables in the Truth system are small). Firstly, a sample ( $\mu$ ) was taken from a Gaussian distribution with a mean of 0 and a standard deviation of 0.05. Then 10 samples were taken from a Gaussian distribution with a mean  $\mu$  and a standard deviation of 0.05 and added to the Truth state. This ensured that the population standard deviation of the initial conditions equalled the standard deviation of the differences between their means and the Truth states, as would be expected if the perturbations came from a well-calibrated error distribution in the estimate of the initial state in a forecasting system.

The forecast anomaly correlation coefficient (ACC) and RMSE at lead time 1 MTU are better than in the No-ANN model for all models with ANNs except those with width-2 and depth-2 or 3 (Figure 4, top; squares





**Figure 5.** Forecast skill as a function of lead time evaluated on the validation data set: the anomaly correlation coefficient (top) and root-mean-square error (bottom). Results are shown for the No-ANN model, coarse-resolution models with depth-1 width-16 and depth-2 width-32 error-correcting ANNs and the Truth model, which shows the maximum potential skill with the given initial condition perturbations. The models with ANNs give modestly higher skill at lead times greater than about 0.75 MTUs, approximately halving the difference relative to the Truth model's skill. ACC = anomaly correlation coefficient; RMSE = root-mean-square error; MTU = model time unit.

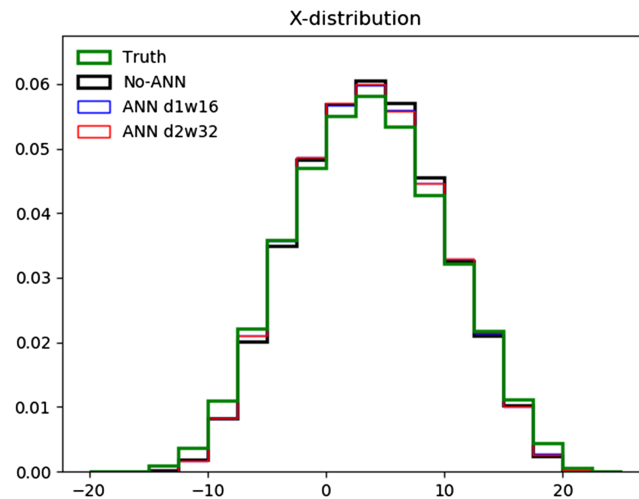
are shaded red where the metric is better than that for the No-ANN model). (Forecasts at a lead time of 1 MTU are roughly analogous to a “medium-range” forecast of the Earth’s atmosphere, given the autocorrelation time scale of the system.) Therefore, in most cases the improvement in representing the single time step tendencies (section 2.2.1) has brought about an improvement of longer range forecast skill relative to the No-ANN model. The improvement seems to be quite modest, however, raising the ACC from about 0.46 to 0.49 and decreasing the RMSE from 5.89 to 5.73 at best. However, note that the ACC for the Truth model initialized with the same initial conditions perturbations is only 0.52 and its RMSE is 5.59. This is the maximum potential skill. Therefore, the best improvements in the ACC and RMSE are slightly over 60% of the difference between the maximum possible skill and that of the No-ANN model. For the median case, they are 49% and 48% of the difference, respectively. This suggests that in a case where forecast skill were much lower than the maximum possible skill than it is here, the absolute skill improvements gained by using ANNs could be much more substantial. (Indeed, ANNs that are trained to simulate the full tendency  $dX_k/dt$  can have skill similar to the models with error-correcting ANNs tested here [not shown]. This suggests that ANNs could learn to correct the errors of a No-ANN model that were degraded to have a much lower skill level, so that the gap between the No-ANN model and the models with ANNs were much larger, though this is not tested here.)

The biases of the time mean of the  $X$  variables diagnosed from 3,000 MTU climate runs are shown in the bottom left panel of Figure 4. The diagnosed biases are mostly similar to those of the No-ANN model, except those for the models with d1w2 and d2w2 ANNs, which have much larger biases. This is the one diagnostic for which sampling variability is substantial. The biases are not statistically significantly different from that of the No-ANN model at the 95% level, except in the cases of the models with d1w2 and d2w2 ANNs. Therefore, it is difficult to be confident about how many of the models with error-correcting ANNs have smaller mean biases without using much longer climate runs, but it seems clear that the

changes in the bias are quite small overall. (The statistical significance was calculated according to a Monte Carlo permutation test; Efron & Tibshirani, 1994. Each time series was divided into blocks of length 100 MTU, which is much larger than the autocorrelation time scale of the data. For each model with an ANN, surrogate time series of length 3,000 MTU were created by selecting blocks randomly without replacement from the simulation by this model and the simulation by the No-ANN model. The probability of the absolute difference between the means of two of these time series being smaller than that between the actual time series was calculated to quantify the statistical significance.)

