A Hybrid Approach to Atmospheric Modeling that Combines Machine Learning with a Physics-Based Numerical Model

Troy Arcomano[1,1,1], Istvan Szunyogh[1,1,1], Alexander Wikner[2,2,2], Jaideep Pathak[3,3,3], Brian R Hunt[2,2,2], and Edward Ott[2,2,2]

[1]Texas A&M University
[2]University of Maryland
[3]Lawrence Berkeley National Laboratory

November 30, 2022

Abstract

This paper describes an implementation of the Combined Hybrid-Parallel Prediction (CHyPP) approach of Wikner et al. (2020) on a low-resolution atmospheric global circulation model (AGCM). The CHyPP approach combines a physics-based numerical model of a dynamical system (e.g., the atmosphere) with a computationally efficient type of machine learning (ML) called reservoir computing (RC) to construct a hybrid model. This hybrid atmospheric model produces more accurate forecasts of most atmospheric state variables than the host AGCM for the first 7-8 forecast days, and for even longer times for the temperature and humidity near the earth's surface. It also produces more accurate forecasts than a model based only on ML, or a model that combines linear regression, rather than ML, with the AGCM. The potential of the approach for climate research is demonstrated by a 10-year long hybrid model simulation of the atmospheric general circulation, which shows that the hybrid model can simulate the general circulation with substantially smaller systematic errors and more realistic variability than the host AGCM.

A Hybrid Approach to Atmospheric Modeling that Combines Machine Learning with a Physics-Based Numerical Model

Troy Arcomano[1], Istvan Szunyogh[1], Alexander Wikner[2], Jaideep Pathak[3], Brian R. Hunt[4], and Edward Ott[2,5]

[1]Department of Atmospheric Sciences, Texas A&M University, Texas, USA.

[2]Department of Physics, University of Maryland, College Park, Maryland, USA.

[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA.

[4]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland, USA.

[5]Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, USA.

Key Points:

- A hybrid model incorporating machine learning produces more accurate forecasts and more realistic climate than the host physics-based model.
- The hybrid model states are more realistically balanced and have substantially lower biases than the host model.
- The hybrid model produces more realistic atmospheric variability than the host model at time scales shorter than about a week.

Corresponding author: Troy Arcomano, troyarcomano@tamu.edu

Abstract

This paper describes an implementation of the *Combined Hybrid-Parallel Prediction (CHyPP)* approach of Wikner et al. (2020) on a low-resolution *atmospheric global circulation model (AGCM)*. The CHyPP approach combines a physics-based numerical model of a dynamical system (e.g., the atmosphere) with a computationally efficient type of *machine learning (ML)* called *reservoir computing (RC)* to construct a *hybrid model*. This hybrid atmospheric model produces more accurate forecasts of most atmospheric state variables than the host AGCM for the first 7-8 forecast days, and for even longer times for the temperature and humidity near the earth's surface. It also produces more accurate forecasts than a model based only on ML, or a model that combines linear regression, rather than ML, with the AGCM. The potential of the CHyPP approach for climate research is demonstrated by a 10-year long hybrid model simulation of the atmospheric general circulation, which shows that the hybrid model can simulate the general circulation with substantially smaller systematic errors and more realistic variability than the host AGCM.

Plain Language Summary

This paper presents a computationally efficient novel approach to construct a *hybrid model* of the atmosphere by combining a physics-based model of the global atmospheric circulation with a machine learning component. The primary purpose of the hybrid model is to produce quantitative weather forecasts on the same grid as the physics-based model. It is found that the hybrid model produces more accurate forecasts than the host physics-based model for the first 7-8 forecast days for most forecast variables, and for even longer times for the temperature and humidity near the earth's surface. Furthermore, the hybrid model is found to simulate the climate with substantially smaller systematic errors and more realistic temporal variability than the host model.

1 Introduction

Numerical weather prediction (NWP) models have been the backbone of operational weather prediction for several decades now (e.g., Lynch, 2006; Harper, 2008). A particular model implements a numerical solution algorithm for the physics-based set of coupled partial differential equations that govern atmospheric motion (e.g., Szunyogh, 2014). The resulting numerical equations form the *dynamical core* of the model. The effects of processes not resolved explicitly by the dynamical core are taken into account by *param-*

eterization schemes that contribute to the forcing terms of the equations. These schemes are based on some combination of theoretical and empirical considerations (e.g., Stensrud, 2007). The initial conditions of the numerical model solutions are observation-based estimates (analyses) of the state of the atmosphere, and the process that produces these estimates is called *data assimilation* (e.g., Szunyogh, 2014). The advances in modeling and data assimilation techniques, alongside with the increase of computing power and the number of observations available for assimilation, led to a *"quiet revolution of NWP"* (Bauer et al., 2015). The incorporation of *machine learning* (ML) techniques into the NWP process promises to lead to further forecast accuracy gains by extracting additional information from the observations.

The earliest applications of machine learning (ML) to atmospheric modeling focused on improving the computational efficiency of the physics-based numerical models (e.g., V. Krasnopolsky et al., 2005; V. Krasnopolsky & Fox-Rabinovitz, 2006; V. M. Krasnopolsky, 2013). These applications employed neural networks to emulate the computationally most expensive physics-based parameterization schemes at a reduced computational cost. The term *hybrid model* was first used in reference to models using this technique. One approach employed by this type of hybrid models is to use a single neural network to emulate the combined effect of multiple parameterized processes, such as cumulus convection, radiation, boundary layer transport, etc. (e.g., V. Krasnopolsky et al., 2010; V. M. Krasnopolsky, 2013; Brenowitz & Bretherton, 2018, 2019; Rasp et al., 2018). For this purpose, the ML systems are often trained on data produced by model simulations at higher resolutions, or with more sophisticated physical parameterization schemes.

Another type of ML-based parameterization scheme (e.g., Gentine et al., 2018; Rasp et al., 2018; Chattopadhyay et al., 2020), is trained on observations or observations-based reanalyses. Such a scheme has the potential to learn about the effects of processes that the higher resolution and more sophisticated model simulations are still unable to capture. ML techniques have also been considered for the estimation of the free parameters of physics-based parameterization schemes (Schneider et al., 2017). This approach takes advantage of the knowledge built into the parameterization schemes, but may suffer from the assumptions and approximations made by the schemes.

The hybrid approach we propose belongs to a class of techniques that are different from those mentioned thus far. Techniques of this class use ML for the *frequent pe-*

riodic interactive correction of the spatiotemporally evolving physics-based numerical model solution after training on observational analyses. The specific approach we propose was originally developed by Pathak, Wikner, et al. (2018) and later adapted to large dynamical systems by Wikner et al. (2020), who named it *Combined Hybrid-Parallel Prediction* (CHyPP). It evolves the hybrid forecasts iteratively, combining a short-term (e.g., 6 h) numerical forecast with a state-dependent ML correction in each "time step" of the "hybrid model integration". CHyPP is not a postprocessing technique, because each "time step" of the evolving hybrid model solution starts from the ML-corrected state of the preceding step, whereas a postprocessing technique does not interact with the evolving model solution. The ML component of CHyPP uses the computationally highly efficient parallel *reservoir computing (RC)* algorithm of Pathak, Hunt, et al. (2018). The other hybrid approaches of the same class use either a random forest (Watt-Meyer et al., 2021) or use a deep learning ML component (Farchi et al., 2021), rather than one based on RC.

Wikner et al. (2020) demonstrated the potential of CHyPP for predicting the evolution of a spatiotemporally chaotic system by experiments with the Kuramoto-Sivashinsky (KS) model (Sivashinsky, 1977), a model that has a single state variable that depends only on a single space dimension in addition to time. We implement CHyPP on the *Simplified Parameterization, primitive-Equation Dynamics (SPEEDY)* (Molteni, 2003; Kucharski et al., 2006) atmospheric global circulation model (AGCM). Ours is the first implementation of the approach on a model that has multiple state variables with a wide range of values and depend on all three spatial dimensions. Because SPEEDY has a substantially lower resolution than a state-of-the-art NWP or climate model, our primary goal is to demonstrate the feasibility and potentials of CHyPP for an atmospheric application, rather than to propose our current model as a potential replacement for a state-of-the-art numerical model. The results of our forecast experiments show that the performance of the hybrid model is superior to that of either SPEEDY, a model based only on ML, or a model that uses linear regression rather than ML for the correction of the short term ("one time step") numerical forecasts.

In what follows, we first describe the hybrid approach and its implementation on SPEEDY in detail (section 2). Then, we discuss the results of the forecast experiments (section 3), and then the climate simulation (section 4). Finally, we summarize our key findings and draw our conclusions (section 5).

2 The Hybrid Model

In CHyPP, the physics-based numerical model state is evolved globally, while the ML correction is done in parallel, in small local domains (Pathak, Hunt, et al., 2018). The model state of a local domain is represented by a local state vector composed of the relevant components of the global state vector. The global hybrid prediction is obtained by piecing together the local hybrid predictions at the end of each Δt-long "time step" of the "hybrid model integration". This approach can be implemented on any numerical model by adjusting the definition of the local state vectors to the spatial discretization strategy of the model. We note that the localization strategy of CHyPP is similar to that employed by the Local Ensemble Transform Kalman Filter (LETKF) data assimilation scheme (Ott et al., 2004; Hunt et al., 2007; Szunyogh et al., 2008), which has been found to scale efficiently even for very high (kilometer) resolution operational weather prediction models (e.g., Schraff et al., 2016).

2.1 The Global State Vector

SPEEDY is a spectral transform AGCM that was developed to produce rapid climate simulations, using simplified, but modern physical parameterization schemes (Molteni, 2003). We implement CHyPP on the standard configuration of Version 41 of the model: the spectral horizontal resolution is T30, while the grid used for the computation of the nonlinear terms and parameterizations has a nominal horizontal spatial resolution of $3.75^\circ \times 3.75^\circ$ with state variables defined at eight vertical σ-levels (0.025, 0.095, 0.20, 0.34, 0.51, 0.685, 0.835, and 0.95), where σ is the ratio of pressure to the surface pressure. The three-dimensionally varying state variables of the model are the two components of the horizontal wind vector, temperature, and specific humidity, while the single two-dimensionally varying state variable is the natural logarithm of surface pressure. The global computational grid and the state variables of the hybrid model are the same as those of SPEEDY.

2.2 The Local State Vectors

In our implementation of CHyPP on SPEEDY, each local state vector represents the atmospheric state in a three-dimensional local domain that has the shape of a rectangular box with a $7.5^\circ \times 7.5^\circ$ (2×2 horizontal grid points) base and extends vertically from ground level to $\sigma = 0.025$. (The boundaries of the horizontal footprint of a local

domain are marked by a blue rectangle in Fig. 1.) In what follows, we describe the computations carried out in parallel for each of the $L = 1,152$ local domains to evolve the hybrid model state from time t to $t + \Delta t$.

Let $\mathbf{v}(t)$ be the local state vector for an arbitrary local domain at time t. The dimension of this state vector is $4\times(8\times4+1)=132$ (resulting from the 4 grid points of a local domain, the 8 σ-levels, the 4 volume distributed state variables, and the natural logarithm of surface pressure state variable). Because the different state variables have different units and ranges of values, where the ranges also depend on the geographical location and vertical level, each grid-point value of each state variable is standardized to have a mean of 0 and a standard deviation of 1 before forming $\mathbf{v}(t)$. The standardization is done by using ERA5 reanalysis data (Hersbach et al., 2020) for the computation of the climatological mean and standard deviation of each grid-point variable. We introduce the notation $\mathbf{v}^p(t)$, $\mathbf{v}^h(t)$, and $\mathbf{v}^a(t)$ for the local state vector of SPEEDY, the hybrid model, and the reanalysis, respectively. We also introduce the notations $\mathbf{v}^{gp}(t)$, $\mathbf{v}^{gh}(t)$, and $\mathbf{v}^{ga}(t)$ for the related global state vectors. For instance, the components of $\mathbf{v}^{ga}(t)$ in an arbitrary local domain are the components of $\mathbf{v}^a(t)$. In what follows, we explain the steps of the computation of $\mathbf{v}^{gh}(t+\Delta t)$ from $\mathbf{v}^{gh}(t)$. A flowchart of these steps is shown in Fig. 2.a.

