Applying machine learning to improve simulations of dynamical systems using empirical error correction

Peter A. G. Watson

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's been proposed that this could replace the role of physically-derived dynamical weather and climate models to give better quality simulations. Here, it is suggested instead to use machine learning in synergy with physically-derived models, by learning how to correct their errors from timestep to timestep. This maintains the physical understanding built into the models, whilst allowing performance improvements, and also requires much simpler algorithms and less training data. This method, implemented using artificial neural networks, is shown to give robust performance improvements in simulating the Lorenz '96 system. 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. Whilst 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. **??**]. The algorithms developed have been found to excel at certain problems that involve predicting an unknown value given values of predictor variables (for example, predicting what objects a photograph contains given its pixel values)—this is similar to the problem of predicting future behaviour 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 post-processing dynamical models' output [e.g. **???**].

Another emerging application is applying machine learning to improve components of the dynamical Earth system models themselves, particularly the parameterisations of unresolved small-scale processes such as radiative interactions and cloud processes. This could allow non-linear interactions between forecast output variables to be improved better than if the output is only processed after the forecast is made. This 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 non-linear) 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 with neurons arranged in layers, with the outputs of neurons in one layer being inputs to neurons in the next layer, can represent any any real continuous function to an arbitrarily small level of accuracy, given a sufficient number of neurons [**?**]. **?** and **?** provide general introductions to the theory and applications of ANNs.

One promising approach has been to use algorithms such as ANNs to reproduce the behaviour of atmospheric parameterisation schemes at a reduced computational cost. For example, **??** found that ANNs could cheaply reproduce the behaviour of the radiative transfer scheme in the European Centre for Medium-Range Weather Forecasts model, although **?** note that this approach did not work sufficiently well after the model's vertical resolution was increased. **??** also found that an ANN could be used to cheaply reproduce the output of the radiation scheme in the National Center for Atmospheric Research (NCAR) Community Atmosphere Model. More recent work has focussed 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 superparameterisation [**???**]. **?** 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 models' predictions. **?** describe how better values of parameters of models could be learnt by algorithms being fed data from observations and high-resolution models. **?** 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.

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 **?** and **?** has been presented as being in competition with improving the conventional physically-derived aspects of Earth system models. **?** argue that improving physically-derived parameterisation schemes is preferable to using ANNs because they will obey conservation laws and symmetries. **?** 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 behaviour and better trust that the model will perform reasonably well in an unseen physical situation. The specific proposal is to use these algorithms as error-correctors in dynamical models. Rather than predict the whole tendency of a system, as in the models considered by **?**, 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). $\epsilon(x)$ should only have a large effect on the simulations when the simulation by the physically-derived model is poor, when 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. The use of ANNs as the error-correctors is focussed on here, but other algorithms could 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. **?** 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 programme, 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 parameterisation 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 parameterisation schemes or increasing the model resolution, and it could reduce costs if it allows the same or greater skill to be obtained with cheaper parameterisations. 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.

In the remainder of this paper, the use of an error-correcting ANN is tested in the chaotic Lorenz '96 dynamical system [**?**] (sometimes also referred to as the Lorenz '95 system). 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. **????**]. The results are also 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. Diagnostics particularly relevant to Earth system modelling are examined, principally the skill of initialised forecasts and climate properties in long simulations.

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 $k = 1, \ldots, K$, and $j = 0, \ldots, J+2$. **?** 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 variable that favours convective activity. Here one of the systems used by **?** is simulated as the "Truth" system, with $K = 8$ and $J = 32$, in which:

$$\begin{aligned} \frac{\mathrm{d}X_k}{\mathrm{d}t} &= -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - (hc/b)\sum_{j=1}^{J} Y_{j,k}, \\ \frac{\mathrm{d}Y_{j,k}}{\mathrm{d}t} &= -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + (hc/b)X_k, \end{aligned} \tag{1}$$

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 0.001MTU using a fourth-order Runge-Kutta time-stepping scheme.

A "training" simulation of this system of length 3000MTUs (not including 10MTUs 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 10MTUs so that memory of the training dataset was effectively lost, and then a further 3000MTUs of data was generated to use as a validation dataset, which is used only to evaluate and not to develop the coarse-resolution models below.