In order to evaluate improvements in the shape as well as the mean of the simulated climatological distribution of  $X$  values, the two-sample Kolmogorov-Smirnov (KS) statistic was calculated between the simulated distribution and the distribution in the truth model validation run. This is simply the maximum difference between the cumulative density functions of the two distributions as a function of  $X$ . The KS statistic is improved in all but the d2w2 case, by up to ~15% (Figure 4, bottom right). Part of the reason that this happens even though the bias in the mean of the distribution is not always improved is that the variance of the  $X$  values is increased relative to that in the No-ANN model (not shown), bringing the  $X$  distribution closer to that of the truth model by this measure, except in the d2w2 case.

Formal statistical significance tests were not carried out for diagnostics other than the mean bias because it seems very unlikely to get the result that models with all but the smallest ANNs seem to have improved diagnostics (Figure 4) if it were not the case that ANNs were truly producing improvements in most cases. Detailed consideration of the sampling uncertainty would be required to assess the relative skill for different ANN structures, but it is not the aim here to do this.



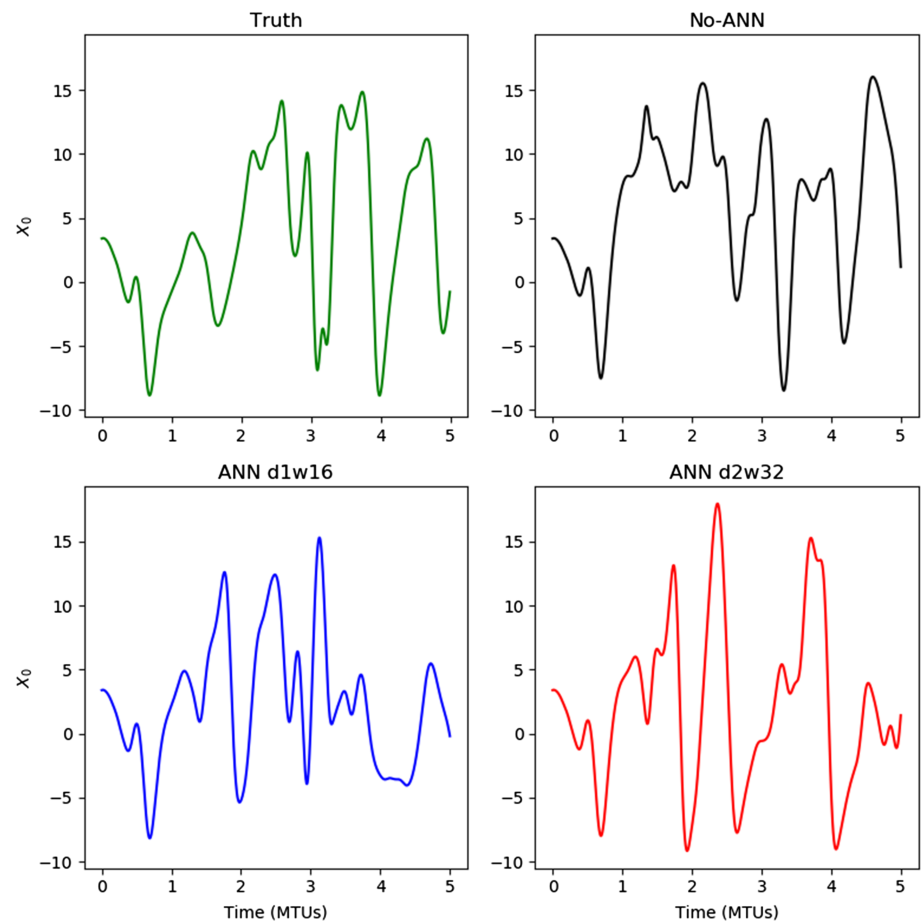
**Figure 6.** Frequency distribution of  $X$  values in long simulations in the validation data set (“Truth”) and in several coarse-resolution models: the No-ANN model and those with error-correcting artificial neural networks (ANNs) with depth-1 and width-16 and with depth-2 and width-32. The simulations by the models with ANNs have a smaller excess of frequencies of values near the center of the distribution than the No-ANN model, but all have too low a frequency of extreme values.