2.3 Reservoir Dynamics

The ML model uses (RC) (Jaeger, 2001; Lukoševičius & Jaeger, 2009; Lukoševičius, 2012) to evolve the ML model component from time t to $t + \Delta t$. In RC, the ML model state is evolved by a high-dimensional dynamical system which, for our RC implementation, is defined by the discrete time map

$$\mathbf{r}(t + \Delta t) = \tanh\left[\mathbf{A}\mathbf{r}(t) + \mathbf{B}\mathbf{u}^{\mathbf{h}}(t)\right]. \tag{1}$$

This dynamical system is the *reservoir*, $\mathbf{r}(t)$ is the *reservoir state vector*, and $\mathbf{u}^h(t)$ is the local input state.

During the training, the input term $\mathbf{u}^h(t)$ in Eq. (1) is replaced by$\mathbf{u}^a(t)$. The local input $\mathbf{u}^h(t)$ in our case is a m-dimensional *extended local state vector*, composed of the components of the local state vector $\mathbf{v}^h(t)$ plus additional components of the global state vector $\mathbf{v}^{gh}(t)$ from the neighboring local domains (see Fig. 1 for illustration), plus the prescribed incoming solar radiation at the top of the atmosphere for the extended

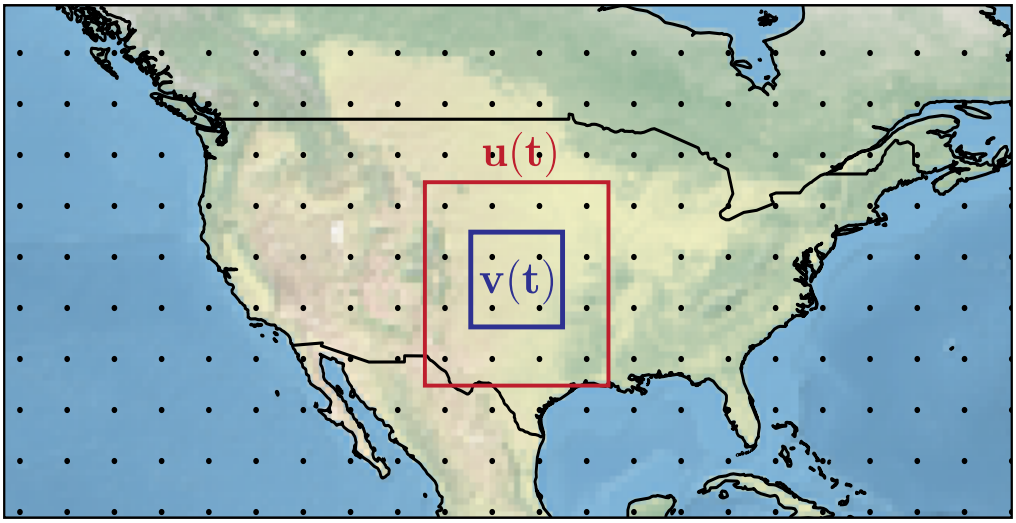

Figure 1. Illustration of the localization strategy. The black dots indicate the horizontal locations of the grid-points of the model. The blue rectangle marks the horizontal boundaries of a particular local domain. The red rectangle indicates the horizontal boundaries of the associated extended local domain.

local domain. The latter component is included to help the hybrid model to learn the diurnal cycle from the input data. (SPEEDY uses the daily average value of the incoming solar radiation at the top of the atmosphere at all times of the day.) For all of the local domains, $m = 16 \times (8 \times 4 + 1 + 1)$, except at the local domains adjacent to the poles where $m = 12 \times (8 \times 4 + 1 + 1)$.

Referring to Eq. (1), the dimension D_r of the vector $\mathbf{r}(t)$ is much higher than that of a local state vector $\mathbf{v}^h(t)$ (e.g., 6,000 vs. 132 in the present article). The *activation function* with a vector argument, $\tanh[\cdot]$, is a vector of the same dimension (D_r) as its argument, and a component of this vector is the hyperbolic tangent of the corresponding component of the argument vector. The matrix $\mathbf{A}$ is a sparse $D_r \times D_r$ *weighted adjacency matrix* that represents a low-degree, directed, random graph (Gilbert, 1959). Each entry of $\mathbf{A}$ is randomly chosen with a probability κ/D_r of being nonzero, where κ is the degree of the graph (the average number of incoming connections per node), and with the nonzero entries of $\mathbf{A}$ randomly drawn from a zero-mean uniform distribution. (The ratio κ/D_r is a measure of the *sparsity* of $\mathbf{A}$.) After randomization, the entries of $\mathbf{A}$ are scaled such that the largest eigenvalue of $\mathbf{A}$ is a prescribed number ρ ($0 < \rho < 1$), which is called the *spectral radius*. The spectral radius controls the length of the memory of

the ML reservoir, and a value $\rho < 1$ typically makes the reservoir state $\mathbf{r}(t)$ depend only on the past states of the modeled system (the atmosphere in our case), and not on the initial reservoir state, when t is sufficiently large. This property of the reservoir is called the *echo state property* (Jaeger, 2001).

The matrix-vector product $\mathbf{B}\mathbf{u}^h(t)$ is called the *input layer* in RC. In our model, $\mathbf{B}$ is a $m \times D_r$ sparse random matrix with an equal number of nonzero entries in each row. These nonzero entries, which are chosen randomly from a uniform distribution on the interval $[-\alpha, \alpha]$, couple the components of $\mathbf{u}^h(t)$ to the reservoir nodes. The *input strength* α is an adjustable parameter that controls the degree of non-linearity experienced by the input signal $\mathbf{u}^h(t)$ from the activation function.

2.4 The Hybrid Model

In addition to providing the input for Eq. (1), the global state $\mathbf{v}^{gh}(t)$ is used as the initial condition for a SPEEDY model forecast $\mathbf{v}^{gh}(t+\Delta t)$. The next local hybrid model prediction is then obtained by

$$\mathbf{v}^h(t+\Delta t) = \mathbf{W} \begin{bmatrix} \mathbf{v}^p(t+\Delta t) \\ \widetilde{\mathbf{r}}(t+\Delta t) \end{bmatrix}, \tag{2}$$

where the components $\widetilde{r}_i(t+\Delta t)$ of the column vector $\widetilde{\mathbf{r}}(t+\Delta t)$, $i = 1, 2, ...D_r$ are defined by $\widetilde{r}_i(t+\Delta t) = r_i(t+\Delta t)$, if i is odd, and $\widetilde{r}_i(t+\Delta t) = r_i^2(t+\Delta t)$, if i is even, and the column vector $\mathbf{v}^p(t+\Delta t)$ represents the local state corresponding to the global SPEEDY forecast $\mathbf{v}^{gp}(t+\Delta t)$. The matrix-vector product on the right-hand side of Eq. (2) is the RC *output layer*. The matrix $\mathbf{W}$ is a *matrix of parameters* to be determined by the training procedure described in Sec. 2.4.1. The local vectors $\mathbf{v}^h(t+\Delta t)$ for each local domain are combined to form the next global hybrid model prediction $\mathbf{v}^{gh}(t+\Delta t)$.

Equation (2) can be written in the equivalent form

$$\mathbf{v}^h(t+\Delta t) = \mathbf{W}_{mod}\mathbf{v}^p(t+\Delta t) + \mathbf{W}_{res}\tilde{\mathbf{r}}(t+\Delta t), \tag{3}$$

which corresponds to $\mathbf{W} = [\mathbf{W}_{mod} \ \mathbf{W}_{res}]$. In the extreme case that $\mathbf{W}_{mod} = \mathbf{0}$, which should be the result of training when the numerical model has no skill according to the training data, the hybrid prediction completely ignores the numerical model forecast $\mathbf{v}^p(t+\Delta t)$. The other extreme case is when $\mathbf{W}_{mod} = \mathbf{I}$ and $\mathbf{W}_{res} = \mathbf{0}$, which should occur when the numerical model is perfect according to the training data. In a typical case,

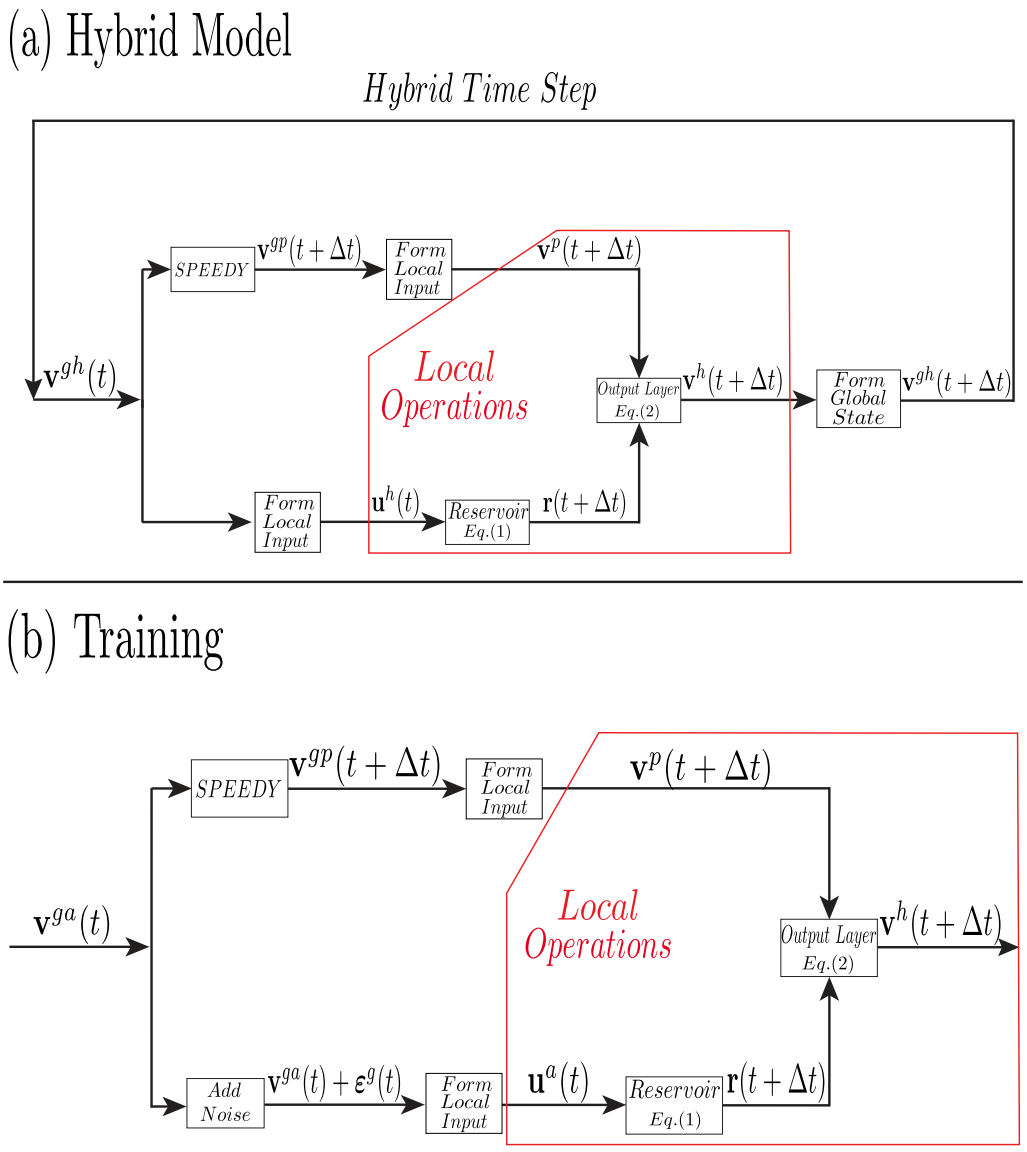

Figure 2. A flow chart of (a) the hybrid model and (b) the training operation of the hybrid model. The notation is defined in Secs. 2.2 and 2.3. The steps inside the red boxes are carried out in parallel for each of the $L = 1,152$ local domains. The training finds the $\mathbf{W}$ that minimizes the cost function of Eq. (4) by solving Eq. (5).

which falls between the two extremes, the ML output and the Δt-long numerical prediction are combined to maximize agreement with the training data.

