Improving Seasonal Forecast Using Probabilistic Deep Learning

Baoxiang Pan1, Gemma J. Anderson1, André Goncalves1, Donald D. Lucas1, Céline J. W. Bonfils1, and Jiwoo Lee1

1Lawrence Livermore National Laboratory, Livermore, CA, USA

Abstract The path toward realizing the potential of seasonal forecasting and its socioeconomic benefits relies on improving general circulation model (GCM) based dynamical forecast systems. To improve dynamical seasonal forecasts, it is crucial to set up forecast benchmarks, and clarify forecast limitations posed by model initialization errors, formulation deficiencies, and internal climate variability. With huge costs in generating large forecast ensembles, and limited observations for forecast verification, the seasonal forecast benchmarking and diagnosing task proves challenging. Here, we develop a probabilistic deep learning-based statistical forecast methodology, drawing on a wealth of climate simulations to enhance seasonal forecast capability and forecast diagnosis. By explicitly modeling the internal climate variability and GCM formulation differences, the proposed Conditional Generative Forecasting (CGF) methodology enables bypassing crucial barriers in dynamical forecast, and offers a top-down viewpoint to examine how complicated GCMs encode the seasonal predictability information. We apply the CGF methodology for global seasonal forecast of precipitation and 2 m air temperature, based on a unique data set consisting 52,201 years of climate simulation. Results show that the CGF methodology can faithfully represent the seasonal predictability information encoded in GCMs. We successfully apply this learned relationship in real-world seasonal forecast, achieving competitive performance compared to dynamical forecasts. Using this CGF as benchmark, we reveal the impact of insufficient forecast spread sampling that limits the skill of the considered dynamical forecast system. Finally, we introduce different strategies for composing ensembles using the CGF methodology, highlighting the potential for leveraging the strengths of multiple GCMs to achieve advantageous seasonal forecast.

Plain Language Summary Seasonal forecast benefits a broad range of societal sectors. However, current dynamical seasonal forecast systems are considerably hindered by observation, model, and computation limitations. We develop a machine learning probabilistic forecast model that learns from climate simulations to infer possible climate patterns a season ahead. The model achieves competitive performance for global precipitation and temperature forecast, compared to the costly dynamical forecasts. More importantly, it offers crucial implications for understanding the limitations and improving the current dynamical forecast systems.

1. Introduction

General circulation models (GCMs) that simulate the dynamics and interactions of various climate subsystems are the workhorses for predictions on weather to climate scales. Over the past decades, advances in computation, observation, and modeling have gradually extended the range of useful deterministic forecasts toward the predictability limit. This limit of range, set by the chaotic nature of the atmosphere, is estimated to be around two weeks (F. Zhang et al., 2019). Forecasts beyond this range are intrinsically probabilistic (Lorenz, 1969). These forecasts seek prediction sources from low-frequency climate signals, which give rise to predictability by constraining the climate variability through the forecasting period.

By better characterizing the key prediction sources and their impacts, GCM-based dynamical seasonal forecasts have demonstrated markedly improved skill, offering crucial benefits to a wide range of societal sectors (Palmer, 2002; Pan, 2019; Zhao et al., 2021). Despite the achievements, current dynamical seasonal forecast systems face challenges in (a) realistically initializing the forecasting models; (b) faithfully representing the climate dynamics that influence seasonal variability; and (c) sufficiently sampling the forecast spread resulting from a chaotic evolution of initial state estimation uncertainties. Given the huge cost to initialize and run large-ensemble dynamical forecasts, the growing complexity of the forecasting GCMs, and the finite climate observations...
for forecast verification, progresses toward tackling the aforementioned challenges and improving dynamical seasonal forecasts become increasingly burdensome.

Aside from the dynamical approach, statistical models have long been applied for seasonal forecasts, offering useful forecast benchmarks and crucial guidance for improving dynamical forecast systems. Rather than explicitly resolving the climate dynamics, these models simulate the statistical dependency relationship between the considered predictand and its prediction sources, based on the available data samples, mostly from observations. Supported by limited observational data, statistical forecasts are often prone to spurious correlations (Caldwell et al., 2014; Gibson et al., 2019), and poorly represent multi-factor impacts. To tackle these limitations, recent studies have begun to build statistical seasonal forecast models using the big data of climate simulations (Stone et al., 2000; Ding et al., 2019, 2018; Ham et al., 2019; Y. Wang et al., 2020).

Why is a statistical forecast model built on climate simulation data applicable for practical seasonal forecast, given that the simulated climate can not account for the correct phase of natural climate variability? This is because, GCMs can roughly well represent the probabilistic dependency between the considered predictand and its prediction source (Ding et al., 2019; Ham et al., 2019). Therefore, we could apply this “model world” revealed seasonal co-variability relationship for real world seasonal forecast. To illustrate this idea, we may consider fitting a statistical emulator in a climate simulation “model world” to infer the distribution of seasonal average precipitation (the predictand), conditioned on the previous season’s upper ocean thermal state (the predictor). If the probabilistic dependency relationship is realistically captured by the GCM, and faithfully reproduced by the statistical emulator, we can apply the learned relationship for real-world forecast.