Altogether this indicates that the use of error-correcting ANNs in this system is able to robustly give improvements in forecast skill and the shape of the climate distribution relative to that of the No-ANN model. However, comparing Figure 4 with Figure 1 shows that improving the error of the predicted tendency does not guarantee that the quality of longer simulations will also improve. The improvements in the climate diagnostics are also smaller than might be anticipated, given the large reduction in the tendency errors that was shown in section 2.2.1.

Figure 5 shows the forecast ACC and RMSE as a function of lead time for the No-ANN model and the models with the d1w16 and d2w32 error-correcting ANNs, which are the same models that were used for Figure 3. The forecast skill is very similar for the different models up to a lead time of about 0.75 MTUs, after which the models with error-correcting ANNs begin to have higher skill than the No-ANN model. After a lead time of  $\sim 1$  MTU, their skill is approximately half way between that of the No-ANN model and the Truth model using the same initial condition perturbations. The maximum skill differences between the models with the error-correcting ANNs and the No-ANN model are about 0.04 in the ACC and 0.2 in the RMSE.

To understand better how the improvements in climate statistics shown in Figure 4 are manifested in the frequency distribution of the  $X$  variables, Figure 6 shows their distribution in the Truth validation data set, in the No-ANN model and in the previously discussed models with the error-correcting ANNs. The simulations produced by the latter have smaller frequencies near the center of the distribution, so that the bias here is smaller, with the frequencies at moderate negative values between about  $-7.5$  and  $-5$  beneficially increased. All of the coarse-resolution models have too low frequencies of large positive and negative  $X$  values, however. This may indicate that it is not possible to simulate the correct frequencies of these extremes without explicitly representing the  $Y$  variables, though it is also possible that it could be improved by applying better machine learning approaches or including stochasticity in the coarse-resolution models (Arnold et al., 2013).

On top of considering statistical summary measures of simulation skill, it is also important to verify that the temporal evolution of the system state is realistically simulated in the models with ANNs. Figure 7 shows a time series of the first  $X$  variable of length 5 MTUs at the start of the validation data set (Truth; Figure 7, top left), in the No-ANN model (top right) and in the models with the d1w16 and d2w32 error-correcting ANNs (bottom). All time series begin with the initial condition of the validation data at time zero, so that the models capture the features of the initial evolution up to about time 1 MTU, and then the model simulations diverge, likely primarily due to chaotic variability. After this point, the coarse-resolution models produce variability that appears qualitatively similar to that in the Truth system.



**Figure 7.** Time series of the first  $X$  variable in the validation run (top left) and in different coarse-resolution models initialized with the state of the validation run at time zero: the No-ANN model (top right) and models with error-correcting artificial neural networks (ANNs) with depth-1 width-16 (bottom left) and depth-2 and width-32 (bottom right). The models with ANNs capture the initial evolution of the true system, and then beyond the predictability limit they exhibit similar variability to the true system. MTU = model time unit.

To quantify the impact of using  $X_{k+2}$  as an input to the ANNs, which is not used as an input to the No-ANN model, results for error-correcting ANNs using the same inputs as the No-ANN model were analyzed. These were also found to robustly improve the metrics discussed above (again, with the exception of the mean climate bias). The improvements in short-range prediction errors relative to the No-ANN model were smaller than found above, by a median of 18% for one time step tendency errors across ANNs with the same structures as those considered above and by about 40% for the ACC and RMSE at a lead time of 1 MTU. The improvements in the climate KS statistic tended to be slightly greater, however, by a median of 5%—this is perhaps because optimizing short-range skill does not always optimize long-range skill. Overall, this illustrates that using inputs that are not presently incorporated in all Earth system model components could be a substantial source of skill added by machine learning algorithms (e.g., data in horizontally adjacent grid columns that are not typically used in subgrid parameterization schemes).