2.4.1 Training

Figure 2.b shows the flow of operations during training. First, we generate a sequence of perturbed global analyses $\mathbf{v}^{ga}(k\Delta t) + \boldsymbol{\varepsilon}^{g}(k\Delta t)$, $k = -K - K_t, -K - K_t + 1, ..., -1$, where $\boldsymbol{\varepsilon}^{g}(k\Delta t)$ is a small-magnitude, zero-mean, normally distributed random noise vector, uncorrelated in time and uncorrelated between components of the noise vector. The role of this noise is to help the ML model learn to return to the bounded set of realistic atmospheric states (the "attractor") in the presence of perturbations that may arise in future forecasts (e.g., Jaeger, 2001; Wikner et al., 2020). The addition of noise to the global analyses during training is essential for the hybrid model to produce stable, realistic predictions; predictions rapidly become unstable without it. Similar behavior has been observed in RC applications involving the prediction of other spatio-temporal systems (e.g., Patel et al., 2021).

The local input state $\mathbf{u}^{a}(k\Delta t)$ is the extended local state vector associated with $\mathbf{v}^{ga}(k\Delta t) + \boldsymbol{\varepsilon}^{g}(k\Delta t)$, for $k = -K - K_t, -K - K_t + 1, ..., -1$ for the particular local domain. The initial state $\mathbf{r}[(-K - K_t)\Delta t]$ of the reservoir can be chosen arbitrarily, because only the evolved reservoir states $\mathbf{r}[(k+1)\Delta t]$, $k = -K, -K+1, \ldots, -1$, are used for training. The purpose of discarding the reservoir state of the first K_t ($K_t \ll K$) iterations is to ensure that the reservoir state $\mathbf{r}(t)$ has sufficient time to settle on its attractor. The unperturbed global analyses $\mathbf{v}^{ga}(k\Delta t)$ are also used as the initial conditions for SPEEDY to obtain $\mathbf{v}^{gp}[(k+1)\Delta t]$ for $k = -K, -K+1, ..., -1$.

Formally, the training is carried out by computing the weight matrix $\mathbf{W} = [\mathbf{W}_{mod} \;\; \mathbf{W}_{res}]$ that minimizes the cost-function

$$J(\mathbf{W}) = \sum_{k=-K+1}^{0} \|\mathbf{v}^{h}(k\Delta t, \mathbf{W}) - \mathbf{v}^{a}(k\Delta t)\|^2 + \beta_{mod}\|\mathbf{W}_{mod} - \mathbf{W}_{prior}\|^2 + \beta_{res}\|\mathbf{W}_{res}\|^2. \quad (4)$$

The local hybrid states $\mathbf{v}^{h}(k\Delta t, \mathbf{W}), k = -K+1, -K+2, ..., 0$, represent the results of Eq. (2) at those times for a particular $\mathbf{W}$, and $\mathbf{v}^{a}(k\Delta t)$ is the local state vector for the unperturbed global analysis $\mathbf{v}^{ga}(k\Delta t)$. (Notice that we use the notation $\mathbf{W}$ for both the variable and the solution of the minimization problem.) The last two terms of the cost function, in which $\|\cdot\|^2$ denotes the sum of the squares of the entries of a matrix (the Frobenius norm), are regularization terms meant to prevent overfitting, with β_{mod} and β_{res} being the regularization parameters for the numerical model and reservoir component, respectively. With these terms, the direct solution of the least-square problem is a *ridge regression* (Tikhonov & Arsenin, 1977). The inclusion of the *prior matrix* $\mathbf{W}_{prior}$,

which was not part of Wikner et al. (2020), allows for a choice like $\mathbf{W}_{prior} = \mathbf{I}$, which dictates that in the absence of training data that demonstrates imperfections in the numerical model, the hybrid model should be equivalent to the numerical model. In our experiments, we tried both $\mathbf{W}_{prior} = \mathbf{I}$ and $\mathbf{W}_{prior} = \mathbf{0}$, and found that the latter yielded better stability. Thus, we report results with $\mathbf{W}_{prior} = \mathbf{0}$, but think that other choices for nonzero $\mathbf{W}_{prior}$ merit further study.

To obtain the direct solution for the matrix $\mathbf{W}$ that minimizes the cost function J, we define matrix $\widetilde{\mathbf{R}}$ by choosing its column k to be $\widetilde{\mathbf{r}}(k\Delta t)$ (see Eq. (2)), and matrix $\mathbf{V}_p$ by choosing its column k to be the $\mathbf{v}^p(k\Delta t)$ local state vector that corresponds to the global SPEEDY forecast from $\mathbf{v}^{ga}((k-1)\Delta t)$. In addition, we define matrix $\mathbf{V}_a$ by selecting its column k to be the local analysis $\mathbf{v}^a(k\Delta t)$. Then, it can be shown that the minimizing $\mathbf{W}$ is the solution of the linear problem

$$\mathbf{W}\begin{bmatrix} \mathbf{V_p}\mathbf{V_p^T} + \beta_{mod}\mathbf{I} & \mathbf{V_p}\widetilde{\mathbf{R}}^{\mathbf{T}} \\ \widetilde{\mathbf{R}}\mathbf{V_p^T} & \widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^{\mathbf{T}} + \beta_{res}\mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{V_a}\mathbf{V_p^T} + \beta_{mod}\mathbf{W}_{prior} & \mathbf{V_a}\widetilde{\mathbf{R}}^{\mathbf{T}} \end{bmatrix} \tag{5}$$

for $\mathbf{W}$.

Because the dimension of the matrix products in this problem does not depend on the length $K\Delta t$ of the training period, the matrix products can be computed incrementally, without simultaneously storing every column of $\widetilde{\mathbf{R}}$, $\mathbf{V}_p$, or $\mathbf{V}_a$ in memory (e.g., Lukoševičius, 2012). That is, in terms of computer memory usage, the resources used by the training do not depend on the length of the training period. This is a highly desirable property for Earth system modeling, in which long training periods are expected to be necessary. In addition, the corresponding columns of $\tilde{\mathbf{R}}$, $\mathbf{V}_p$, and $\mathbf{V}_a$ can be obtained by training on multiple time series of training data. For example, suppose that the global analyses $\mathbf{v}^{ga}(t)$ have a temporal resolution Δt_a that is finer than the Δt temporal resolution of the hybrid model with $\Delta t = J\Delta t_a$, where J is an integer. Then, the number of time series available for training is J; i.e., the first term in Eq. (4) can be replaced by

$$\sum_{j=0}^{J-1} \sum_{k=-K+1}^{0} \|\mathbf{v}^h(k\Delta t - j\Delta t_a, \mathbf{W}) - \mathbf{v}^a(k\Delta t - j\Delta t_a)\|^2. \tag{6}$$

2.4.2 Synchronization and Prediction

Let $K_f\Delta t$ be the forecast start time. Starting the hybrid forecast requires the availability of the global analysis $\mathbf{v}^{ga}(K_f\Delta t)$ and the reservoir state $\mathbf{r}(K_f\Delta t)$ for each local domain. Because according to the "echo state property" $\mathbf{r}(K_f\Delta t)$ is determined by the

past states of the atmosphere, it can be obtained by synchronizing the evolution of the reservoir states with the analyses for a sufficiently long time period that ends at $K_f\Delta t$. Let $K_s\Delta t$ be the start time of the synchronization. Synchronization is achieved by evolving the reservoir equation using $\mathbf{u}^h(k\Delta t) = \mathbf{u}^a(k\Delta t)$ in Eq. (1) for $k = K_s, K_{s+1}, \ldots, K_f$.

Piecing together the local hybrid forecasts for all local domains yields the global "one-step" hybrid forecast $\mathbf{v}^{gh}[(K_f{+}1)\Delta t]$ (Fig. 2.a). The forecast can be extended arbitrarily far into the future by using an iterative process for $k = K_f + 1, K_f + 2, \ldots,$ in which the extended local state vector $\mathbf{u}^h(k\Delta t)$ extracted from $\mathbf{v}^{gh}(k\Delta t)$ is used as $\mathbf{u}^h(k\Delta t)$ in the Eq. (1) to compute $\mathbf{r}[(k{+}1)\Delta t]$. The global "one-step" hybrid forecast $\mathbf{v}^{gh}(k\Delta t)$ is also used as the initial condition of the $\mathbf{v}^{gh}[(k + 1)\Delta t]$ SPEEDY component of the hybrid forecast. In a cycled forecast system of an operational NWP center, in which analyses are prepared and forecasts are started with a regular frequency (e.g., 6 h), the reservoir state can be kept continuously synchronized with the real-time evolution of the atmosphere.

2.5 Implementation with ERA5 Reanalysis Data

We use interpolated hourly global ERA5 reanalyses to train and synchronize the hybrid model. We do the horizontal interpolation of the reanalysis fields onto the computational grid of SPEEDY by a 2-dimensional quadratic B-spline interpolation. We then compute the value of σ at each horizontal grid point and use a 1-dimensional cubic B-spline for the vertical interpolation of the model state variables to the eight prescribed constant σ levels of SPEEDY. The training starts at 0000 UTC on January 1, 1990 and ends at 2300 UTC on June 26, 2011 ($K \approx 3.14{\times}10^4$), with the data discarded for the first 6.25 days ($K = 31355$ and $K_t = 25$).

2.6 Selection of the Hyperparameters

Hyperparameters are adjustable parameters (e.g. κ, ρ, α, D_r,β_{res}, β_{mod}, $\boldsymbol{\varepsilon}$, and Δt) that control overall characteristics of the hybrid model and require "tuning" to produce desirable results. There exists "tricks of the trade" practical rules for the selection of the hyperparameters of an RC model (Lukoševičius, 2012). These general rules also work for the hyperparameters of the hybrid model. First, the hybrid model is only weakly sensitive to κ and ρ. While we use $\kappa = 6$, other small values of κ (e.g., $\kappa = 3$) work sim-

ilarly well. We use a value of ρ that monotonically increases toward the poles from 0.3 at the equator to 0.7 at 45°, so that the reservoir mimics the general property of the atmospheric dynamics that its memory is shorter in the tropics than the extratropics. Changing these values by $\pm 0.1 - 0.2$ has little effect on the model performance. We choose $D_r = 6,000$, because we find that further increasing the reservoir size does not lead to substantial further improvement of the model performance. We find the hybrid model performance to be somewhat sensitive to the value of α, which controls the amount of nonlinearity of the reservoir dynamics. Setting $\alpha \leq 0.3$ or $\alpha \geq 0.7$ yields noticeable degradation of the errors compared to the value we use, $\alpha = 0.5$. For each of the options $\mathbf{W}_{prior} = \mathbf{I}$ and $\mathbf{W}_{prior} = \mathbf{0}$, we tried various powers of 10 for the regularization parameters β_{res} and β_{mod}; we found that $\mathbf{W}_{prior} = \mathbf{0}$ yielded better stability, and found that $\beta_{res} = 10^{-4}$ and $\beta_{mod} = 10^{0}$ led to good model performance. Among the several values we tried, in increments of 0.05, for the standard deviation of the components of the random noise $\boldsymbol{\varepsilon}$ added to the training data, we chose the smallest value (0.20) for which all hybrid forecasts were stable. The time step Δt is another important hyperparameter to tune; we chose $\Delta t = 6$ h, because using $\Delta t = 1$ h or $\Delta t = 3$ h (with other hyperparameters tuned accordingly) led to clearly poorer model performance. Moreover, we use a time step of $\Delta t/24 = 0.25$ h for the numerical integration of SPEEDY, because longer time steps degraded the 6 h forecast performance of SPEEDY. Since the temporal resolution of the ERA5 reanalyses is 1 h ($\Delta t_a = 1$), the training is done on $\Delta t/\Delta t_a = 6$ time series of data.