Building statistical forecast models using climate simulation data offers several overlooked but crucial benefits, as these statistical forecast models allow us to bypass some of the aforementioned dynamical forecasting barriers, and provide a top-down viewpoint to examine how a complicated GCM encodes the seasonal predictability information. Specifically, we can largely avoid the impact of state shocks in initializing an ocean-atmosphere coupled GCM, by directly modeling the seasonal co-variability relationship revealed in a non-interrupted simulation of this GCM. We can also flexibly leverage the strengths of multiple GCMs by learning from their ensemble simulations. Finally, supported by potentially unlimited climate simulation data, we can now apply powerful statistical models, that is, deep neural networks, to explicitly model the interplay between the seasonal predictability signal and the internal climate variability noise, offering an efficient complement to the costly large ensemble forecast approach for understanding seasonal predictability.

To realize the benefits above, we should clearly distinguish and represent different sources of uncertainties in formulating a statistical forecast model. Existing works along this novel line of research usually apply “off-the-shelf” machine learning regression techniques for deterministic seasonal forecast (Stone et al., 2000; Ham et al., 2019; Y. Wang et al., 2020). However, in these problem settings, the statistical forecast models focus on extracting the predictability signal from the reference climate simulation data, and they ignore the clarification or representation of forecasting uncertainties resulting from internal climate variability and GCM formulation deficiencies (Ding et al., 2018; Ham et al., 2019). Without a clear distinction of different forecasting uncertainties, we can not make utmost use of the data, nor can we gain useful insights about the forecasting barriers imposed by these uncertainties (Anderson & Lucas, 2018; Pan, Anderson, Goncalves, et al., 2020; Pan et al., 2021).

Our objective in this work is to develop a data-driven probabilistic modeling framework that leverages the big data of climate simulations for probabilistic seasonal forecasting. We explicitly represent the impact of internal climate variability and GCM formulation deficiencies to estimate the forecasting uncertainty. By doing so, we build a clear connection between the top-down, data-driven forecasting methodology and the bottom-up, dynamical forecasting methodology. A comparison between these two methodologies helps reveal different aspects of the forecasting barriers, pinpointing directions toward seasonal forecast improvement.

To achieve this objective, we apply probability theory to rigorously represent the different sources of uncertainty in designing our data-driven seasonal forecasting model. This practice goes beyond the traditional pattern recognition problem setting, and poses a number of computational and modeling challenges (Ghahramani, 2015). Recently, approximate probabilistic inference powered by deep neural networks has enabled probabilistic modeling on big data (Kingma & Welling, 2013). Existing works that leverage this idea have demonstrated imminent potential for various probabilistic modeling and forecasting tasks, such as video prediction (Walker et al., 2016), wind turbine fatigue prediction (Mylonas et al., 2021), and stochastic parameterization of unresolved processes in
climate models (Mukkavilli et al., 2020). We make use of this novel technique to learn from climate simulation for probabilistic seasonal forecasting.

The rest of this paper is organized as follows. We formulate the problem and introduce the data-driven probabilistic modeling framework in Section 2. Section 3 describes the experimental setup, including the data, model configuration, and evaluation strategies. In Section 4, we present the seasonal forecasting results by comparing our methodology with dynamical forecasts. Based on these results, we discuss the implications for understanding and improving current seasonal forecasting systems. We draw conclusions in Section 5.

2. Methodology

This section introduces the data-driven probabilistic forecasting methodology, which hereafter we refer to as the Conditional Generative Forecasting (CGF) model. As is illustrated in Figure 1, underpinning the CGF model is a novel deep learning approach that incorporates stochastic gradient variational inference (Kingma & Welling, 2013) for probabilistic representation of the internal climate variability noise (Figure 1 middle, Section 2.2), and the entity embedding technique (Guo & Berkhahn, 2016) that accounts for GCM formulation differences (Figure 1 top, Section 2.3). Within this framework, we use convolutional neural nets (CNNs, LeCun & Bengio, 1995) to learn feature representations of the geospatial data (Figure 1 left, right, and bottom, Section 2.4). While this methodology is detailed below in a general manner to fit for a broad range of applications, we consider
a case study of global seasonal forecasting for precipitation and near-surface (2 m) air temperature to demonstrate its capability in Section 3.

2.1. Notation and Problem Formulation

We consider climate simulation data that appear as tuples \((X, Y, M) = \{(x_1, y_1, m_1), \ldots, (x_N, y_N, m_N)\}\). Here, \(X\) and \(Y\) are array-structured data representing the predictor and predictand variables, \(M\) is categorical data that indexes the data generating GCMs, \(N\) denotes sample size. In our case, \(X\) are low-frequency climate variables that strongly regulate the seasonal variability of \(Y\). For instance, \(X\) could be the upper ocean thermal state, \(Y\) could be following season's precipitation or temperature anomaly. Our objective is to approximate the conditional distribution of \(p(Y|X, M)\), using multiple tuples of training data generated by different GCMs.

2.2. Stochastic Gradient Variational Inference

How to efficiently and accurately approximate \(p(Y|X, M)\), in the presence of high dimensional geospatial data, data inhomogeneity introduced by GCM formulation differences, and large data set? We define a random variable \(z\) with fixed distribution, that is, standard Gaussian, and pass it, together with the predictor information \(X\) and GCM index information \(M\), through a parametric function \(p_\theta\). By varying \(\theta\), we can change this distribution and make it close to the target distribution of \(p(Y|X, M)\) (Arjovsky et al., 2017). Once the model is trained, we can make a probabilistic forecast by simply generating samples of \(z\), and passing them together with \(X\) and \(M\) through the trained model to create plausible samples of \(Y\).