2.1.1 Coarse-resolution model

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

$$\frac{\mathrm{d}X_k^*}{\mathrm{d}t} = -X_{k-1}^*(X_{k-2}^* - X_{k+1}^*) - X_k^* + F - U(X_k^*) \tag{2}$$

with $X_k = X_{k+K}$. The time step is also increased to 0.005MTU, so that the system has a coarsened time resolution as well.

The function $U(X_k^*)$ is derived using the same method as **?**. 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 0.005MTU, derived from the run of the truth system sampled every 0.005MTU. Its parameters were fit to minimise the root mean square error of predictions of these tendencies made using equation 2, taking values $a_0 = -0.207$, $a_1 = 0.577$, $a_2 = -0.00553$ and $a_3 = -0.000220$.

Hereafter this model is referred to as "No-ANN".

2.1.2 Models with ANNs

To produce coarse models with error-correcting ANNs, ANNs with a multilayer perceptron architecture [**??**] 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:

$$\epsilon_k = \frac{\mathrm{d}X_k}{\mathrm{d}t} - \frac{\mathrm{d}X_k^*}{\mathrm{d}t}.$$

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 takes advantage of the ability of ANNs to use inputs that are difficult to know how to include by physical reasoning—here, for example, it is difficult to hypothesise how to include X_{k+2} by inspection of equations 1, even though it may contain useful information. 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 [**?**]. 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 improve ANN performance [**?**]. The values of the mean and standard deviation are derived during training and the same values are used when the validation data is used as inputs, to ensure validation results do not depend on knowing any aspect of the validation data.

To compare the use of ANNs to correct the tendencies predicted by equation 2 with ANNs that replace equation 2, multilayer perceptron ANNs were also trained to predict the full true system tendency. Again, these predict the tendency for one X variable at a time, using the same inputs as the ANN error-correctors.

The ANNs all use rectified linear unit activation functions, which were found to work more robustly than hyperbolic tangent functions for the case of training ANNs to predict the full tendencies of the system. 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 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. 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.

2.1.3 Training ANNs

Results are presented for models using ANNs trained on 1000MTUs of truth model data, using the tendency over every 0.005MTU interval, in order to test their potential skill when data availability is not a limitation. Using all 3000MTUs of the training data was not found to increase the skill of ANNs substantially when tested on a few chosen ANN structures. It is also shown in section 2.2.1 that the performance of the ANNs at predicting tendencies in the training and validation truth datasets is similar, indicating that the ANNs are not sustantially 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 minimise the squared prediction error using stochastic gradient descent with the Adam algorithm [**?**]. 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 0.0001 twice consecutively after iterating over the whole training dataset.

Note that improvements upon the coarse-resolution model can also be obtained with much smaller amounts of training data. Models with ANN correctors trained on just 2MTUs of training data predicted tendencies for the validation dataset with root mean square errors (RMSEs) that were robustly less than those predicted from the No-ANN model alone. Pure ANN models typically require at least 20MTUs 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 0.4MTUs. This illustrates one advantage of using ANNs to correct physical models rather than do the full computation, namely that a lot less data is needed, and there is correspondingly less of a risk of overfitting the data.

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 they are very similar when only half of the data is used, except where explicitly noted.

2.2.1 One-timestep tendency forecast errors

Figure 1 shows the RMSE of tendency predictions for a single coarse time step (0.005MTUs) for coarse-resolution models with error-correcting ANNs, for ANNs with depths up to three and widths that are integer powers of 2 between 2 and 64 (width-1 ANNs did not generally perform well, as expected). The RMSE is calculated for 10,000 randomly chosen time steps in each of the training and validation datasets, 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 model's tendency errors to the X variables may be quite complex. The maximum error reduction on the validation dataset is 42% for the largest (d3w64) ANN, showing that ANNs can greatly reduce the error. The RMSEs on the validation dataset are not more than 3% above those on the training dataset, indicating that no subtantial overfitting is occurring.

For comparison, figure 2 shows RMSEs of tendency errors of coarse 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 two 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.

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 versus the true tendencies. The sets of tendencies shown are from the No-ANN model and two example coarse 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]). 100,000 scatter points are shown for tendencies predicted in each of the training and validation datasets.