3 Forecast Experiments

We compute forecast error statistics based on 100 21-day forecasts, with start times equally spaced every 4 days between 0000 UTC, June 27, 2011 and 0000 UTC, July 28, 2012. We evaluate the forecast performance of the hybrid model by comparing it to that of a variety of benchmark forecasts started from interpolated ERA5 reanalyses.

3.1 Benchmark Forecasts

The set of benchmark forecasts includes numerical forecasts produced by SPEEDY, a model based only on ML, and a model in which the 6 h SPEEDY forecasts are corrected by linear regression rather than by ML. We call the latter benchmark SPEEDY-LLR, where LLR stands for *local linear regression*.

Comparing the performance of the hybrid model to that of a model based only on ML is important, because ML-only models (e.g., Arcomano et al., 2020; Rasp & Thuerey, 2021; Weyn et al., 2020) are considered a potential alternative to the hybrid approaches for the utilization of ML in Earth syetem modeling. Our ML model is formally the same as our hybrid model except that we use the constraint $\mathbf{W}_{mod} = \mathbf{0}$ in Eq. (3), with Eqs. (4) and (5) modified accordingly, and the hyperparameters are different: $D_r = 9,000$, $\beta_{res} = 10^{-6}$, $\Delta t = 3$ h, and $\boldsymbol{\varepsilon}$ has a standard deviation of 0.28. (The smaller reservoir size necessary to obtain good results from the hybrid as compared to the ML-only model is an important advantage of the hybrid model.) While this ML-only model is formally identical to the one described by Arcomano et al. (2020), its forecast performance is better, thanks mainly to using a time step of Δt =3 h rather than Δt =1 h and the addition of the incoming solar radiation to the input of the reservoir.

The SPEEDY-LLR is the same as the hybrid model except that $\mathbf{W}_{res} = \mathbf{0}$. In this model, a larger regularization parameter is necessary to produce stable forecasts for at least 10 days. We use $\beta_{mod} = 1600$, which provides the most accurate short and medium range (1-5 days) forecasts that also remain stable for at least 10 days. The stability of the SPEEDY-LLR forecasts can be improved by further increasing β_{mod}, but only at the price of degrading the short and medium range forecast accuracy. (For $\beta_{mod} \to \infty$, SPEEDY-LLR becomes SPEEDY, which produces stable forecasts for indefinitely long lead times). Since, SPEEDY-LLR does not include the nonlinear ML correction of the hybrid model (the second term on the right side of Eq. (3)), training is a simple linear regression of the numerical model forecast. With the help of this benchmark, we can assess the relative importance of making periodic corrections to the numerical forecasts based on linear regression of the model state alone versus making those corrections by the proposed hybrid technique.

To assess whether a model forecast has skill, the figures also include comparisons to forecasts based on persistence and daily climatology. The persistence forecasts are based on the assumption that the state of the atmosphere at the beginning of the forecast persists for the entire duration of the forecast, while the climatological forecasts are based on the daily climatological mean for the calendar day at the particular geographical location and pressure level for years 1990-2010.

3.2 The Measure of the Forecast Error

The error of each forecast is measured by the area-weighted *root-mean-square error*,

$$RMSE = \sqrt{\frac{1}{N_{lon}N_{lat}} \sum_{i=1}^{N_{lon}} \sum_{j=1}^{N_{lat}} a(j)(V_{i,j}^{f} - V_{i,j}^{a})^2}, \tag{7}$$

where,

$$a(j) = \frac{\cos\left(\varphi(j)\right)}{\frac{1}{N_{lat}} \sum_{j=1}^{N_{lat}} \cos\left(\varphi(j)\right)}. \tag{8}$$

Here the subscript i, j refers to the value of a scalar state variable V for a specific forecast lead time at a particular pressure level at grid point i, j of the verification region defined by N_{lon} discrete longitudes and N_{lat} discrete latitudes. The RMSE is averaged over the 100 forecasts to obtain a single scalar measure of the forecast error for each state variable, pressure level, and forecast lead time. In what follows, the term *forecast error* refers to this scalar measure. We call a forecast more accurate than another, if the forecast error is lower for the former than the latter forecast. In addition, we say that a model forecast has *forecast value*, if its forecast error is lower than that of both persistence and climatology (the latter two are available without the substantial cost of preparing model forecasts). The qualitative behavior of the errors of the model forecasts with respect to the errors of these two references is well understood. In particular, if the model has realistic climatology, in the sense that it represents the atmospheric variability (the variability of the atmospheric state) correctly, the error of the model forecasts and the error of persistence saturate at the same level. While the error is initially lower for persistence than climatology, its saturation value is higher by a factor of $\sqrt{2}$ (e.g., section 3.8 of Szunyogh (2014)).

3.3 Comparisons of the Forecast Accuracy

3.3.1 Synopsis of the Forecast Verification Results

Figures 3 and 4 illustrate the temporal evolution of the forecast errors for the first five forecast days in the NH midlatitudes and Tropics, respectively. The errors are shown for the temperature (top row), meridional component of the wind vector (middle row) and specific humidity (bottom row) at forecast lead times day 1 (left column), day 3 (middle column), and day 5 (right column). In general, the hybrid forecasts (blue curves) have forecast value, except for the specific humidity at day 5 in the NH midlatitudes, for which they are only about as accurate as the forecasts based on climatology. In addition, the

hybrid forecasts are either more accurate than all benchmark forecasts, or similarly accurate to the most accurate benchmark forecast. The hybrid model performance in the SH midlatitudes (not shown) is similar to that in the NH midlatitudes. The advantage of the hybrid model compared to the different benchmarks, however, strongly depends on the forecast variable and lead time. Next, we discuss this dependence, as it provides important insight into the mechanisms by which CHyPP improves the numerical forecasts.

3.3.2 Hybrid Versus SPEEDY Forecasts

Compared to SPEEDY, the advantage of the hybrid model is the largest for the temperature. While all hybrid temperature forecasts have substantial forecast value for the first 5 forecast days, the SPEEDY day 5 temperature forecasts have no forecast value in the Tropics and in the stratosphere in the NH midlatitudes. In addition, the SPEEDY forecasts have little forecast value at day 5 in the midlatitudes. The benefit of the ML correction is particularly striking in the tropical upper troposphere, where the SPEEDY forecasts have a large error with a maximum of 6 K at 200 hPa, while the error of the hybrid forecasts remains below 1 K.

In addition to the temperature, the hybrid forecasts are also substantially more accurate than the SPEEDY forecasts for the specific humidity, especially, in the lower troposphere, where parameterizations play an important role in modeling the effects of moist atmospheric processes. While in the NH midlatitudes the hybrid forecasts degrade only to the level of the forecasts based on climatology by day 5, the error of the SPEEDY forecasts reaches saturation by that time.

In the two midlatitudes, the state variable for which the advantage of the hybrid model is the smallest compared to SPEEDY is the meridional component of the wind vector. This result is not surprising, as numerical models are known to capture synoptic-scale Rossby wave dynamics, which dominate the variability of weather in the midlatitudes. In contrast, in the Tropics, where wave dynamics is coupled to the parameterized process of deep convection, the advantage of the hybrid model for the meridional wind component is more substantial.

To explore the scale-dependence of the performance of the hybrid and benchmark forecasts, we examine the spectrum of the errors for the meridional component of the

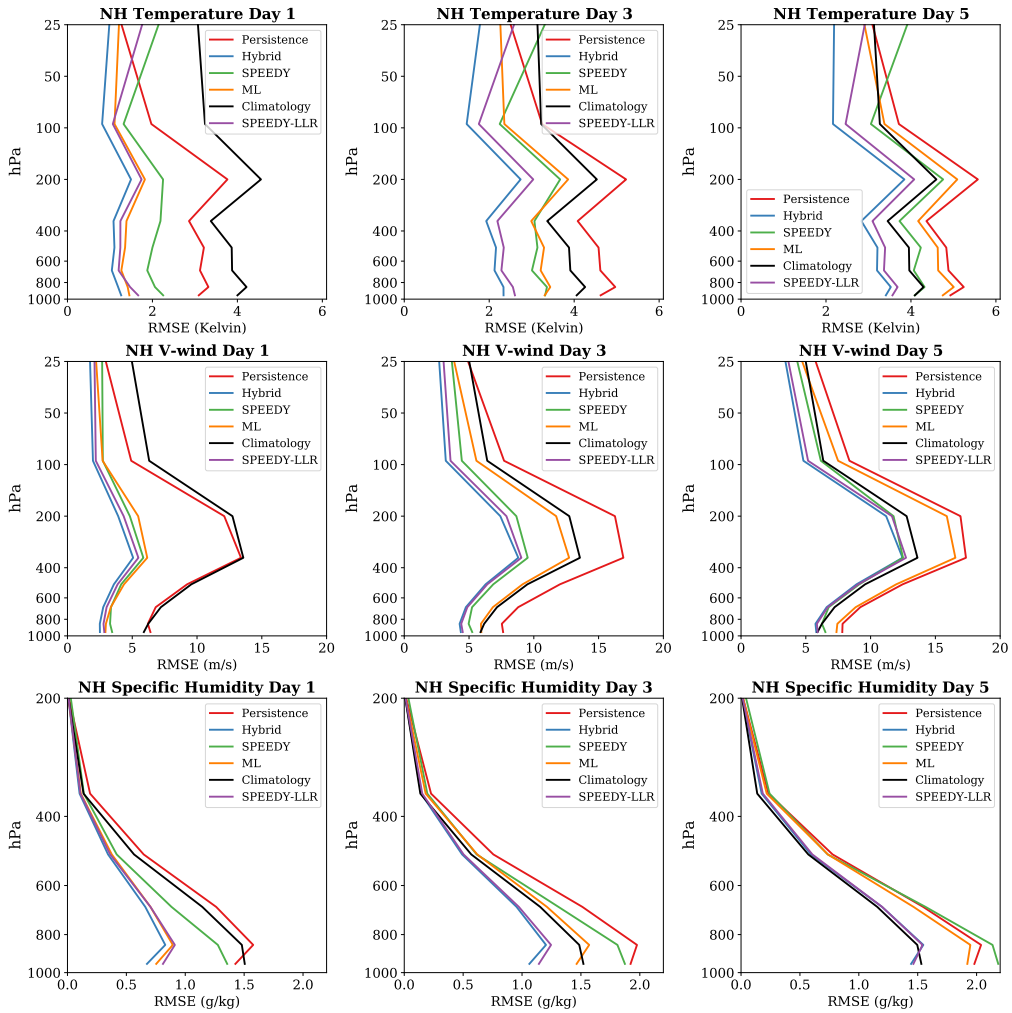

Figure 3. Northern Hemisphere midlatitudes (between 30°N and 70°N) forecast verification results. Results are shown for the (blue) hybrid model, (green) SPEEDY, (orange) ML-only model, (purple) SPEEDY-LLR model, (red) persistence, and (black) climatology. Shown is the area-weighted root-mean-square error at the different atmospheric levels for (top row) the temperature, (middle row) meridional wind, and (bottom row) specific humidity at (left column) day 1, (middle column) day 3, and (right column) day 5 forecast time.

wind at 500 hPa with respect to the zonal wave number (Figure 5). (This figure also shows results for day 10, in addition to the results for forecast days 1, 3, and 5.) The left panel shows the results for the hybrid and the SPEEDY model. Because SPEEDY is a spectral transform model with cut-off wave number 30, the spectrum for SPEEDY has no power at all beyond that wave number, and it is heavily dampened at wave numbers larger than about 20. Therefore, the errors of the hybrid forecasts, which have realistic power