Given the data, a theoretically suitable objective for training \(p_\theta\) is to maximize the sum over the log likelihood of individual data points: \(\sum_{i=1}^{N} \log p_\theta (y_i|x_i, m_i)\). Applying the law of total probability, each log likelihood here can be written as:

\[
\log p_\theta (y_i|x_i, m_i) = \log \int p(y_i|x_i, m_i, z) \, p(z|x_i, m_i) \, dz
\]

(1)

to maximize the log likelihood based on Equation 1, we need to integrate over \(z\) for each data point, which is computationally intractable. To tackle this computational difficulty, we resort to stochastic gradient variational inference (Kingma & Welling, 2013): for each data point \((x_i, y_i, m_i)\), we narrow the integration space of \(z\) in Equation 1, by only considering \(z\) values that are likely to generate \(y_i\) given \((x_i, m_i)\). This information is given by the unknown posterior distribution of \(p(z|x_i, y_i, m_i)\), which is approximated by an inference neural network \(q_\phi\). We further approximate the likelihood of \(z\), that is, \(p(y|x_i, m_i, z)\), using another neural network \(p_\psi\). Here \(q_\phi\) and \(p_\psi\) are assumed to be Gaussian distribution with parameters \(\phi\) and \(\psi\). The error of this approximation is quantified by the residual term in the following equation (see Supporting Information S1 for derivation):

\[
\log p_\theta (y|x_i, m_i) = E_{z \sim q_\phi} (\log p_\psi) - D_{KL} (q_\phi \| p(z|x_i, m_i)) + D_{KL} (q_\psi \| p(z|x_i, y_i, m_i))
\]

(2)

here \(E\) is probabilistic expectation, \(D_{KL}\) is the Kullback-Leibler divergence, which is a non-negative measurement of how one probability distribution differs from a second, reference probability distribution. \(p(z|x_i, m_i)\) is the conditional prior of \(z\), which is assumed to be standard Gaussian. Under the variational inference framework, we implicitly maximize \(\log p(Y|x_i, m_i)\) by maximizing the first two terms (the reconstruction term and the regularization term) on the right hand side of Equation 2. This objective function is coined as the Evidence Lower Bound (ELBO) in the variational inference literature.

The reason for using the ELBO as our objective function is explained as follows. First, the ELBO represents a lower bound estimate of \(\log p_\theta (Y|x_i, m_i)\), with its tightness quantified by the non-negative residual term. By maximizing the ELBO, we will approximately maximize \(\log p_\theta (Y|x_i, m_i)\), while strengthening the tightness of this lower bound estimate. Second, we can apply standard stochastic gradient ascent to maximize the ELBO, scaling up this methodology to big data problem setting. Specifically, we can compute the expected gradient with respect to the reconstruction term, using Monte-Carlo reparameterization; we can compute a closed-form gradient with
respect to the regularization term, given the Gaussian distribution assumptions for $q_x$ and $p(z|x, m_i)$. A detailed explanation is given in the Supporting Information S1.

How can we apply the trained model to generate a probabilistic forecast? The answer becomes clear as we view the inference process from a Bayesian perspective. In Bayesian inference, we derive the posterior probability of an unknown variable, that is, $p(z|x, y, m_i)$, as a consequence of two antecedents: a likelihood function revealed by data, that is, $p_y = p(y|x, m, z)$, and a prior probability, that is, $p(z|x, m_i)$. In the ELBO objective function, the impact of the likelihood is reflected in the reconstruction term: we encourage accurate estimation of $y$ using information from $(x, m_i)$ and the most likely $z$. The impact of the prior is reflected in the regularization term: we force the posterior of $z$ to fully exploit the prior distribution space, encouraging each sample from the prior to generate a plausible $y$ outcome. As we optimize our model toward these two goals, the model is trained to map the $y$ agnostic conditional prior of $p(z|x, m_i)$ to the target $p(y|x, m_i)$. During forecast, without knowing $y$, a priori, we can now have good estimate of $p(y|x, m_i)$ by sequentially sampling the conditional prior of $p(z|x, m_i)$ and the likelihood of $p_y$. For the latter, we actually use the mean of $p_y$ rather than a sample of this learned Gaussian distribution as our sampling result. This widely-applied practice yields an average estimate of all possible outcomes conditioned on a sampled latent $z$ (Mukkavilli et al., 2020; Mylonas et al., 2021; Walker et al., 2016). These expected $Y$ outcomes are therefore considered as probabilistic forecasts.

### 2.3. Encoding the Climate Model Information Using Entity Embedding

Different GCMs may represent the probabilistic dependency between $X$ and $Y$ in different ways, as a result of their different configuration of dynamical core and parameterization schemes. We learn an entity embedding vector (Guo & Berkhahn, 2016) for $M$ to account for these differences in our CGF model (Figure 1). Assume $M = \{m_1, \ldots, m_N\}$ encapsulates $N$ different categorical variables representing $N$ different data-generating GCMs. Our objective is to learn continuous vector representation for each categorical variable, so that the difference among these embedding vectors accounts for the difference of how the GCMs represent the probabilistic dependency between $X$ and $Y$. To achieve this, we map each categorical variable $m$ to a vector $I_{m}: m \mapsto I_m$. Mathematically, this is equivalent to applying a linear transformation on top of the one-hot encoding of the categorical variables (Guo & Berkhahn, 2016). The mapped embeddings are the weights of the linear transformation layer, which can be trained in the same way as the other parameters of the neural net. A detailed explanation of this entity embedding technique can be found in Guo & Berkhahn, 2016. An application of this technique for weather forecast bias correction can be found in Rasp & Lerch, 2018.