The models with error-correcting ANNs predict tendencies that are close to the true tendencies in both the training and validation datasets, including for the extremely positive and negative tendencies. The predicted tendencies are generally closer to the true tendencies than for the No-ANN model throughout the whole range of true tendency values, although the No-ANN model also does not make any particularly large errors. This indicates 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 dataset even when there are few examples of the latter in the training data, as opposed to simply fitting the training data and performing poorly at extrapolation. This is generally the case for all of the different ANN structures, even for the smallest ANN that was tested (d1w2; not shown).

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 models with error-correcting ANNs of different depths and widths, evaluated using the validation dataset only (this is the case for all model quality metrics shown from now on).

Forecast diagnostics were computed from 10-member ensembles of simulations initialised from each of 300 states of the X-variables sampled from the Truth validation run, each separated by 1MTU, 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 (μ) was taken from a Gaussian distribution with mean zero and standard deviation 0.05. Then ten samples were taken from a Gaussian distribution with mean μ and standard deviation 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 1MTU 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 are shaded red where the metric is better than that for the No-ANN model). (Forecasts at a lead time of 1MTU 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 timestep tendencies (section 2.2.1) has brought about an improvement of longer range forecast skill relative to the No-ANN model. The improvement is only 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.

The biases of the time-mean of the X variables diagnosed from 3000MTU climate runs (figure 4, bottom left) is improved in more than half of cases, but there are some models with error-correcting ANNs for which this score is worse than in the model without. Note that this is the one diagnostic for which sampling variability seems substantial (not shown)—if time series of half the length are used, the bias can appear larger than for the No-ANN model for some models that are diagnosed to have smaller biases based on the 3000MTU runs. So 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 improvements in the bias are quite modest overall.

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 $\sim$15% (figure 4, bottom right). This happens because even though the bias in the mean of the distribution is not always improved, the variance of the X values is increased relative to that in the No-ANN model, bringing the X-distribution to that of the truth model by this measure, except in the d2w2 case.

Altogether this indicates that the use of error-correcting ANNs in this system is able to quite robustly give improvements in short-range 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 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 as the error-correcting ANNs used for figure 3. The forecast skill is very similar for the different models up to a lead time of about 0.75MTUs, after which the models with error-correcting ANNs begin to have higher skill than the No-ANN model. 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.

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 coarse models. Figure 6 shows a time series of X_0 of length 5MTUs at the start of the validation dataset ("Truth"; figure 6, 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 1MTU, and then the model simulations diverge, likely primarily due to chaotic variability. After this point, the coarse models produce variability that appears qualitatively similar to that in the Truth system.

To understand better how the improvements in climate statistics shown in figure 4 are manifested in the frequency distribution of the X variables, figure 7 shows their distribution in the validation dataset, 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 centre 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 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 models [**?**].

3 Discussion and conclusions

It has been shown that an artificial neural network can learn to correct errors of a coarse-resolution model of a chaotic, dynamical system (the Lorenz '96 system), resulting in improved skill at forecasting and simulating climate properties. Improvements are found for a wide range of ANN structures, showing that they are quite robust. Reductions in the errors of predicted single-timestep tendencies become gradually larger as the ANN complexity increases (figure 1), indicating that there is not a substantial problem due to the training getting stuck in local minima. The models with ANNs also give good predictions of extreme tendencies that were not seen in the model training stage (figure 3).

This gives support to 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 [**?**]. However, it is far easier to use ANNs to correct the output of a given model than to train ANNs to simulate the entire system, because much less data is required for training and far smaller ANNs can be used. Also, Earth system models typically relate dozens of inputs and outputs at every grid point, but an error-correcting system can produce performance improvements whilst 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 utilises the physical understanding embedded in the existing parameterisation 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 main drawback of this approach compared to simulating 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 a model's parameterisation schemes that do reduce the computational cost [e.g. **??????**].) 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 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). On longer time scales, the skill improvement was modest, with typical improvements of up to ∼15% in the climate KS statistic and a few percent in the anomaly correlation coefficient and root mean squared error of "medium-range" forecasts at a lead time of 1MTU. This may be because the model without an ANN was already actually quite skilful at predicting the Truth system's behaviour—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. **??**], suggesting that there may be much more room for improvement using machine learning.