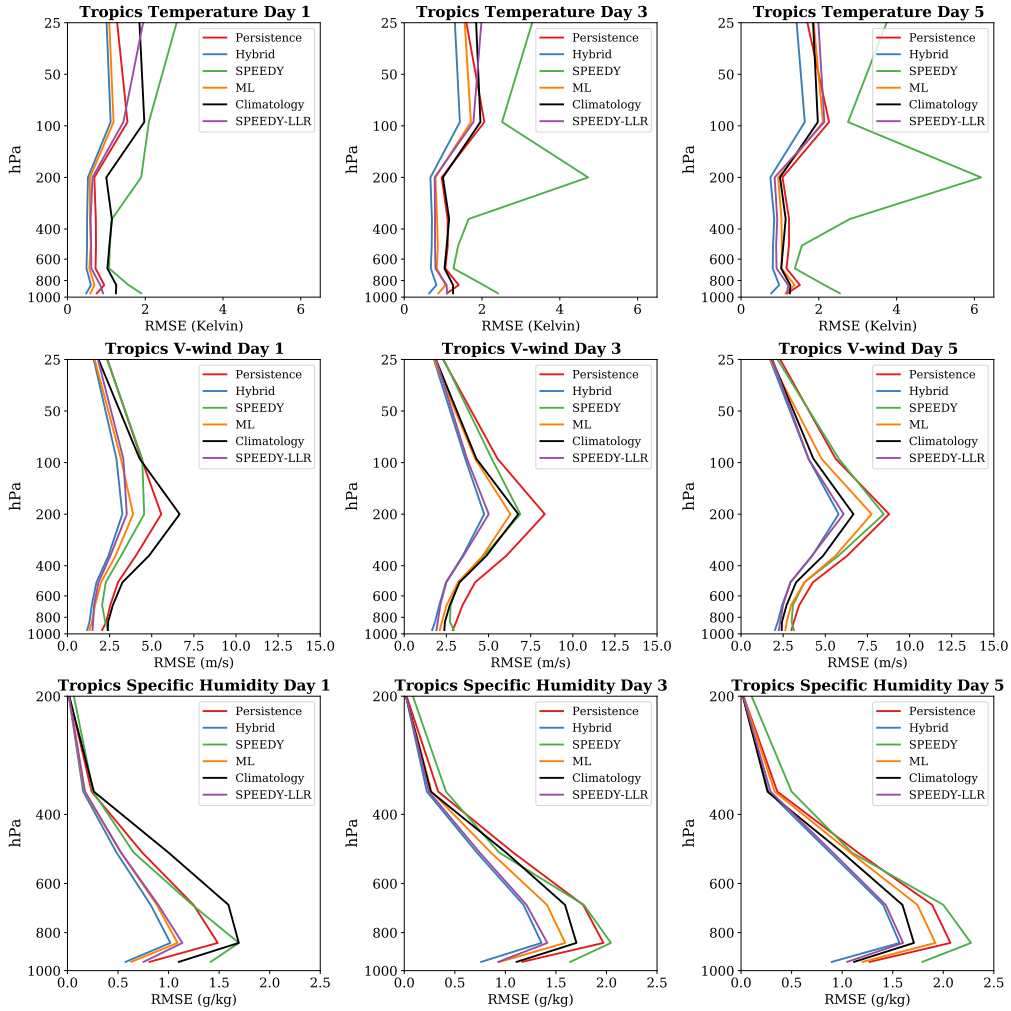

Figure 4. As in Fig. 3 for the Tropics (between 30°S and 30°N)

at all wave numbers, are expected to saturate at a level that is higher than that for SPEEDY at the tail-end of the spectrum. At day 1, the hybrid forecasts have a clear advantage over the SPEEDY forecasts at the synoptic and large scales (zonal wave numbers lower than about 20). A smaller, but spectrally similar advantage still exists at day 3, while the advantage of the hybrid forecasts disappears, except at wave numbers 5 and 6, by about day 5.

3.3.3 Hybrid Versus ML-only Forecasts

While the errors of the ML-only forecasts (orange curves in Figs. 3-5) are only slightly larger than that of the hybrid forecasts at day 1, they grow much faster in the next four

days and the ML forecasts typically have no value by day 3. This result suggests that while the RC-based ML technique can produce accurate forecasts in the short range (day 1-2), it is more effective in assisting SPEEDY than directly predicting the weather beyond that range. A comparison of the left and middle panels of Fig. 5 suggests that the information provided by SPEEDY to the hybrid is particularly beneficial at the large scales (wave numbers lower than about 6).

3.3.4 Hybrid Versus SPEEDY-LLR Forecasts

Next to the hybrid model, the benchmark that performs the best in the medium (day 2-5) forecast range is the SPEEDY-LLR (purple curves). While the hybrid forecasts are more accurate than the SPEEDY-LLR forecasts, the forecast error differences between the two models are modest, except for those in the stratosphere. The fact that the forecast error differences are smaller for the hybrid model versus SPEEDY-LLR than for the hybrid model versus SPEEDY indicates that the periodic interactive correction of the SPEEDY forecasts itself makes an important contribution to the good performance of the hybrid model. The additional forecast improvement, however, is not the only benefit of using ML rather than local linear regression for the forecast correction: while the hybrid forecasts remain stable indefinitely (see section 4), some of the SPEEDY-LLR forecasts fail as early as day 11 lead time, with about 60% of the forecasts reaching the intended 21 days.

It should be noted that the fact that local linear regression can efficiently correct the errors of a 6 h forecast is not completely surprising, considering that linear regression can be used to model the short-term forecast error dynamics for even a state-of-the-art NWP model (Bishop et al., 2017), in which nonlinear effects are expected to play a more important role even at short lead times. It is a nontrivial result, however, that the information provided by such a linear approach can be used for the periodic, interactive correction of an evolving numerical forecast. It is also a nontrivial result that an RC-based ML technique stabilizes the resulting hybrid model indefinitely, and leads to further forecast improvement in the short and medium (day 1-5) range.

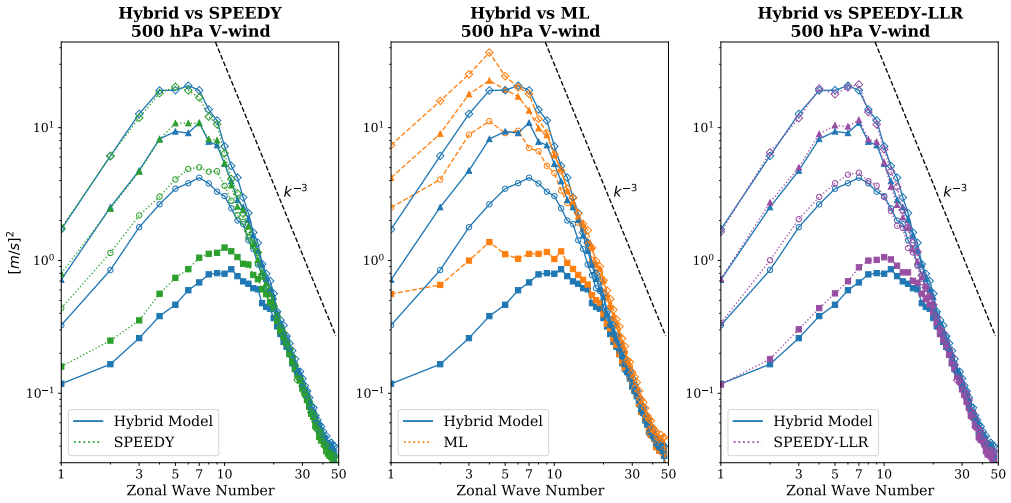

Figure 5. Spectral distribution of the 500 hPa meridional wind forecast error in the NH mid-latitudes (between 30°N and 70°N) with respect to the zonal wave number. The power spectra of the forecast errors are shown (left) for the the hybrid model (blue) vs SPEEDY (green), (middle) the hybrid model (blue) vs the ML-only model (orange), and (right) hybrid model (blue) vs SPEEDY-LLR (purple) at day 1 (solid square), day 3 (open circle), day 5 (solid triangle), and day 10 (open diamond).

3.4 Global Mean and Spatially Varying Errors

To gain further insight into the ways the hybrid approach improves forecast performance, we decompose the global RMSE into a bias and a standard deviation component. (The sum of the squares of the two components is equal to the square of the root-mean-square error.) The bias measures the global mean error, while the standard deviation measures the spatially varying part of the forecast error. The time evolution of the two error components, averaged over the 100 forecasts is shown for three representative state variables in Fig. 6.

For the temperature near the surface (at 950 hPa, top panel), SPEEDY rapidly develops a warm bias that oscillates around a mean of 0.75 K with the diurnal cycle. This bias is the result of SPEEDY using a single daily average value of the incoming solar radiation at the top of the atmosphere at all times of the day. The hybrid model greatly reduces the magnitude of the bias and also removes its diurnal oscillation. The biases of the ML model and SPEEDY-LLR are comparable to that of the hybrid model in magnitude, but the SPEEDY-LLR bias exhibits diurnal variability.

The spatially variable component of the low-level temperature error remains lower for the hybrid model than for SPEEDY throughout the 14-day period shown in the figure. The same component is initially similarly low for the hybrid and ML-only model, but it increases much more rapidly for the ML-only model. (Even with this rapid increase, the ML-only forecasts remain more accurate than the SPEEDY forecasts until about day 4). This component is initially lower for the hybrid model than for SPEEDY-LLR, but their accuracies are essentially the same after about day 8. Also, while the curves for SPEEDY and the hybrid model saturate at the same level as persistence, the curve for the ML-only model saturates at a higher level, indicating that the ML-only model overestimates the spatial variability of the low-level temperature at the longer forecast times.

SPEEDY rapidly develops a positive specific humidity bias near the surface (950 hPa, middle panel) that saturates at about 1 g/kg at day 7 lead time. Both the hybrid model and the other two benchmarks eliminate most of this bias. The spatially varying component of the error behaves similarly to that for the low level temperature, with the hybrid model outperforming the benchmarks for lead times from 1-7 days.

For the meridional wind component in the upper troposphere (200 hPa, bottom panel) none of the models develop a noteworthy bias. Thus, the differences in forecast performance are solely due to differences in the spatially varying component of the forecast error. This error component is still smaller for the hybrid model than SPEEDY for the first 9 forecast days, and than for the other benchmarks for the the first 6 forecast days.