### 3. Discussion

#### 3.1. Additional Considerations for Earth System Modeling

The approach used here of training ANNs to reduce single time step tendency errors could not be applied exactly analogously to learn to better represent the dynamics of the Earth system because observations at a given location are typically spaced 6 hr or more apart, and state-of-the-art dynamical Earth system models use time steps that are much shorter. Maintaining a short time step is desirable so that the model equations can better approximate the true equations, which are continuous in time. It may also be necessary to ensure numerical stability. Therefore, an approach is required that could update the learning algorithm's param-

ters based on what would improve forecast skill over multiple time steps. Brenowitz and Bretherton (2018) achieve this when emulating convection-resolving simulations in a single column model by optimizing a cost function that takes into account errors in a prediction over multiple time steps—this is in order to make their system stable, but it may also help to improve longer range prediction skill. In a free-running system, the impact of parameter perturbations on output from previous time steps would need to be taken into account, on top of their impact in the time step corresponding to the observation. The “backpropagation through time” algorithm (Werbos, 1990) that is used in recurrent ANNs (Funahashi & Nakamura, 1993) could be used. In a model with multiple grid points, if the Earth system learning algorithm is “local,” then the effect of varying the algorithm’s parameters on predictions at nearby grid points probably needs to be taken into account as well as the effect on the predictions through multiple time steps—the backpropagation needs to be done “backward through time and sideways through space.” This is because tendencies at a given grid point depend on the system state at nearby grid points, and so prediction errors at those points at earlier time steps need to be accounted for. (It seems desirable for the algorithm to take inputs only from local grid points in order to be easier to implement in parallel computing environments, Dueben & Bauer, 2018, and to respect symmetry of the physical equations with respect to spatial translation.) For error-correcting algorithms, this approach requires the tangent linear approximation of the remainder of the model, which is related to the adjoint models that are often used in data assimilation (Errico, 1997) and have been developed for some models of Earth system components (e.g., Janisková & Lopez, 2013; Lea et al., 2015).

The data used for training algorithms also need to be considered. Dueben and Bauer (2018) suggest using reanalysis data. Although reanalysis data is imperfect, it is likely to have smaller climate biases than existing dynamical models, enabling the algorithms to yield performance improvements. A possible next step would be to recalculate the reanalysis using the improved model, combining information from this model and observations to get a yet better estimate of climate statistics. This could then be used to train better algorithms, and so on, yielding further upward steps in performance, as well as an optimal estimate of past weather given our observations.

### 3.2. Application to Problems Beyond Increasing Prediction Skill With a Stationary Climate

Error-correcting algorithms in dynamical models may be useful for addressing problems besides improving simulation skill. For example, if they do a good job at correcting large model errors, then it may be possible to understand from them how model components like conventional parameterization schemes can be improved, making use of advances in interpreting the workings of algorithms like ANNs (e.g., Ribeiro et al., 2016). They could also help to constrain stochastic parameterizations (Palmer, 2001; Watson et al., 2015) by placing an upper bound on the size of the component of the tendencies that is not predictable given the variables on the coarse grid, an irreducible error for a given model resolution, which can be modeled stochastically. Generative-adversarial algorithms (Goodfellow et al., 2014) could also find better ways to model the effects of unresolved flow stochastically (Xie et al., 2018; *Gagne II, DJ et al.*, “Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz ’96 Model,” in preparation).

The ability to vary the complexity of algorithms like ANNs in a systematic way to create a model ensemble also allows for testing of the seamless prediction paradigm—the idea that models that have better short-range prediction skill also have better long-range skill, which would mean that metrics of weather forecast skill would be informative about models’ abilities to simulate the climate response to anthropogenic forcing (Matsueda et al., 2016; Palmer et al., 2008). Alternative methods of creating an ensemble of models such as by perturbing model parameters may generally struggle to give any skill improvements, so it cannot be seen if climate simulation skill improves as short-range prediction skill gets better. In the Lorenz ’96 system studied here, correlations between the single-tendency prediction error in the validation data set (Figure 1, right) and the forecast and climate skill diagnostics shown in Figure 4 have magnitudes between 0.63 and 0.70. Correlations between the forecast RMSE at lead time 1 MTU (Figure 4, top right) and the climate mean and KS statistic (Figure 4, bottom panels) are 0.57 and 0.81, respectively. This quantifies the relationship between short-range and long-range skills in this system when using error-correcting ANNs, showing that improvements in predictions at shorter lead times do indeed tend to be associated with improvements in long-range predictions. However, as noted earlier, the correspondence is not perfect, and the improvements made to long-term climate diagnostics by using ANNs are considerably smaller than what might be expected given the improvements made to single time step tendency predictions. Therefore, the seamless prediction paradigm does not apply fully. It would be very interesting to see how well it applies