3.1 Additional considerations for Earth system modelling

The approach used here of training ANNs to reduce single-timestep tendency errors could not be applied exactly analagously to learn to better represent the dynamics of the Earth system because observations at a given location are typically spaced six hours 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 parameters based on what would improve forecast skill over multiple time steps. **?** achieve this when emulating a superparameterisation scheme by optimising a cost function that takes into account errors in a prediction over multiple time steps in a single column model setting—this is in order to make their system stable, but it may also help to improve longer range prediction skill. Another approach would be to use something similar to a recurrent ANN [**?**], in which parameters can be updated using a "backpropagation through time" algorithm [**?**]. To apply either approach 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 "backwards 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 and to respect symmetry of the underlying equations with respect to spatial translation.) For error-correcting algorithms, use of the backpropagation algorithm requires the tangent linear approximation of the remainder of the model, which are related to the adjoint models that are often used in data assimilation [**?**].

The data used for training algorithms also needs to be considered. **?** 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 reanalyis 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 parameterisation schemes can be improved, making use of improvements in interpreting the workings of algorithms like ANNs [e.g. **?**]. They could also help to place 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 would be valuable for constraining the perturbations added by stochastic parameterisations [**??**]. Generative-adversarial algorithms [**?**] could also find better ways to model stochastic terms.

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 [**?**]. 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 dataset (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 1MTU (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 skill 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-range skill by using ANNs are considerably smaller than what might be expected given the improvements made to single-timestep 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 [**?**].

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. **?** found that their emulation of a convection parameterisation could not reproduce the effect of climate change well when it was trained only in a "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 over- or under-estimating the response to a particular forcing [**?**]. 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. **??**]. 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 [**??**], for which extrapolation beyond observed conditions is not so much of a concern.

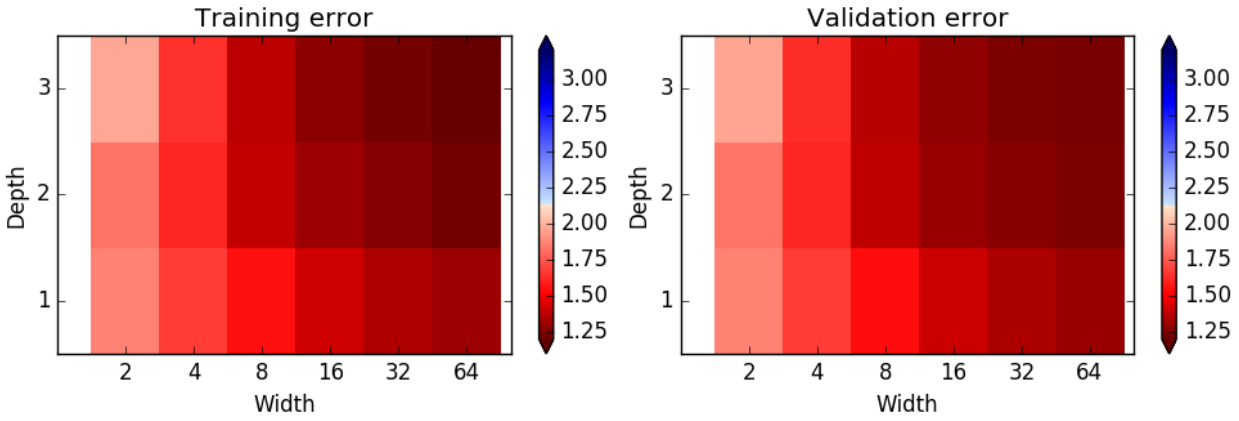

Figure 1: The root mean square error of the tendency predictions over one coarse time step (0.005MTUs) of error-correcting ANNs with different widths and depths evaluated on the training data (left) and validation data (right). Red indicates better performance relative to the No-ANN model and blue worse performance. The ANNs give robust outperformance of the coarse-resolution model.