### 2.4. Learning Feature Representations of Climate Variables Using Convolutional Neural Net

To build the CGF model, we use CNNs (LeCun & Bengio, 1995) to learn useful feature representations for the high-dimensional $X$ and $Y$ (Figure 1 left and bottom). Convolutional neural nets apply convolution and aggregation steps to systematically reduce small-scale information to global information, which have demonstrated particular advantage in modeling geospatial data (Miao et al., 2019; Pan, Hsu, AghaKouchak, & Sorooshian, 2019; Sadeghi et al., 2019).

The convolution steps in CNNs are implemented by cross-correlating the input feature maps using parameterized kernels. The aggregation steps are implemented by down-sampling the feature maps to synthesize local features for large-scale features. Specific for our model, we use CNNs to parameterize the inference neural net $q_x$, which yield the posterior of $z$ using the $(X, Y, M)$ information. We apply transposed convolutional neural net (TCNN, Figure 1 right) to parameterize the likelihood neural net $p_y$, which generates high-resolution predictand from low-resolution representations of $X$, $M$, and $z$. The transpose convolution operator is implemented by first padding an input grid’s neighbors with 0, followed by a regular convolution.

### 2.5. The Final Model

We obtain the CGF model by assembling the building blocks introduced in Sections 2.2–2.4. In this modeling framework, we apply stochastic gradient variational inference to efficiently approximate the target conditional distribution, which generates samples of the predictand conditioned on the predictor information and the GCM.
information. We apply the entity embedding technique to account for the impact of GCM formulation differences. We apply CNNs to learn feature representations of the high dimensional geospatial data.

It is crucial to distinguish the training mode and the forecasting mode of the CGF model. During training, the inference net \( q_\phi \) is optimized to encode \( Y \) into the latent \( z \) space, so that, first, we can sample from \( q_\phi \) to faithfully reconstruct \( Y \); second, we can fully exploit the conditional prior of \( z \) to have each of its sample generating a plausible \( Y \) outcome. In forecast mode, without knowing \( Y \) a priori, we discard \( q_\phi \), and sample \( z \) from the conditional prior of \( p(z|X, M) \). This is guaranteed by the variational inference framework explained in Section 2.2. The \( z \) samples are thereafter applied to create expected \( Y \) outcomes, which serve as our probabilistic forecasting ensemble members.

### 3. Experimental Setup

To test the proposed methodology, we set up an experiment to predict the global extended boreal winter (October-March) mean precipitation and near-surface (2 m) air temperature anomalies, using predictor of the previous July upper ocean thermal status. The applied data, model configuration, and evaluation strategy are described below.

#### 3.1. Data Set

We create a unique data set consisting 52,201 years of climate simulation to support data-driven seasonal forecast. This data set is derived from the historical climate simulations and pre-industrial control climate simulations, generated by 30 ocean-atmosphere coupled GCMs that participated in the Phase 5 (Taylor et al., 2012) and Phase 6 (Eyring et al., 2016) of the Coupled Model Intercomparison Project (CMIP5/CMIP6). Details of the applied GCMs are given in the Supporting Information S1. We regrid, clean, and normalize the data to facilitate data-driven modeling. The data and the data processing methods are described below.

We consider the monthly upper ocean thermal state as the prediction source, which is quantified by the upper ocean potential temperature profile sampled at 8 ocean top layers (5, 15, 25, 35, 45, 55, 75, 100 m). We consider monthly precipitation and near-surface (2 m) air temperature as the predictand. Both the predictor and the predictand variables are bi-linearly regridded to 2° × 2° global maps (90 × 180 grid points). We convert the data generated by different GCMs to same units (Kelvin for ocean potential temperature and 2 m air temperature, mm/day for precipitation). For the ocean potential temperature profile data, we mask the land area with 0 values.

Different normalization approaches could be applied to the regridded data. Here, we standardize the predictor at each grid point, using mean and standard deviation calculated in individual simulations. For the predictand, as we usually care about multi-month average variability, we first take the temporal average regarding the forecast setting, then, we standardize the temporally-averaged data at each grid point.

Besides the climate simulation data, we also include observational data to verify the model performance in making a practical seasonal forecast. We use ocean potential temperature reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) ocean reanalysis system 5 (ORASS, Zuo et al., 2017). Regarding the observed predictands, the precipitation observational records are obtained from the Global Precipitation Climatology Project (GPCP, Huffman et al., 1997); the 2 m air temperature reanalysis records are obtained from the fifth generation of the ECMWF atmospheric reanalyses (ERAS, Hersbach et al., 2020).

We benchmark the model performance using the North American Multi-model Ensemble (NMME) Phase-II retrospective forecast (Kirtman et al., 2014). Four dynamical seasonal forecast systems from the NMME project are considered here, namely the Community Climate System Model Version 4 (CCSM4, Jahn et al., 2012), the Community Earth System Model Version 1 (CESM1, Hurrell et al., 2013), the fourth generation of Canadian Coupled Global Climate Model (CanCM4, Merryfield et al., 2013), and the Forecast-oriented Low Ocean Resolution Model (FLOR-B01, Msadek et al., 2014). Each model has 10 ensemble members. All data cover the period of 1982–2012, and are bi-linearly regridded to 2° × 2° resolution.
3.2. Model Configuration