3.5 Atmospheric Balance

Maintaining the delicate balance between the wind (momentum) and mass field in a numerical model, especially at short forecast lead times, has been one of the biggest challenges of atmospheric modeling since the dawn of NWP (e.g., Lynch, 2006). In a modern NWP model, a weakened balance is a short-lived transient property and the magnitude of the initial transient can be greatly reduced by *initialization* techniques (e.g., section 8 of Lynch (2006)). In the hybrid model and SPEEDY-LLR, however, no initialization is done before a corrected 6 h forecast is used as the initial condition of the next 6 h numerical forecast. Hence, the corrections inevitably upset the balance in the numerical component of the hybrid forecasts every 6 h. The forecast verification results dis-

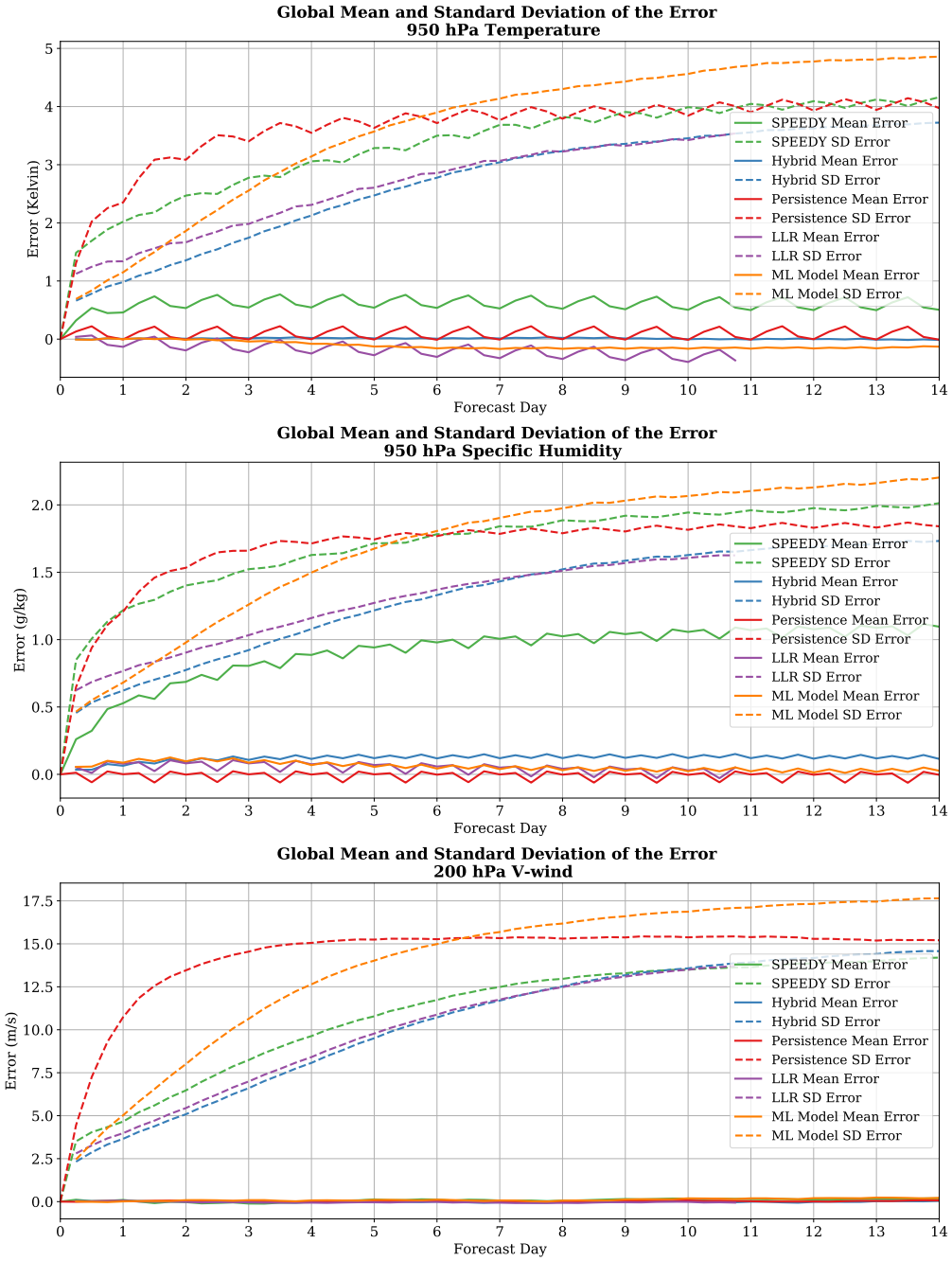

Figure 6. The time evolution of the (dashed) standard deviation and (solid) mean of the forecast errors. Each color indicates forecasts by a particular model: (blue) hybrid model, (green) SPEEDY, (purple) SPEEDY-LLR model, (orange) ML model, and (red) persistence. Results are not shown for SPEEDY-LLR beyond day 11, at which time some of the the forecasts for that model fail.

cussed thus far suggest that these imbalances do not outweigh the positive effects of the corrections on the accuracy of the hybrid forecasts. But, can the hybrid model produce realistic surface pressure tendencies by also correcting the surface pressure field for the effects of gravity waves excited by the imbalances? We investigate this possibility by examining the global root-mean-square of the surface pressure tendency in the forecasts for the hybrid and the benchmark models (Fig. 7). We assume that the value computed for ERA5 (red curve), which is about 0.4 hPa/h, provides a realistic estimate of the global root-mean-square of surface pressure tendency in the atmosphere.

As can be expected from a numerical model started from an uninitialized initial condition, the initial tendency for SPEEDY (about 1 hPa/h) is higher than desired. As forecast time increases, the the magnitude of the mean tendency drops, first rapidly, and then at a decreasing rate until it settles below the natural level, at about 0.28 hPa/h. The latter behavior suggests that the diffusion built into the model to combat imbalances over-smooths the temporal variability of the forecasts beyond day 1. While the magnitude of the mean tendency for the hybrid forecasts (about 0.38 hPa/h) is initially slightly smaller than the natural value, and further decreases in the first 72-84 h (to about 0.36 hPa/h), it is closer to the natural value than those for the benchmark forecasts. The SPEEDY-LLR is less effective than the hybrid model in eliminating the initial transient and it also produces an average tendency at the later forecast times (about 0.30 hPa/h) that is further below the natural level. The ML-only model behaves similarly to the hybrid model for the first two forecast days, but the saturation value is clearly lower (about 0.33 hPa/h) than for the hybrid model.

3.6 Sensitivity to Training Length

To test the sensitivity of the performance and stability of the hybrid model to the training length, we carry out a series of experiments with the same hyperparameters as before, but for shorter training periods. In particular, we train the model on 2 years, 5 years, or 10 years of reanalysis data, with the training always ending at 2300 UTC, June 26, 2011, as for the original forecast experiments. (We recall that the length of the training for the original experiments is 20.5 years.) The results of these experiments for the usual 100 21-day forecast cases for select variables are summarized in Fig. 8.

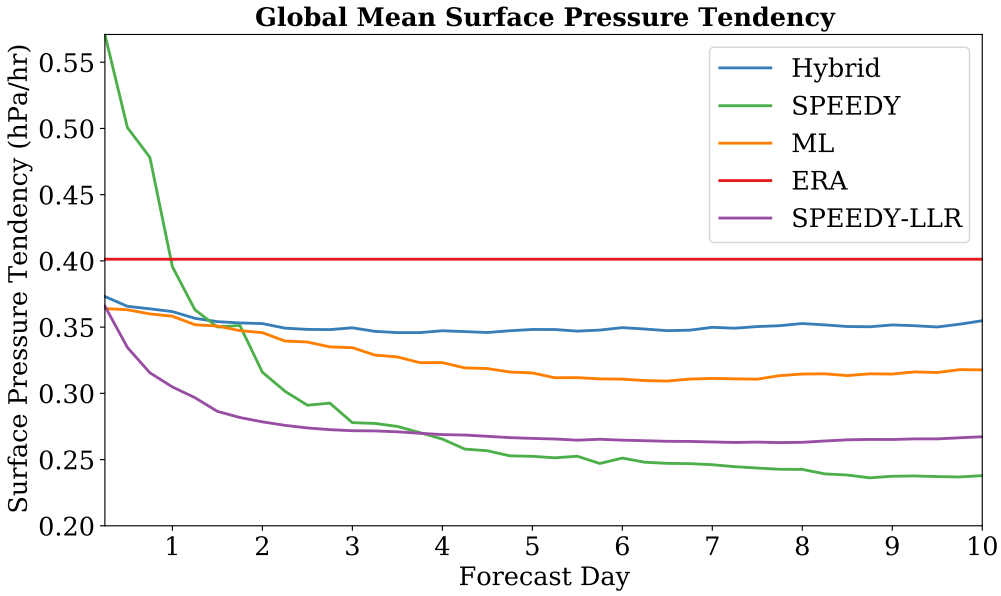

Figure 7. Atmospheric balance in the model forecasts. Shown is the global root-mean-square of the approximate surface pressure tendency computed by finite-differences based on 6-hourly data for the (blue) hybrid model, (green) SPEEDY, (orange) ML-only model, and (purple) SPEEDY-LLR model. The (red) value computed for 2011-2012 based on the ERA5 reanalyses is also shown for reference.

While training the hybrid model for only 2 years already significantly improves the forecast performance for the near-surface temperature and specific humidity compared to that of SPEEDY, extending the training length further improves the forecasts. The hybrid model trained for 2 years does not improve the meridional wind component in the upper troposphere, and actually degrades the forecasts beyond 3 days. A longer training makes the hybrid model perform better initially than SPEEDY. The length of the superior performance of the hybrid model becomes longer as the length of the training period increases. The results shown in Fig. 8 also suggest that a further modest improvements of the forecast performance could be achieved by using a training period even longer than 20.5 years.

4 Climate Simulation Experiment

To evaluate the long term stability of the hybrid model and its ability to simulate the climate, we compute an 11 year long free run with the model. For this simulation

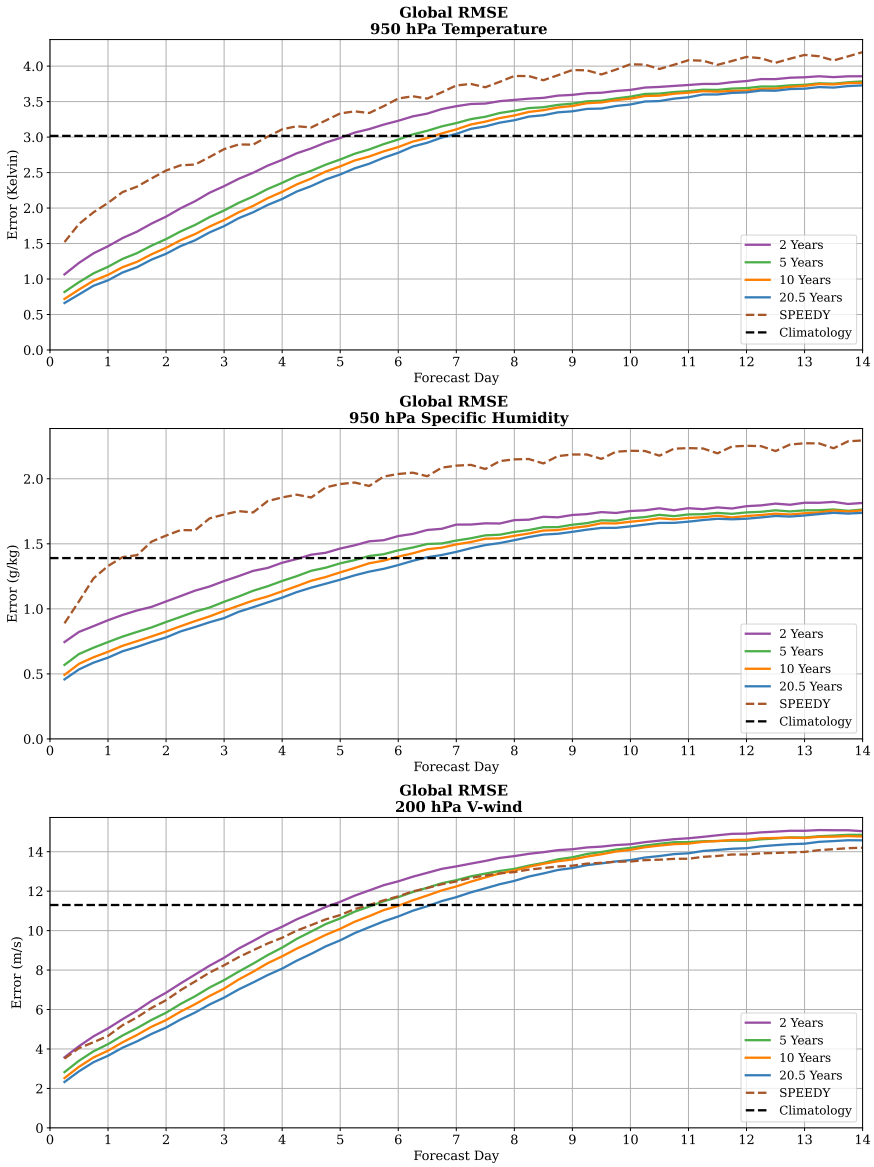

Figure 8. Time evolution of the global root-mean-square forecast error for different lengths of the training of the hybrid model. Results are shown for a (purple) 2 years, (green) 5 years, (red) 10 years, and (blue) 20.5 years training period. For reference, the forecast errors are also shown for (brown dashes) SPEEDY and (black dashes) climatology.

experiment, the hybrid model is trained on ERA5 reanalyses for the 19-year period from January 1, 1981 to December 27, 1999. The simulation starts from the ERA5 reanalysis valid at 0000 UTC, January 1, 2000. To suppress the effects of initial transients and the initial condition on the model diagnostics, we discard the data from the first year of the simulations before computing the diagnostics. To compare the performance of the

hybrid model and SPEEDY in simulating the climate, we assume that the two simulations attempt to simulate the climate of the 10-year period from 2001-2010 as represented by ERA5.