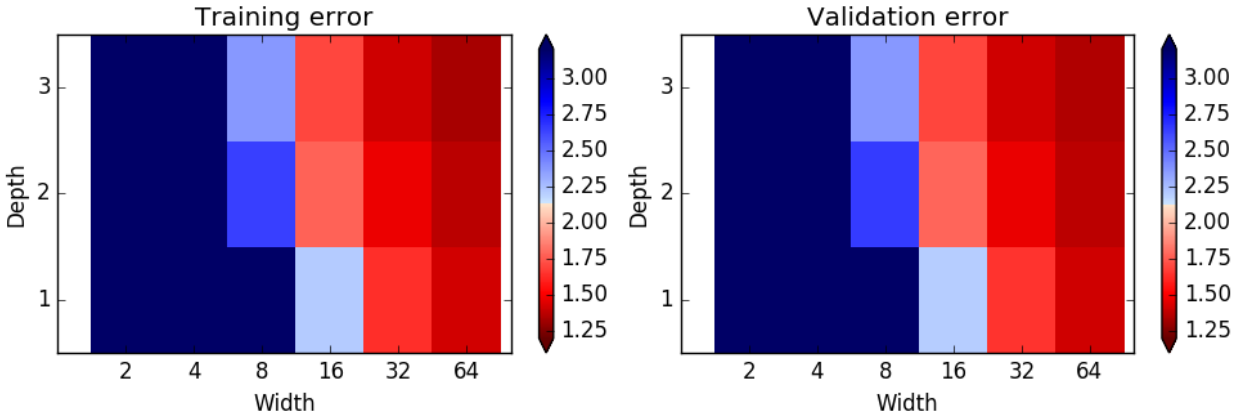

Figure 2: RMSEs of tendency errors, as in figure 1, but for ANNs predicting the full X tendencies. More complex ANNs are required to outperform the coarse-resolution model than for error-correcting ANNs.

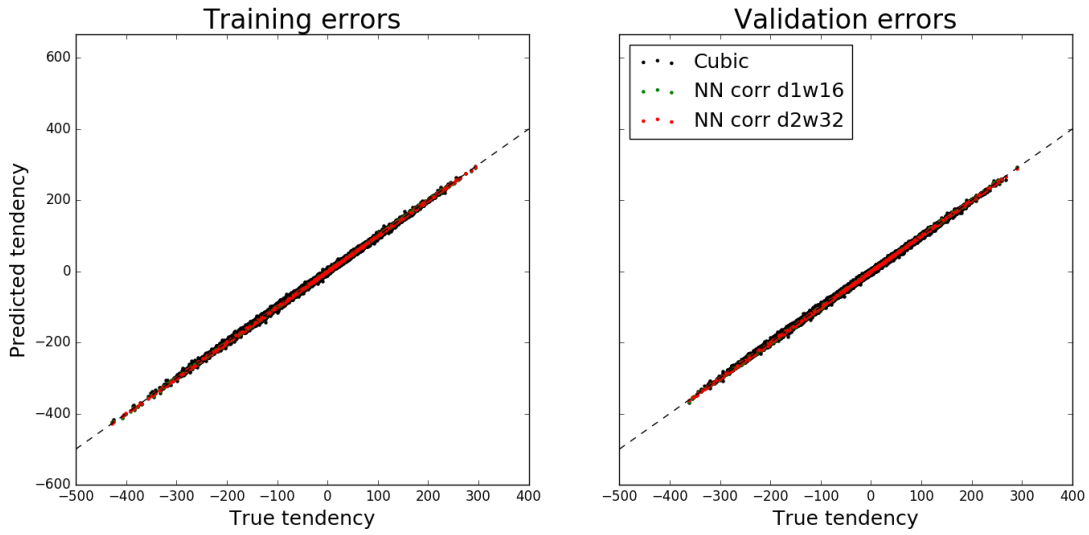

Figure 3: The tendencies predicted by several coarse-resolution models plotted against the true tendencies in the training dataset (left) and validation dataset (right). The coarse-resolution models are the No-ANN model and the models with depth-1 width-16 and depth-2 width-32 error-correcting ANNs. The dashed lines indicate the values for perfect predictions. The tendency predictions for the models using ANNs are close to the truth, including for extremely high and low values, indicating that they have learnt to represent the true dynamics well.