The structure and hyperparameter settings of the deployed CGF model are detailed in the Supporting Information S1. To train this model, we first divide the climate simulation data into the training, validation, and test sets. The test set includes the last 30 years' historical climate simulation for each GCM. These data are selected to maximally match the practical forecast verification period (1982–2012). We randomly shuffle the remaining data, and allocate 90% and 10% of them into the training set and the validation set. Note that, due to the existence of multi-annual climate patterns, such as the Pacific Decadal Oscillation, this random shuffling strategy for training/validation data set splitting may not guarantee the data to be independent between these two datasets. We revisit this issue in Section 5.2. We apply the Adaptive Moment Estimation (ADAM, Kingma & Ba, 2014) stochastic gradient ascent for training. We consider a mini-batch size of 32, and a maximum training epochs of 1000. An early stopping strategy is applied to avoid overfitting: training is stopped when the relative increase of the validation ELBO objective is less than 0.1% for 5 consecutive training epochs. To apply the trained model for probabilistic forecast, we consider a 300-ensemble member forecast. This ensemble forecast is created by first sampling the conditional prior of \( p(z|X, \mathbf{M}) \), and thereafter feed the \( z \) samples into \( p_{\psi} \) to generate plausible expected \( Y \) outcomes.

3.3. Evaluation Strategy

We consider the following two scenarios to evaluate the forecasting capability of the CGF model. First, we test how the CGF model represents the seasonal co-variability relationship between \( X \) and \( Y \) for a specific GCM. Here, the Canadian Earth System Model (CanESM) is considered as an example. To carry out this test, we fix \( \mathbf{M} \) in the CGF model to be the entity embedding vector for CanESM. Thereafter, we evaluate how the CanESM supported CGF model, denoted as CGF_{CanESM}, applies for the “model-world” forecast, as well as for the real-world forecast. Performance of the CGF_{CanESM} real-world forecast is compared with the similar GCM supported dynamical forecast (CanCM4). Second, we test how the CGF model represents the seasonal co-variability relationship for GCM ensembles. We consider the following two strategies for composing GCM-ensemble forecasts using the CGF methodology. First, we apply all the GCM entity embedding vectors to create forecast ensembles, this is equivalent to assigning equal weight to each GCM for building the CGF ensemble forecast. Second, we search an optimal GCM entity embedding vector that best fits the observational data. To obtain the optimal GCM entity embedding vector, we freeze all the neural net parameters in \( q_{\phi} \) and \( p_{\psi} \), and apply stochastic gradient optimization to search the entity embedding vector that best fits the observational data. We compare the two CGF based ensemble forecasting methodology with the NMME multi-GCM ensemble dynamical forecast.

We evaluate the forecasting performance using two widely-applied deterministic ensemble-mean scores, namely the anomaly correlation coefficient (ACC), and the normalized root mean square skill score (NRMSS). Also, we consider two common probabilistic skill scores, namely the area under the Receiver Operating Characteristics curve score (AUC, Khairin & Zwiers, 2003, see Supporting Information S1 for details), and the continuous ranked probability skill score (CRPSS, Hersbach, 2000). A summary of these skill scores, the computation methods, and the ranges is given in Table 1.

We apply bootstrap for significance test of models' forecasting skill differences. We first generate 1,000 bootstrap samples of forecasting cases using random sampling with replacement. Thereafter, we calculate the ratio of forecasting skill scores between two models for each bootstrap sample. The confidence interval is subsequently determined using a percentile method. All tests are single-side at confidence level of 95%.

4. Results and Discussion

The results include two parts corresponding to the two forecast evaluating cases: (a) comparing a single GCM (CanESM) based CGF and similar GCM based dynamical forecast (CanCM4), and (b) comparing GCM-ensemble based CGF and GCM-ensemble based dynamical forecast. For both cases, we quantify the models' forecasting skills using deterministic and probabilistic skill scores listed in Table 1. Implications for diagnosing and improving the forecasting systems are discussed thereafter.
4.1. Forecasts Based on Individual GCM

4.1.1. Forecasting Skills

Figures 2 and 3 show the forecasting skill for extended boreal winter precipitation and 2 m air temperature anomalies, using the CanESM supported CGF model (CGF\textsubscript{CanESM}) and the corresponding dynamical forecasting model (CanCM4). All forecasts here are made from July for October to March seasonal average prediction, with forecast lead time of 3–8 months. Evaluations are carried out at 2° × 2° grid scale. For both Figures 2 and 3, the four rows show forecasting skill quantified by the ACC, NRMSS, AUC, and CRPSS skill scores. The first column shows the skill of CGF\textsubscript{CanESM} for the CanESM “model-world” forecast, based on the test set data of the last 30 years’ CanESM historical simulation. The second column shows the skill of CGF\textsubscript{CanESM} for the real-world forecast, based on the observational data for the period of 1982–2012. The last column shows the skill of CanCM4 for the real-world forecast, based on the same observational data for the same period. We test the significance of skill differences between CGF\textsubscript{CanESM} and CanCM4 for real world forecast. Stippling in the second column shows locations where the CGF\textsubscript{CanESM} forecasts show significantly (95% confidence level, same for below) higher skill than the CanCM4 forecasts. Stippling in the third column shows locations where the CanCM4 forecasts show significantly higher skill than the CGF\textsubscript{CanESM} forecasts.

To interpret the evaluation results, we first discuss the commonalities shared by these forecasts, then we compare different forecasts to highlight the pros and cons of the considered forecasting approaches, and finally we discuss the implications for forecast improvements.