in Earth system models, given the correspondence that has been identified between biases in short-range forecasts and simulated climate (Ma et al., 2013; Sexton et al., 2019).

Another interesting question is whether using statistical learning algorithms within Earth system models could help to give more accurate simulations of the impacts of anthropogenic climate change. This is challenging because this requires making predictions about conditions that are dissimilar from those we have observed, so that a good representation of the underlying dynamics of the system is necessary. O’Gorman and Dwyer (2018) found that their emulation of a convection parameterization could not reproduce the effect of climate change well when it was trained only in a stationary “control” climate. However, statistical approaches such as optimal fingerprinting are well established in work on detection and attribution of climate change and can be used to estimate the extent to which a given model is overestimating or underestimating the response to a particular forcing (Bindoff et al., 2013). The climate change signal in individual weather events also appears clearer when dynamical variability is controlled for, which has been done previously using weather analogues (e.g., Cattiaux et al., 2010; Yiou et al., 2007). This suggests that there is scope for learning the effects of anthropogenic emissions more precisely within a model that can also accurately take into account all of the other influences on individual weather events. Even if such a model would not be trusted for projecting the impacts of large climatic changes without people being able to understand the calculations behind its predictions, it may still be useful for problems such as the attribution of observed extreme weather events (Allen, 2003; National Academies of Sciences Engineering and Medicine, 2016), for which extrapolation beyond observed conditions is not so much of a concern.

#### 4. Conclusions

It has been shown that ANNs can learn to correct errors of a coarse-resolution model of a chaotic dynamical system (the Lorenz ’96 system), resulting in stable simulations that have both improved skill in initialized forecasts and better long-term climate statistics. Improvements are found for a wide range of ANN structures, showing that they are quite robust.

The ANNs used here could reduce errors in single time step predictions by up to about 40%, and it seems that the errors could be reduced yet further if the ANNs were increased in size (Figure 1), though it is not the aim here to find the best possible performance. Errors of predicted single-time step tendencies become gradually smaller as the ANN complexity increases (Figure 1), and there does not appear to be a substantial problem due to the training getting stuck in poor local minima. The models with ANNs also give good predictions of extreme tendencies that were not seen in the model training stage (Figure 3).

In initialized medium-range forecasts, the improvement in the absolute ACC and RMSE was only a few percent at lead times of  $\sim 1$  MTU. However, this was  $\sim 50\%$  of the maximum possible improvement in the median case, determined by comparison with forecasts made by the Truth model with the same initial condition perturbations. The improvement of climate statistics was modest, with improvements up to  $\sim 15\%$  in the climate KS statistic and no discernible improvement in the time-mean state. This may be because the model without an ANN was already actually quite skilful at predicting the Truth system’s behavior—for example, Figure 3 shows that its predicted tendencies are always quite close to the true tendencies. For models of Earth’s atmosphere, coarse-graining studies find much worse agreement between tendencies predicted by models and estimated true tendencies (e.g., Shutts & Pallarès, 2014; Shutts & Palmer, 2007), suggesting that there may be much more room for improvement using machine learning. However, as discussed in section 3.1, getting large benefits in longer range skill may require training algorithms to target improvements on time scales longer than single time steps.