4.1 Zonal Mean Biases

Figures 9 and 10 show the zonal mean biases of the simulations by SPEEDY (left panels) and the hybrid (right panels) for the boreal winter (December, January, and February) and boreal summer (June, July, and August), respectively. These figures can be used, not only to compare the quality of the two simulations, but also to assess the average magnitude of the corrections made by the ML component of the hybrid model. In particular, the difference between a left panel and the corresponding right panel is the zonal mean of the ML correction for a particular state variable.

The top left panels show that SPEEDY has a large upper tropospheric warm bias for the tropical regions, during both the boreal winter and summer. In both polar regions SPEEDY has a cold bias for the upper troposphere and stratosphere during the boreal winter and a warm (cold) bias in the southern (northern) polar region during the boreal summer. The magnitude of the bias is not surprising given the coarse resolution and simplified parameterizations used in SPEEDY (Molteni, 2003). The top right panels show that the hybrid model greatly reduces, but does not completely eliminate, these biases when the model is cycled over a long period of time. The bias reduction is particularly notable in the the tropics and the midlatitudes. The largest remaining biases are in the polar regions.

The hybrid model reduces the zonal component of the wind bias, especially in the stratosphere and upper troposphere, and in the lower troposphere in the SH midlatitudes in the boreal summer. The only exception is the introduction of a positive zonal component of the wind bias in the stratosphere in the tropics. The hybrid model also greatly reduces the large positive humidity bias of SPEEDY with maxima in the tropics.

Figure 11 shows the mean surface pressure biases for the simulations by SPEEDY (left panels) and hybrid model (right panels) for the boreal winter (top row) and boreal summer (bottom row). The mottled short scale patterning seen in the two left panels of the figure are due to the spectrally truncated topography of SPEEDY, which is much smoother than the topography determining the interpolated ERA5 reanalyses used for

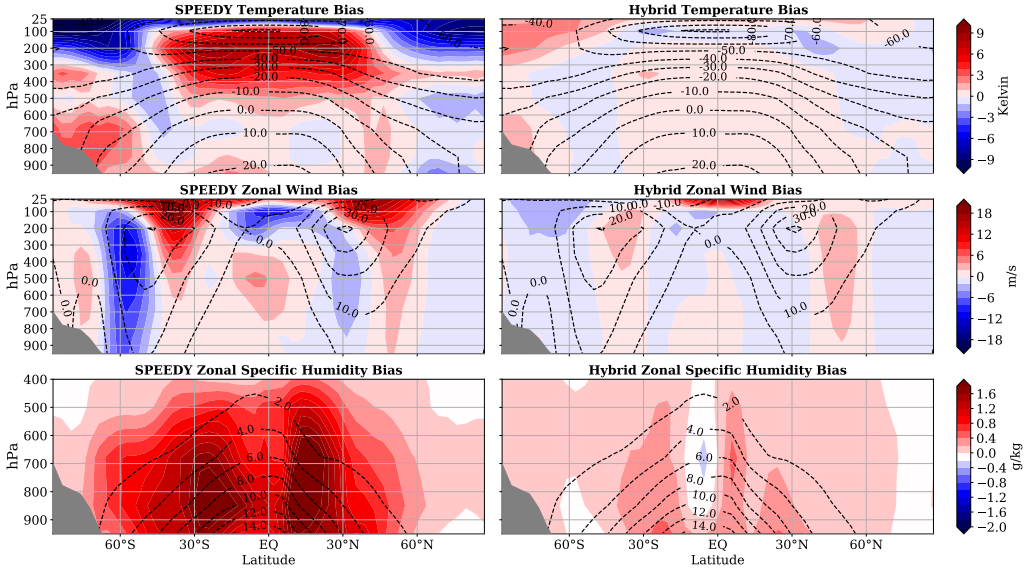

Figure 9. Comparison of the zonal mean biases of the SPEEDY and hybrid simulation simulations for the boreal winter (December, January, February). Results are shown for (top) the temperature (middle) zonal wind, and (bottom) specific humidity for (left) SPEEDY and (right) the hybrid model.

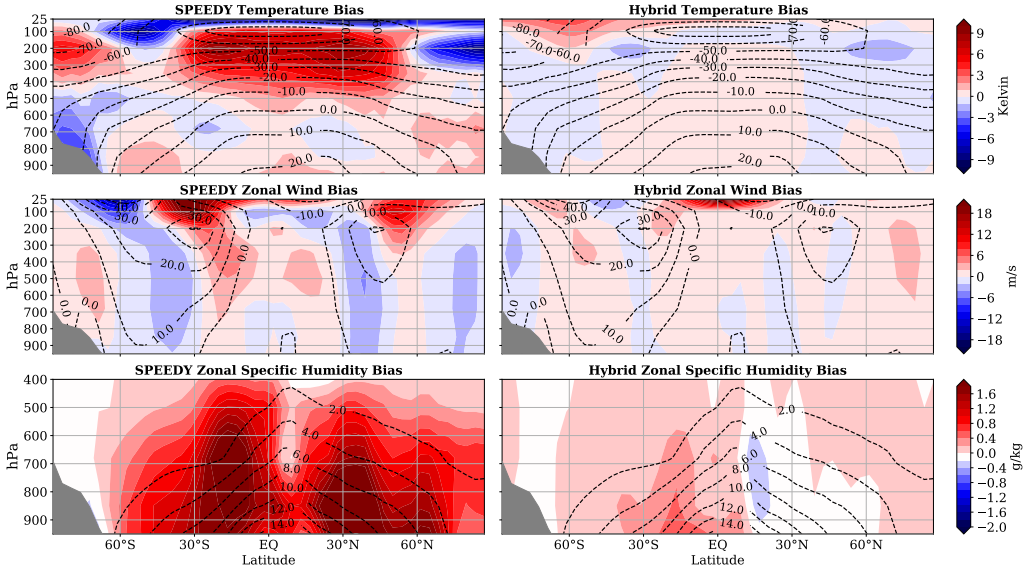

Figure 10. Same as Fig. 9, except for the boreal summer (June, July, August).

the evaluation of the simulations, and for the training of the hybrid model. In combination with the artifacts caused by the spectral truncation in SPEEDY, the large local differences in the mountainous regions lead to substantial surface pressure biases in the

SPEEDY simulations. The hybrid model corrects the large local biases, but still has smaller magnitude large scale biases. The wave-number-two structure of the large-scale hybrid model bias in the NH suggests that these biases are related to the low resolution representation of the topography and the land-sea contrasts in the numerical model. The remaining biases are also relatively large in the polar regions, especially in the boreal summer. We speculate that the bias of the hybrid model in the polar regions might be related to our particular strategy to do the localization on a cylindric (Mercator) map projection. On the other hand, the bias is not concentrated at the poles for the variables shown in Figures 9 and 10.

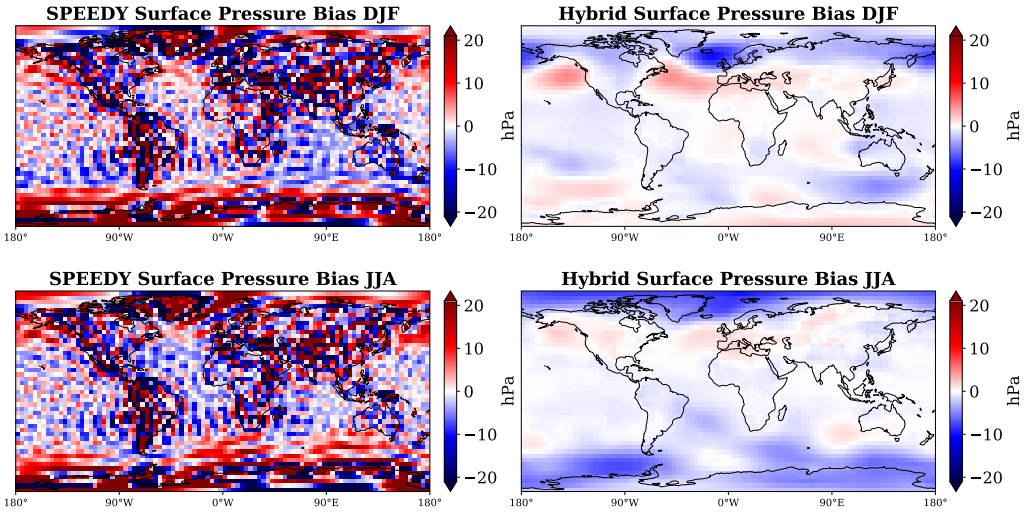

Figure 11. The mean surface pressure bias in the SPEEDY and hybrid climate simulations. Shown is the bias for (top) the boreal winter (December, Januar, February) and (bottom) boreal summer (June, July, August) for (left) SPEEDY and (right) the hybrid model.

4.2 Temporal variability

To investigate the temporal variability of the atmosphere in the SPEEDY and hybrid climate simulations, we examine the temporal dependence of the 950 hPa temperature at the four model grid points that fall in the Sahara Desert. The top two panels of Fig. 12 show the power spectra of the temporal variability for the two models. These power spectra are computed by applying a Hamming filter first, and then a discrete Fourier transform to the 10 years of 6-hourly simulation data, and finally computing the square

of the absolute value of the Fourier coefficients. The results show that both simulations correctly capture the variability at time scales longer than about a week. At the shorter time scales, however, SPEEDY increasingly underestimates the variability. The ML correction greatly reduces, but does not completely eliminate, this problem: the hybrid model underestimates the variability at the scales between one week and one day only slightly, and reduces the underestimation by SPEEDY at the even shorter scales. Most importantly, unlike SPEEDY, the hybrid model has a strong diurnal cycle. It should be noted that an earlier version of the hybrid model, which did not include the incoming solar radiation at the top of the atmosphere as an input to the reservoir, lost the diurnal cycle at around the end of year 4. This motivated us to add the incoming solar radiation as an input parameter, even though it had no significant effect on the forecast accuracy. We find it a noteworthy, nontrivial result that the earlier version of the hybrid model was able to learn the diurnal cycle strictly from the training data.

The fact that a simulation correctly captures the variability at a number of frequencies does not guarantee that the phases of the temporal changes (e.g. the timing of the seasons) are also correct. To exclude the possibility of such a flaw of the simulations, we plot (bottom panel of Fig. 12) the time series of the average 950 hPa temperature for the same four Saharan grid points for the last full year of the simulations. The points along these curves should fall within two standard deviations from the mean for the given date and time (the interval marked by gray shading) with a 95% observed frequency. Based on the full ten years of data, the observed frequency is 88.2% for SPEEDY and 98.0% for the hybrid model.