All the forecasts here demonstrate a roughly similar spatial distribution pattern of skill: the highest skill is achieved for the tropical Pacific, which could be attributed to the high predictability of upper ocean thermal status and the corresponding atmospheric responses; high-latitude and land regions generally achieve lower forecast skill. Specific for precipitation (Figure 2), considerable forecasting capability is achieved for certain extratropical regions, including the Middle East Asia, Western Coast of North America, Southeast U.S., South Africa, Australia, and the southeastern region of South America. Compared to precipitation, prediction for 2 m air temperature (Figure 3) generally achieves a higher level of skill, offering a broader coverage of useful forecast skill for the ocean regions and the coastal land regions.

Next, we compare the applicability of the CGF\textsubscript{CanESM} for “model-world” forecast (first column of Figures 2 and 3), and for real-world forecast (second column of Figures 2 and 3). The former informs us how well the CGF\textsubscript{CanESM} model captures the seasonal co-variability relationship as revealed by the climate simulation training data. The latter tells how well we can transfer this learned relationship for real-world forecast. The difference between these
two groups of evaluations can be attributed to the domain shift between the CanESM “model-world” climatology and the real-world climatology. This domain shift issue frequently arises as we apply a machine learning model trained from a source data distribution (i.e., climate simulations) for a different target data distribution (i.e., climate observations). In the context of climate modeling, this domain shift can be attributed to the GCM formulation biases (Pan, Anderson, Lucas, et al., 2020; Pan et al., 2021). As can be expected, CGF_{CanESM} shows an overall better applicability for the “model-world” forecast, as compared to the real-world forecast, particularly for 2 m air temperature (Figure 3).

The regions where the CGF “model-world” forecast and the CGF real-world forecast show significant differences mark the hotspots where the supporting GCM is considerably biased. A noteworthy example is the Eastern Maritime Continent (5°S–5°N, 140°E–160°E): in the “model world,” CGF_{CanESM} obtains a spatial average ACC score of 0.71/0.92 for precipitation/2m air temperature forecast (Figures 2a and 3a); in the real world, the spatial average ACC score drops to −0.06/0.31 for precipitation/2m air temperature (Figures 2b and 3b). The abnormally low forecasting capability for the Eastern Maritime Continent has been documented in the literature (Becker et al., 2014, 2020; Pan et al., 2021). As can be expected, CGF_{CanESM} shows an overall better applicability for the “model-world” forecast, as compared to the real-world forecast, particularly for 2 m air temperature (Figure 3).

Finally, we compare the CGF_{CanESM} and CanCM4 for real-world forecast (last two columns of Figures 2 and 3). These two forecasting approaches stand for two different ways in applying a GCM for seasonal forecast. The
former represents a data-driven, top-down approach that directly simulates the probability dependency between the considered predictor and predictand as revealed by the GCM simulations; the latter represents a process-based, bottom-up approach that attempts to represent the evolution of all physical processes, starting from realistically initialized climate states. Although the seasonal predictability signal learned by the CGF$_{\text{CanESM}}$ may be oversimplified, and inherits the systematical biases from the CanESM, we find CGF$_{\text{CanESM}}$ is in general comparable to CanCM4, and significantly outperforms CanCM4 for a broad range of regions (stippling in the second column of Figures 2 and 3).

We provide three explanations for this phenomenon. First, given the multi-scale nature of geophysical fluid dynamics, we can largely decouple the high-frequency weather noise and the low-frequency seasonal predictability signal (Ghil et al., 2015; S. Wang et al., 2017; Y. Zhang et al., 2021). This raises the possibility of applying the CGF methodology to simulate the interplay between the seasonal predictability signal and the internal variability noise from the climate simulation data. Second, by learning from uninterrupted runs of GCM simulations, the data-driven forecast avoids introducing state shocks, which may offer a more reliable initial state estimate for forecast. Third, CGF$_{\text{CanESM}}$ leverages the big data of long-term simulations for probabilistic seasonal forecast, which can efficiently exploit the forecast spread using large ensembles. In comparison, the CanCM4 dynamical forecast only applies 10 ensemble members. This small ensemble size leads to high sampling variability in estimating the plausible seasonal forecasting outcomes, which diminishes the forecasting performance of the CanCM4. We illustrate this aspect in the following section.

Overall, the comparison results here suggest that the CGF methodology can accurately simulate the seasonal co-variability relationship encoded in a GCM, by learning from its simulation data. This learned relationship applies well for practical seasonal forecast. Using this CGF as a powerful benchmark, we demonstrate that the considered dynamical forecast system, CanCM4, has not fully exploited the prediction capacity of its supporting

Figure 3. Similar as Figure 2 but for 2 m air temperature forecast.
4.1.2. Impact of Ensemble Size

We investigate the impact of ensemble size on models’ forecasting performance. We consider the same forecasting cases as discussed in the previous section: the CGF \textsubscript{CanESM} for “model-world” and real-world forecast, as well as the CanCM4 for real-world forecast. For each case, starting from single member forecast, we gradually double the ensemble member size, and monitor the change of spatial average ACC score for precipitation and 2 m air temperature forecast. The regions considered here are the whole globe, the tropics (30°S–30°N), and the southern/northern mid-latitudes (30°S–60°S/30°N–60°N). While it is costly to create more CanCM4 dynamical forecast members, it is much more efficient to increase the ensemble size of the CGF \textsubscript{CanESM} forecast, by simply further sampling the latent space. This convenience offers an opportunity to extend our analysis to very large ensemble forecast cases. The results are shown in Figure 4.