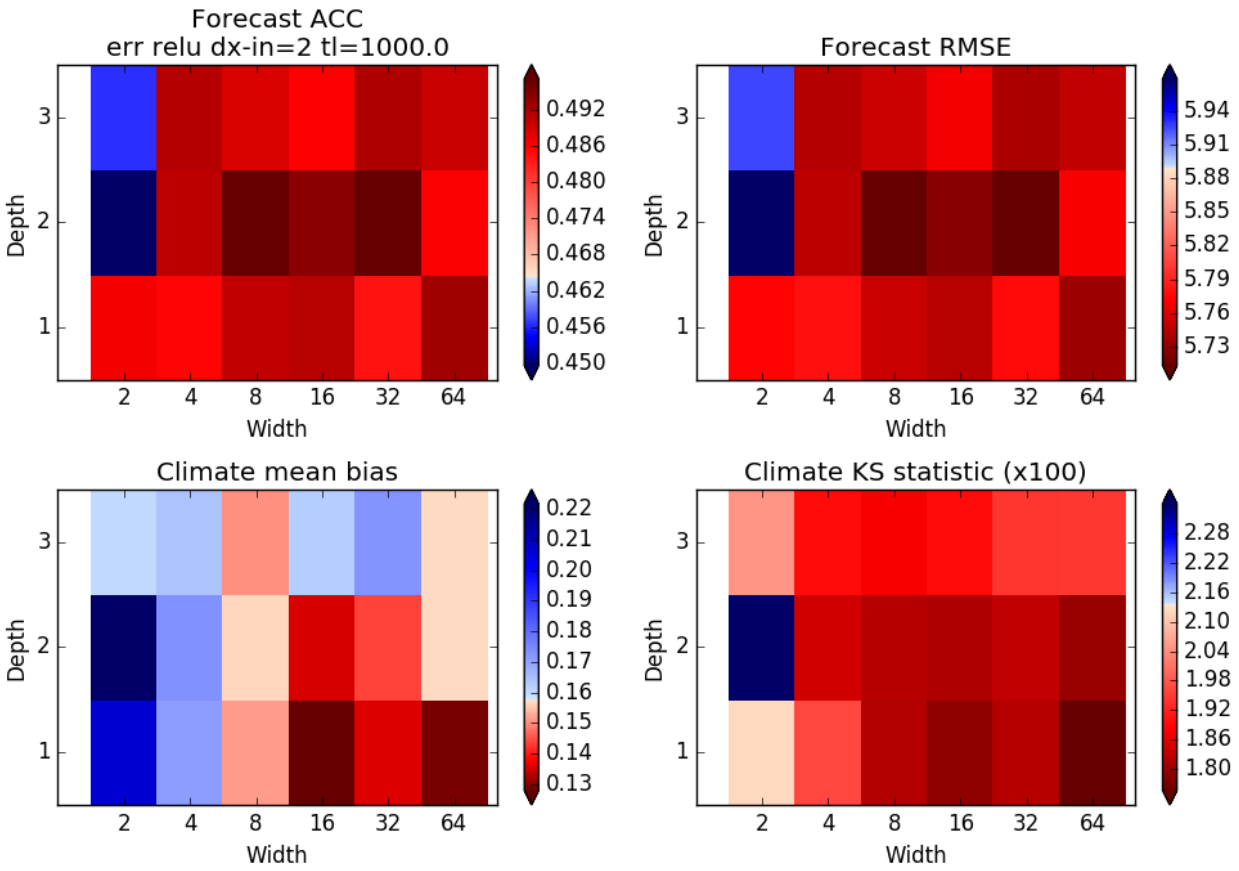

Figure 4: Forecast skill and climate simulation diagnostics for models with error-correcting ANNs with different widths and depths, evaluated using the validation dataset as a reference: the forecast anomaly correlation coefficient (top left) and root mean square error (top right), both at lead time 1MTU, and the climate time-mean bias (bottom left) and Kolmogorov-Smirnov (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 outperform the model without.