5 Conclusions

In this paper, we described results from the first implementation of the hybrid modeling approach CHyPP of Wikner et al. (2020) on a realistic atmospheric model. We used a low-resolution AGCM based on the full set of primitive equations, along with ERA5 reanalysis data for training and verification, to demonstrate the potentials of CHyPP for both NWP and climate modeling. The spatio-temporal structure of the improvements of the forecasts and simulations suggests that the ML component of the model primarily corrects for errors caused by the limitations of the parameterization schemes of the AGCM. While state-of-the-art numerical models have much higher resolutions and more advanced parameterization schemes than SPEEDY, the weather forecasts and climate

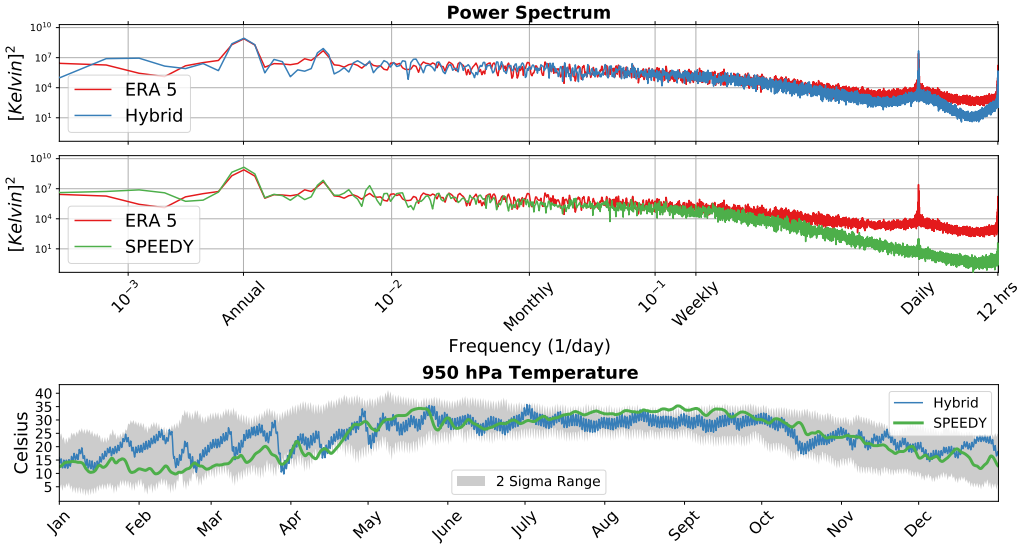

Figure 12. Temporal variability of the 950 hPa temperature in the Sahara Desert for the ten years of simulations. Shown are the power spectra for (top) the hybrid model and ERA5 and (middle) SPEEDY and ERA5. The bottom panel shows the time series of simulated temperatures for the last full year of the simulations. The gray shading represents the range of plus/minus two standard deviations from the mean in the ERA5 reanalyses for 2001-2010.

simulations they provide still have substantial biases. We expect the hybrid approach to effectively reduce these biases.

Because the ML component of the hybrid model is based on RC, training the model is computationally highly efficient. Specifically, the training described in this paper requires only 30 minutes wall-clock time using 1,152 Intel Xeon E5-2670 v2 processors on a supercomputer that is much less powerful than those at the operational NWP centers. Using the same computational resources, preparing a 21-day forecast takes about 52 seconds, while carrying out a one-year simulation takes about 15 minutes. These numbers are only 25% higher than those for SPEEDY, and the extra time is mainly due to the overhead associated with the frequent restart of SPEEDY.

Due to the parallel nature of the computational algorithm, we expect it to scale well for higher model resolutions and larger number of processors. A modification of the current implementation of our method that might be helpful for scaling is vertical localization. By "vertical localization" we mean the use of local domains that, as well as being limited in horizontal extent as shown in Fig. 1, are also of limited height and are

stacked vertically with overlap from ground-level to the top of the atmosphere. Though we do not use vertical localization in this article, we plan to test it soon for potential improvements with SPEEDY.

The ideal size of a local domain still needs to be determined through additional experimentation, both for SPEEDY and for higher-resolution models. Thus, it is hard to make a precise quantitative projection for scaling, but here is a comparison that indicates feasibility for operational models. The current computer of ECMWF has 129,960 processors (about 100 times more than what we used), and their operational model has 6.5×10^6 horizontal grid points (about 180 times more than SPEEDY) ("IFS Documentation CY47R1 - Part III: Dynamics and Numerical Procedures", 2020). If the local regions for the ECMWF model would be defined by four horizontal and all vertical grid points, as in our paper, each processor would have to handle less than twice as many local regions at ECMWF than in our model. Also, there is no obvious reason to believe that the computational overhead of the hybrid model would be substantially higher than the 25% we found for SPEEDY. The high computational efficiency of the approach would allow for a large number of experiments to find the optimal configuration of a future operational hybrid model. Developing an efficient systematic approach to find a near optimal combination of the hyperparameters, nevertheless, would be highly desirable and is one of the subjects of our ongoing research efforts. =An unknown factor that could have a very favorable impact on future scaling considerations is the ongoing rapid technological developments of alternative, fast, cheap physical implementations of reservoir computing, e.g., implementations based on photonics or on Field Programmable Gate Arrays.

We emphasize that while the ML component of the hybrid model is highly efficient in correcting the biases of the forecasts and simulations prepared by the host model, it is not a ML-based postprocessing technique. While a technique of the latter type corrects the numerical-model-based forecasts of a specific forecast variable or phenomenon (e.g., Rasp & Lerch, 2018; Chapman et al., 2019; Kim et al., 2021) without interacting with the numerical model, the ML component of the hybrid model makes frequent periodic interactive corrections to the numerical model solution. Hence, it also greatly improves the representation of the spatiotemporal variability of the atmospheric state by the model.

We expect that the performance of the hybrid model can be further improved by investigating the relationship between the parameters of the ML model and the representation of basic atmospheric processes. Such an investigation could lead to further improvements of the model, similar to the way studies of the interactions between numerics and dynamics (e.g., Arakawa & Lamb, 1977) led to much improved physic-based numerical models. For instance, one potentially important fundamental question is the optimal relationship between the size of the local domains, the overlap between the local domains in the input of the reservoir, and the length of the time step Δt. The fact that the ML component is more effective in correcting localized errors than errors at the larger scales in the current version of our hybrid model may be partly the result of using local domains and an overlap that are less than optimal for the selected time step. In our experiments, the size of the overlap was primarily dictated by the structure of our code and the available computer resources, but larger local domains and a larger overlap could be used in the future.

An intriguing possibility is to use the hybrid model for data assimilation in addition to forecasting, as data assimilation could greatly benefit from the higher accuracy and smaller biases of the short term hybrid forecasts used as background. Furthermore, integrating ML and data assimilation may allow in the future to do online training of the ML component of the hybrid model on real-time observations rather than canned reanalyses data. The availability of such training procedure would make it possible to extend the hybrid modeling approach to numerical models for which high-quality reanalysis data are not available (e.g., an AGCM that also includes a sophisticated model of the upper atmosphere well beyond the lower stratosphere). It could also allow the ML component of the model to adjust to variability and changes of the climate. We have made a first step toward this ambitious goal, in which we iteratively use the hybrid model to prepare an updated set of analyses, which is then used to train the next iteration of the hybrid model (Wikner et al., 2021). Our plan is to test this approach with the hybrid model of the current paper.

Acknowledgments

This work was supported by DARPA contract DARPA-PA-18-01 (HR111890044). The work of T. A. and I. S. was also supported by ONR award N00014-18-2509. The work of Alexander Wikner was supported in part by the National Science Foundation (NSF)

(Award No. DGE-1632976). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. This paper greatly benefitted from stimulating discussions with Sarthak Chandra, Michelle Girvan, Garrett Katz, and Andrew Pomerance. The constructive comments of the three anonymous reviewers helped us to greatly improve the presentation of our ideas and results. The new data generated for the paper are available online (http://doi.org/10.5281/zenodo.5103176).

References

Arakawa, A., & Lamb, V. R. (1977). Computational design of the basic dynamical processes of the UCLA general circulation model. In J. Chang (Ed.), *Methods in computational physics: Advances in research and applications* (Vol. 17, p. 173-265). Elsevier.

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020). A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, *47*, e2020GL087776.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*, 47-55.

Bishop, C. H., Frolov, S., Allen, D. R., Kuhl, D. D., & Hoppel, K. (2017). The local ensemble tangent linear model: an enabler for coupled model 4d-var. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 1009-1020. doi: https://doi.org/10.1002/qj.2986

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*, 6289-6298.

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*, 2728-2744.

Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, *46*(17-18), 10627-10635. doi: https://doi.org/10.1029/2019GL083662

Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2020). Data-driven super-parameterization using deep learning: Experimentation with multiscale lorenz

96 systems and transfer learning. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002084. doi: https://doi.org/10.1029/2020MS002084

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3067-3084. doi: https://doi.org/10.1002/qj.4116

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.

Gilbert, E. N. (1959). Random graphs. *Ann. Math. Statist.*, *30*, 1141-1144.

Harper, K. C. (2008). *Weather by the numbers the genesis of modern meteorology.* The MIT Press. doi: 10.2307/j.ctt5hhddq

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. doi: https://doi.org/10.1002/qj.3803

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, *230*(1), 112-126. doi: https://doi.org/10.1016/j.physd.2006.11.008

IFS documentation cy47r1 - part III: Dynamics and numerical procedures. (2020). In *IFS documentation cy47r1.* ECMWF.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, *148*.

Kim, H., Ham, Y. G., Joo, Y. S., & Son, S. W. (2021). Deep learning for bias correction of mjo prediction. *Nature Communications*, *12*(1), 3087. doi: 10.1038/s41467-021-23406-3

Krasnopolsky, V., & Fox-Rabinovitz, M. S. (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, *19*, 122-134.

Krasnopolsky, V., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2010). Development of neural network convection parameterizations for numerical climate and

weather prediction models using cloud resolving model simulations. In *The 2010 International Joint Conference on Neural Networks (IJCNN)* (p. 1-8).

Krasnopolsky, V., 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.

Krasnopolsky, V. M. (2013). Applications of NNs to developing hybrid earth system numerical models for climate and weather. In *The application of neural networks in the earth system sciences: Neural networks emulations for complex multidimensional mappings* (p. 81-143). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-6073-8_4

Kucharski, F., Molteni, F., & Bracco, A. (2006). Decadal interactions between the western tropical pacific and the north atlantic oscillation. *Climate Dynamics*, *26*(1), 79-91.

Lukoševičius, M. (2012). A practical guide to applying echo state networks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade: Second edition* (p. 659-686). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127-149.

Lynch, P. (2006). *The emergence of numerical weather prediction: Richardson's dream*. Cambridge University Press.

Molteni, F. (2003). Atmospheric simulations using a GCM with simplified physical parametrizations. I: model climatology and variability in multi-decadal experiments. *Climate Dynamics*, *20*(2), 175-191.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., ... Yorke, J. A. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, *56*(A), 415-428.

Patel, D., Canaday, D., Girvan, M., Pomerance, A., & Ott, E. (2021). Using machine learning to predict statistical properties of non-stationary dynamical processes: System climate, regime transitions, and the effect of stochasticity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*(3), 033149. doi: 10.1063/5.0042598

Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, *120*(2), 024102.

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), 041101.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885-3900.

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*, 9684-9689.

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405.

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*(24), 12,396-12,417.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., & Potthast, R. (2016). Kilometre-scale ensemble data assimilation for the cosmo model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1453-1472.

Sivashinsky, G. I. (1977). Nonlinear analysis of hydrodynamic instability in laminar flames—i. derivation of basic equations. *Acta Astronautica*, *4*(11), 1177-1206. doi: https://doi.org/10.1016/0094-5765(77)90096-0

Stensrud, D. J. (2007). *Parameterization schemes: Keys to understanding numerical weather prediction models*. Cambridge, UK: Cambridge University Press.

Szunyogh, I. (2014). *Applicable atmospheric dynamics: Techniques for the exploration of atmospheric dynamics*. doi: 10.1142/8047

Szunyogh, I., Kostelich, E. J., Gyarmati, G., Kalnay, E., Hunt, B. R., Ott, E., ... Yorke, J. A. (2008). A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus*, *60*(A), 113-130.

Tikhonov, A. N., & Arsenin, V. I. (1977). Solutions of ill-posed problems.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon,

J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: https://doi.org/10.1029/2021GL092555

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002109.

Wikner, A., Pathak, J., Hunt, B., Girvan, M., Arcomano, T., Szunyogh, I., ... Ott, E. (2020). Combining machine learning with knowledge-based modeling for scalable forecasting and subgrid-scale closure of large, complex, spatiotemporal systems. *Chaos*, *30*(5), 053111.

Wikner, A., Pathak, J., Hunt, B. R., Szunyogh, I., Girvan, M., & Ott, E. (2021). Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*(5), 053114.