For all the considered cases and forecasting methodologies, the spatial average ACC score improves as we increase the ensemble size. For CGF\textsubscript{CanESM} based forecasts (purple and green), this skill improvement diminishes as the ensemble size becomes large enough. Compared to 2 m air temperature forecast (Figures 4e–4h), the prediction for precipitation (Figures 4a–4d) requires more ensemble members to reach the “diminishing point” (Clark et al., 2011), beyond which an increase of ensemble size offers no significant skill improvement. For the CanCM4 dynamical forecast (red), a doubling of ensemble size from 1 to 8 improves the ACC score, following similar trajectories as the CGF\textsubscript{CanESM} based results (purple and green). Therefore, we can reasonably expect that, a further improvement of dynamical forecasting skill is achievable if more ensemble members were available. This corroborates the previous conjecture that, due to an insufficient sampling of the forecast spread, the CanCM4 dynamical forecast has not fully exploited the prediction capacity of its supporting GCM.

It is worth noticing that, for 2 m air temperature prediction (Figures 4e–4h), a single member based CGF\textsubscript{CanESM} forecast in general shows higher ACC score compared to single member CanCM4 forecast. This is because, we build each of the CGF ensemble member using the mean of the posterior \( P_\psi \), which has already averaged the possible outcomes conditioned on a sampled latent \( z \). Meanwhile, for precipitation prediction (Figures 4a–4d), a single member based CGF\textsubscript{CanESM} forecast shows significantly lower ACC score compared to single member CanCM4 forecast, especially for the northern mid-latitude (Figure 4d), where ocean covers a less ratio, and the atmosphere exhibits strong variability with rapid regime shifts during boreal winter. This may reflect an under-fitting of the applied CGF method for the thorny problem of precipitation forecast.
4.2. Forecasts Based on GCM Ensembles

The CGF methodology proposed here can flexibly leverage the strengths of multiple GCMs by learning from their ensemble simulations. Here we test two different strategies for composing GCM-ensemble forecasts using the CGF methodology. The first ensemble strategy, denoted as CGF\textsubscript{Ens}, is to apply all the GCM entity embedding vectors to create forecast ensemble members: we apply each of the 30 learned GCM entity embedding vectors in the CGF methodology to generate 10 ensemble members, together CGF\textsubscript{Ens} has 300 ensemble members. The second ensemble strategy, denoted as CGF\textsubscript{M}, searches an optimal GCM entity embedding vector $\mathbf{M}$ that best fits the observational data, and uses this entity embedding vector in the CGF methodology to create a 300 ensemble member forecast. The optimal GCM entity embedding vector $\mathbf{M}$ is obtained by fine-tuning the entity embedding module (Figure 1 top) of the CGF model based on observational data: we freeze all the neural net parameters except the entity embedding module, and apply stochastic gradient optimization to search the entity embedding vector that best fits the observational data. As a result, this fine-tuning alleviates the GCM formulation deficiencies as reflected in the CGF methodology. We compare the two CGF based ensemble forecasts with the NMME multi-GCM ensemble dynamical forecasts (NMME\textsubscript{Ens}). NMME\textsubscript{Ens} includes four quasi-independent dynamical forecast systems, namely CCSM4, CESM1, CanCM4, and FLOR-B01. Each of these dynamical forecast systems contributes 10 ensemble members. Together NMME\textsubscript{Ens} has 40 ensemble members. We assign equal weight to each ensemble member here.

Figures 5 and 6 show the forecasting skills for extended boreal winter precipitation and 2 m air temperature anomalies, using CGF\textsubscript{Ens} (first column), CGF\textsubscript{M} (second column), and NMME\textsubscript{Ens} (third column). All forecasts are evaluated using observational data for the period 1982–2012, at 2° × 2° grid resolution. The rows show forecasting skill quantified by the ACC, NRMSS, AUC, and CRPSS skill scores. Stippling in the first/column marks locations where the CGF\textsubscript{Ens}/CGF\textsubscript{M} forecasts demonstrate significant higher skill than NMME\textsubscript{Ens} forecasts. Stippling in the third column marks locations where the NMME\textsubscript{Ens} forecasts show significant higher skill than the CGF\textsubscript{M} forecasts.

The evaluation results in Figures 5 and 6 generally demonstrate similar skill distribution patterns as the corresponding single-GCM based real-world forecasting results shown in Figures 2 and 3, with degradation or enhancement of skills depending on the considered variable, location, forecasting methodology, ensemble strategy, and skill score. To better uncover the impact of including forecasting ensemble members from independent constituent models, we compare the latitudinal, longitudinal, and global average skills of single GCM-based forecasts (CGF\textsubscript{CanESM} and CanCM4) and GCM-ensemble based forecasts (CGF\textsubscript{Ens}, CGF\textsubscript{M}, and NMME\textsubscript{Ens}). Also, we compute the number of gridpoints that the forecasting models demonstrate significant (95% confidence level) skill differences. It is worth noticing that, many grid points showing significant skill differences here have very low skills, indicating that the models find low predictability for these regions. Therefore, even if there is significant skill improvement, the forecasts may still not be considered useful. Results for precipitation and 2 m air temperature are reported in Figures 7 and 8, respectively.

For the dynamical forecast, the NMME\textsubscript{Ens} forecast (solid red) significantly outperforms the CanCM4 forecast (dashed red) regarding most of the considered measures. This skill enhancement can be attributed to (a) extensive sampling of plausible forecasting outcomes using the collected ensemble members, and (b) error cancellation among the quasi-independent forecasting models. Since the individual dynamical forecast systems in NMME\textsubscript{Ens} may not have extensively sampled the plausible forecasting outcomes to fully exploit the models' prediction capacities (Section 4.1.2), and expose the models' formulation deficiencies, we can not assign the deserved credit of skill enhancement to the error cancellation effect.