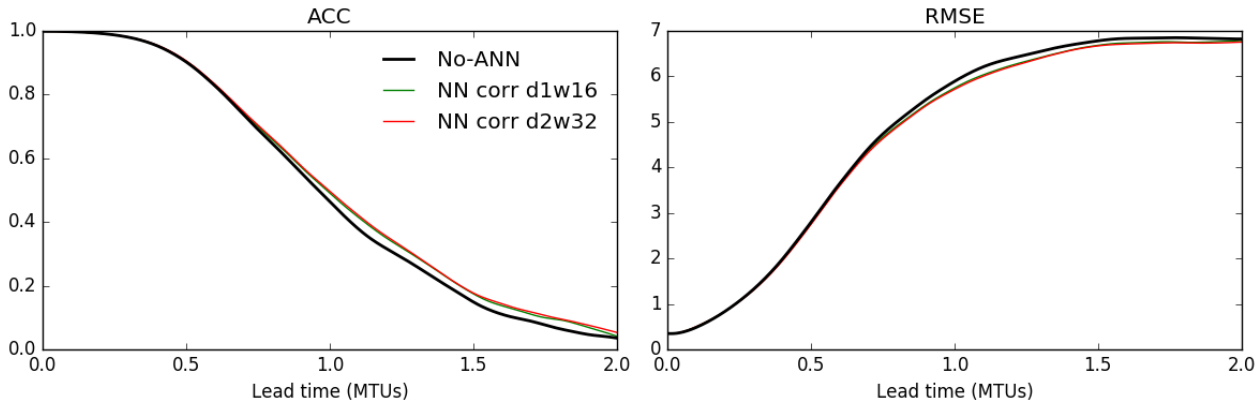

Figure 5: Forecast skill as a function of lead time evaluated on the validation dataset: the anomaly correlation coefficient (left) and root mean square error (right). The truncated models are the No-ANN model and those with depth-1 width-16 and depth-2 width-32 error-correcting ANNs. The neural network models give modestly higher skill at lead times greater than about 0.75MTUs.