These results support the idea that ANNs (or other machine learning algorithms) could help to reduce errors in dynamical Earth system simulations by learning a better representation of the physical equations from observations or from more realistic models that are too expensive to use generally in weather and climate prediction (Dueben & Bauer, 2018). However, it can be far easier to use ANNs to correct the output of an existing model than to train ANNs to simulate the entire system, because far smaller ANNs can be used and much less data is required for training the ANNs (it was also shown by Jia et al., 2018, that error-correcting ANNs require less data, in the context of simulating lake temperatures). Also, Earth system models typically relate dozens of inputs and outputs at every grid point, but an error-correcting system can produce performance improvements while only considering a subset of the models’ inputs and outputs, meaning it is possible to begin demonstrating improvements without reproducing the complexity of the full model. This



is valuable because the more complex the ANN that is required, the harder it is generally to find a training method that produces good results. This method also utilizes the physical understanding embedded in the existing parameterization schemes, and the error-corrections should only become large in situations when the schemes do not perform well, reducing concerns about their reliability. This makes this approach more appropriate for use in a research program to investigate the potential for ANNs to reduce model errors and to begin producing operational improvements. The next step is to test whether the method works as well in models of components of the Earth system.

The main drawback of this approach compared to training an algorithm to simulate the full system is that the computational cost of the model cannot be reduced. Using algorithms like ANNs to learn to represent the full system's dynamics may therefore be the approach adopted in the long run, but developing systems to learn to correct model errors will give invaluable insights about how to achieve this in the medium term and help to demonstrate whether attempting to learn a better representation of the full dynamics from observations or expensive models is likely to give a substantial improvement in forecast skill. (There is also nothing to preclude an error-correcting algorithm being used in conjunction with emulators of an existing model's parameterization schemes or high-resolution simulations that do reduce the computational cost; e.g., Brenowitz & Bretherton, 2018; Chevallier et al., 1998; Gentine et al., 2018; Krasnopolsky et al., 2010; O'Gorman & Dwyer, 2018; Rasp et al., 2018). Models of the Lorenz '96 system using ANNs to predict the full tendency were found to achieve similar performance to the models with error-correcting ANNs, just requiring larger ANNs to do so (not shown). Therefore, there is nothing in the results presented here to preclude using ANNs in place of physically derived models eventually. The two methods can be used in a complementary way in a research program.

#### Acknowledgments

I thank Peter Dueben and members of Tim Palmer's research group, particularly Matthew Chantry and Jan Ackmann, for stimulating discussions about this work and also Myles Allen, Tim Woollings, and Tim Palmer for supervisory support. I also thank two anonymous reviewers for their constructive comments. I received funding from European Research Council grant 291406 and Natural Environment Research Council grant NE/P002099/1. No external data sources are required to reproduce the results presented in the manuscript. A Jupyter notebook for training and evaluating the models with ANNs can be found at the GitHub website ([https://github.com/PAGWatson/Lorenz96\\_and\\_neural\\_networks](https://github.com/PAGWatson/Lorenz96_and_neural_networks)).

#### References

- Alexander, M. J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., et al. (2010). Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum flux from observations and models. *Quarterly Journal of the Royal Meteorological Society*, 136(650), 1103–1124. <https://doi.org/10.1002/qj.637>
- Allen, M. (2003). Liability for climate change. *Nature*, 421, 891–892. [https://doi.org/10.1016/S0262-4079\(10\)62047-7](https://doi.org/10.1016/S0262-4079(10)62047-7)
- Arakawa, A. (2004). The cumulus parameterization problem: Past, present, and future. *Journal of Climate*, 17, 2493–2525.
- Arnold, H., Moroz, I., & Palmer, T. (2013). Stochastic parametrizations and model uncertainty in the Lorenz '96 system. *Philosophical Transactions of the Royal Society A*, 371, 20110479–20110479. <https://doi.org/10.1098/rsta.2011.0479>
- Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., et al. (2013). Detection and attribution of climate change: From global to regional. In T. Stocker (Ed.), *Clim. Chang. 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess. Rep. Intergov. Panel Clim. Chang.* (Vol. chap. 10, pp. 867–952). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. <https://doi.org/10.1017/CBO9781107415324.028>
- Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11, 376–399. <https://doi.org/10.1029/2018MS001472>
- Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45, 6289–6298. <https://doi.org/10.1029/2018GL078510>
- Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., & Codron, F. (2010). Winter 2010 in Europe: A cold extreme in a warming climate. *Geophysical Research Letters*, 37, L20704. <https://doi.org/10.1029/2010GL044613>
- Chevallier, C., Chérut, F., Scott, N., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology*, 37, 1385–1397. [https://doi.org/10.1175/1520-0450\(1998\)037<1385:ANNAFA>2.0.co;2](https://doi.org/10.1175/1520-0450(1998)037<1385:ANNAFA>2.0.co;2)
- Chevallier, F., Morcrette, J. J., Chérut, F., & Scott, N. A. (2000). Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, 126(563), 761–776. <https://doi.org/10.1002/qj.49712656318>
- Cooper, F. C., & Zanna, L. (2015). Optimisation of an idealised ocean model, stochastic parameterisation of sub-grid eddies. *Ocean Modelling*, 88, 38–53. <https://doi.org/10.1016/j.ocemod.2014.12.014>
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. <https://doi.org/10.1007/BF02551274>
- Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11, 3999–4009. <https://doi.org/10.5194/gmd-2018-148>
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC press.
- Errico, R. M. (1997). What is an adjoint model? *Bulletin of the American Meteorological Society*, 78(11), 2577–2592. [https://doi.org/10.1175/1520-0477\(1997\)078<2577:WIAAM>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2)
- Forsell, U., & Lindsag, P. (1997). Semi-physical and neural network modeling: An example of its usefulness. *IFAC Proceedings Volumes*, 30(11), 767–770. [https://doi.org/10.1016/s1474-6670\(17\)42938-7](https://doi.org/10.1016/s1474-6670(17)42938-7)
- Funahashi, K.-i., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6), 801–806. [https://doi.org/10.1016/S0893-6080\(05\)80125-X](https://doi.org/10.1016/S0893-6080(05)80125-X)
- Gaitan, C., Balaji, V., & Moore, B. III (2016). Can we obtain viable alternatives to Manning's equation using genetic programming? *Journal of Artificial Intelligence Research*, 5(2), 92–101. <https://doi.org/10.5430/air.v5n2p92>
- Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45, 5742–5751. <https://doi.org/10.1029/2018GL078202>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. pattern Recognit* (pp. 770–778).
- Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2016). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98, 589–602. <https://doi.org/10.1175/BAMS-D-15-00135.1>
- Janisková, M., & Lopez, P. (2013). Linearized physics for data assimilation at ECMWF. In S. K. Park, & L. Xu (Eds.), *Data Assim. Atmos. Ocean. Hydrol. Appl.* (Vol. II, pp. 251–286). Berlin, Heidelberg: Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-35088-7\\_11](https://doi.org/10.1007/978-3-642-35088-7_11)
- Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2018). Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. *ArXiv e-prints*, 1810.13075.
- Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017a). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2318–2331. <https://doi.org/10.1109/tkde.2017.2720168>
- Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017b). Physics-guided neural networks (PGNN): An application in lake temperature modeling. *ArXiv e-prints*, 1710.11431.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv e-prints*, 1412.6980.
- Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, 133(5), 1370–1383. <https://doi.org/10.1175/MWR2923.1>
- Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Hou, Y. T., Lord, S. J., & Belochitski, A. A. (2010). Accurate and fast neural network emulations of model radiation for the NCEP coupled climate forecast system: Climate simulations and seasonal predictions. *Monthly Weather Review*, 138, 1822–1842. <https://doi.org/10.1175/2009MWR3149.1>
- Krasnopolsky, V., & Lin, Y. (2012). A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US. *Advances in Meteorology*, 2012(649), 450.
- LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backProp (2nd ed.). In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks tricks trade. Lect. Notes Comput. Sci.* (Vol. 7700, pp. 9–48). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-642-35289-8\\_3](https://doi.org/10.1007/978-3-642-35289-8_3)
- Lea, D. J., Mirouze, I., Martin, M. J., King, R. R., Hines, A., Walters, D., & Thurlow, M. (2015). Assessing a new coupled data assimilation system based on the Met Office coupled atmosphere-land-ocean-sea ice model. *Monthly Weather Review*, 143(11), 4678–4694. <https://doi.org/10.1175/MWR-D-15-0174.1>
- Lorenz, E. (1996). Predictability—A problem partly solved, *Proc. Semin. Predict.* (Vol. 1, pp. 1–18). UK: ECMWF, Reading.
- Ma, H.-Y., Xie, S., Klein, S. A., Williams, K. D., Boyle, J. S., Bony, S., et al. (2013). On the correspondence between mean forecast errors and climate errors in CMIP5 models. *Journal of Climate*, 27(4), 1781–1798. <https://doi.org/10.1175/JCLI-D-13-00474.1>
- Matsueda, M., Weisheimer, A., & Palmer, T. N. (2016). Calibrating climate change time-slice projections with estimates of seasonal forecast reliability. *Journal of Climate*, 29(10), 3831–3840. <https://doi.org/10.1175/JCLI-D-15-0087.1>
- McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98(10), 2073–2090. <https://doi.org/10.1175/BAMS-D-16-0123.1>
- Morcrette, J.-J., Mozdzyński, G., & Leutbecher, M. (2008). A reduced radiation grid for the ECMWF integrated forecasting system. *Monthly Weather Review*, 136(12), 4760–4772. <https://doi.org/10.1175/2008MWR2590.1>
- National Academies of Sciences Engineering and Medicine (2016). *Attribution of extreme weather events in the context of climate change*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/21852>
- Nielsen, M. A. (2015). Neural networks and deep learning. <http://neuralnetworksanddeeplearning.com/>
- O’Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change and extreme events. *ArXiv e-prints*, 1806.11037. <https://doi.org/10.1007/s10666-012-9340-4>
- Palmer, T. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, 127, 279–304.
- Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., & Rodwell, M. J. (2008). Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society*, 89, 459–470. <https://doi.org/10.1175/BAMS-89-4-459>
- Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos*, 28(4). <https://doi.org/10.1063/1.5028373>
- Rasp, S., & Lerch, S. (2018). Neural networks for post-processing ensemble weather forecasts. *ArXiv e-prints*, 1805.09091v1.
- Rasp, S., Pritchard, M. S., & Gentile, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, 115(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Prabhat Deep learning and process understanding for data-driven Earth system science. *Nature*, 566, 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, ACM (pp. 1135–1144).
- Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, 3(4), e1602614. <https://doi.org/10.1126/sciadv.1602614>
- Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717), 2830–2841. <https://doi.org/10.1002/qj.3410>
- Schmidt, M., & Lipson, H. (2009). Distilling free-norm natural laws from experimental data. *Science*, 324(5923), 81–85. <https://doi.org/10.1126/science.1165893>
- Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44, 12,396–12,417. <https://doi.org/10.1002/2017GL076101>
- Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., et al. (2019). Finding plausible and diverse variants of a climate model. Part 1: Establishing the relationship between errors at weather and climate time scales. *Climate Dynamics*. <https://doi.org/10.1007/s00382-019-04625-3>
- Shutts, G., & Pallarès, A. C. (2014). Assessing parametrization uncertainty associated with horizontal resolution in numerical weather prediction models. *Philosophical Transactions of the Royal Society A*, 372(20130), 284. <https://doi.org/10.1098/rsta.2013.0284>