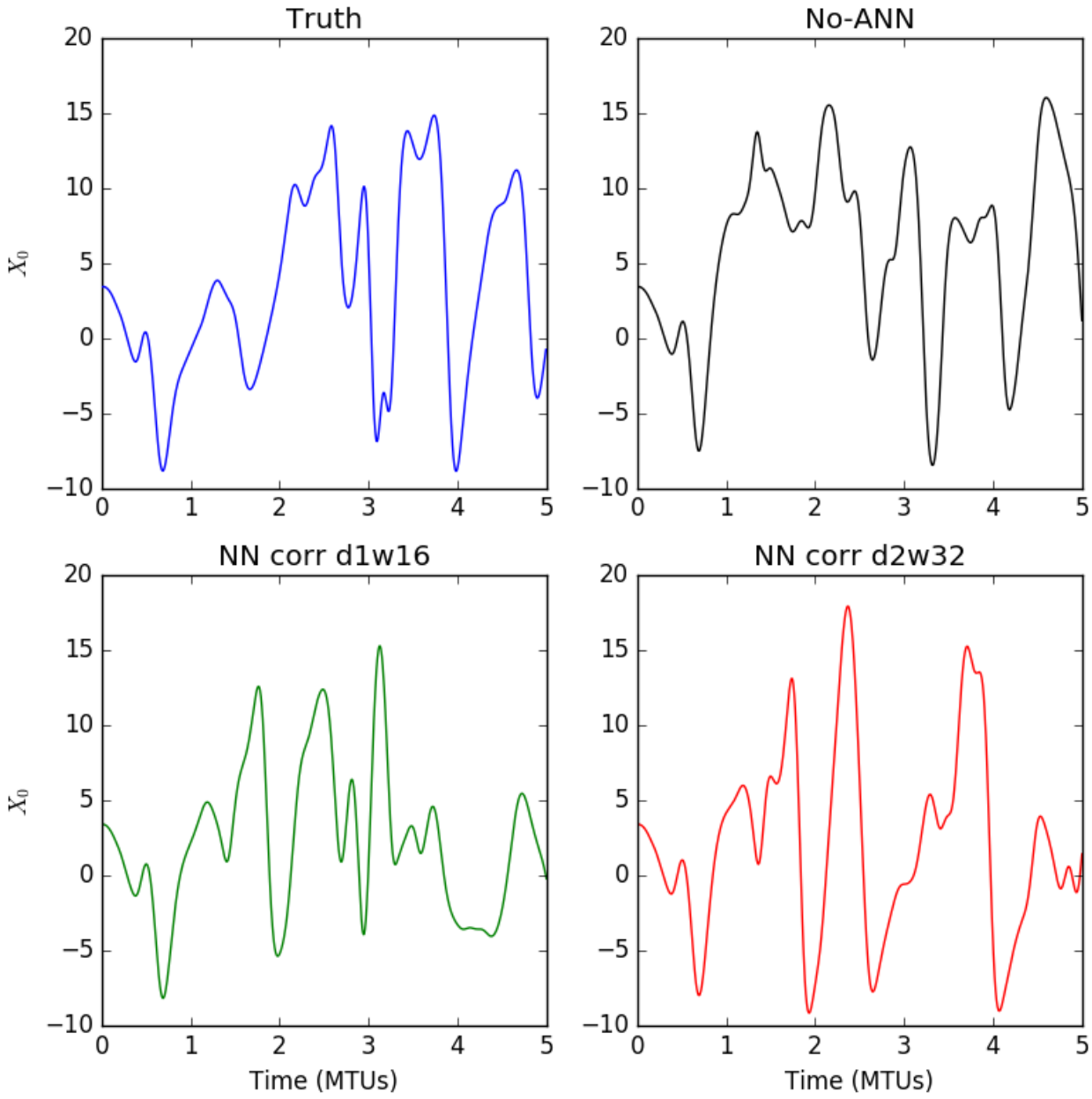

Figure 6: Time series of the first X variable in the validation run (top left) and in different coarse-resolution models initialised with the state of the validation run at time zero: the No-ANN model (top right) and models with error-correcting ANNs with depth-1 width-16 (bottom row, left) and depth-2 and width-32 (bottom row, 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.

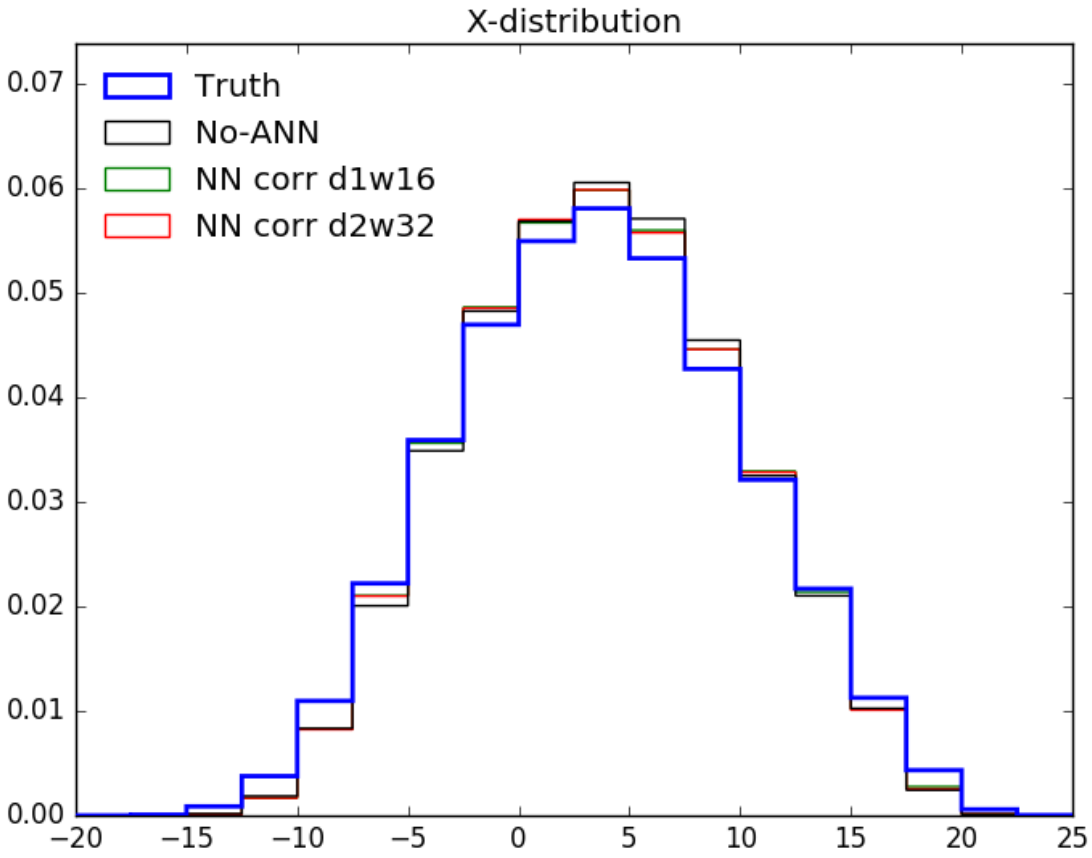

Figure 7: Frequency distribution of X values in long simulations in the validation dataset ("Truth") and in several coarse-resolution models: the No-ANN model and those with error-correcting ANNs with depth-1 and width-16 and with depth-2 and width-32. The neural networks' simulations have a smaller excess of frequencies of values near the centre of the distribution than the No-ANN model, but all have too low a frequency of extreme values.

Acknowledgements: 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 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.