- Shutts, G. J., & Palmer, T. N. (2007). Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *Journal of Climate*, 20, 187–202. <https://doi.org/10.1175/JCLI3954.1>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354.
- Watson, P. A. G., Christensen, H. M., & Palmer, T. N. (2015). Does the ECMWF IFS convection parameterization with stochastic physics correctly reproduce relationships between convection and the large-scale state? *Journal of the Atmospheric Sciences*, 72, 236–242. <https://doi.org/10.1175/JAS-D-14-0252.1>
- Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10), 1550–1560. <https://doi.org/10.1109/5.58337>
- Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, 131(606), 389–407. <https://doi.org/10.1256/qj.04.03>
- Wu, T., & Tegmark, M. (2018). Toward an AI physicist for unsupervised learning. *ArXiv e-prints*, 1810.10525.
- Xie, Y., Franz, E., Chu, M., & Thuerey, N. (2018). tempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow. *ACM Transactions on Graphics*, 37(4), 95. <https://doi.org/10.1145/3197517.3201304>
- Xu, T., & Valocchi, A. J. (2015). Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers & Geosciences*, 85, 124–136. <https://doi.org/10.1016/j.cageo.2015.05.016>
- Yiou, P., Vautard, R., Naveau, P., & Cassou, C. (2007). Inconsistency between atmospheric dynamics and temperatures during the exceptional 2006/2007 fall/winter and recent warming in Europe. *Geophysical Research Letters*, 34, L21808. <https://doi.org/10.1029/2007GL031981>
- Zhang, S., & Lin, G. (2018). Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society*, 474(20180), 305. <https://doi.org/10.1098/rspa.2018.0305>