The CGF methodology offers an efficient way to extensively sample the plausible forecasting outcomes, hence better revealing the capability of individual forecasting GCMs and their values in composing the multi-model ensemble forecast. Contrary to the dynamical forecast results, CGF\textsubscript{Ens} (blue) generally yields a lower skill compared to CGF\textsubscript{CanESM} (dashed black), particularly for precipitation (Figure 7). This evidence suggests that, by simply taking into account of ensemble members generated by multiple GCMs, and assigning equal weight to each forecasting member, we may not necessarily cancel out the GCMs' errors, and outperform a single, well-performing GCM based large ensemble forecast. This is particularly obvious for precipitation, given the notable discrepancies in how different GCMs represent the complicated precipitation process.
CGF_M represents a different strategy for composing GCM-ensemble forecast: by fine-tuning the entity embedding component of the CGF model to fit the observational data, CGF_M finds an optimal GCM entity embedding vector $\tilde{M}$, and applies $\tilde{M}$ to create its forecasting ensembles. Compared to CGF_CanESM (dashed black), CGF_M (green) offers advantageous deterministic and probabilistic skill. Besides, it significantly outperforms NMMEEns (solid red) for 2 m air temperature forecast regarding a majority of the considered measures here. For precipitation forecast, CGF_M tends to offer lower skill compared to NMMEEns, especially for the deterministic skill scores (ACC and NRMSS). This suggests the difficulty for coordinating information provided by multiple GCMs for precipitation forecast using the CGF methodology.

The optimal entity embedding vector employed by CGF_M differs from any of the learned GCM entity embedding vectors, given that no GCM simulations perfectly represent the seasonal co-variability relationship. The distances between these entity embedding vectors roughly reflect the “model world” and the real world mismatch regarding the considered seasonal co-variability relationship. Here, we use a commonly applied nonlinear dimensionality reduction technique, named t-distributed stochastic neighbor embedding (t-SNE, Van der Maaten & Hinton, 2008), to visualize the distances among these entity embedding vectors. In t-SNE, we represent each high-dimensional entity embedding vector with a 2D point. The points are aligned in such a way that similar entity embedding vectors are represented by nearby points, and dissimilar entity embedding vectors are represented by distant points with high probability.
The neighboring 2D points in Figure 9 tend to be the neighboring points in the entity embedding space, suggesting that the underlying GCMs show similar seasonal co-variability relationships. For precipitation forecast (Figure 9a), the CMIP5 models (red) tend to slant to the top left, while the CMIP6 models (blue) tend to slant to the bottom right, forming two slightly discernible clusters of these two generations of GCMs. Meanwhile, the optimal entity embedding vector \( \hat{\mathbf{A}} \) lies at the top right of the map, which is far away from the center of both clusters. This suggests that a majority of the considered GCMs have significant biases in realistically representing the seasonal co-variability relationship between ocean thermal status and precipitation. Also, this corroborates the previous finding that we cannot easily cancel out the precipitation prediction biases by simply averaging the forecasts using all GCMs. The two GCMs that are closest to \( \hat{\mathbf{A}} \) are the CMIP5 GISS-E2-H-CC model and the CMIP6 IPSL-CM6A-LR model. This may suggest the advantage of these two models in seasonal precipitation forecasts.

For 2 m air temperature forecast (Figure 9b), the CMIP5 (red) and CMIP6 (blue) models do not show discernible clusters. The optimal entity embedding vector \( \hat{\mathbf{A}} \) lies roughly in the middle of all the considered GCM entity embedding vectors. This suggests that, on average, the considered GCMs can realistically represent the seasonal co-variability relationship between ocean thermal status and 2 m air temperature. The 2D points that are close to \( \hat{\mathbf{A}} \) are possibly the GCMs that offer better 2 m air temperature forecasts.

5. Concluding Remarks

5.1. Contributions

Seasonal forecast of key climate variables benefits a broad range of societal sectors. To improve seasonal forecasts, research has proceeded along two complementary paths: dynamical forecasts, which apply general circulation models (GCMs) to simulate plausible climate state evolution trajectories; and statistical forecasts, which empirically simulate the predictor-predictand co-variability in climate observations. Historically, statistical
forecast often precedes dynamical forecast in finding crucial sources of predictability, pinpointing directions for dynamical forecast improvement (Dobrynin et al., 2018; Pan, Hsu, AghaKouchak, Sorooshian, & Higgins, 2019; Seo et al., 2009). During the past decades, dynamical seasonal forecasts have achieved consistent progress, due to advances in computation, observation, and modeling. Meanwhile, statistical seasonal forecasts are gradually losing their advantage, given that the limited observational data can not support the development of powerful statistical models.

Currently, most of the seasonal forecasting information is provided by dynamical forecast. Yet, with the growing complexity of the dynamical forecast systems, and the declining role of statistical forecast as effective guidance and competitive benchmark, improvement in dynamical seasonal forecast becomes increasingly difficult. Here, we recognize three critical challenges in dynamical seasonal forecasts. First, it is difficult to faithfully initialize the forecasting GCMs, given the observation limitations and state shock problem (Balmaseda & Anderson, 2009). Second, the forecasting GCMs may not accurately represent the climate dynamics that control seasonal variability.
Third, current forecasts employing finite ensemble members may not have sufficiently sampled the possible forecasting outcomes that result from a chaotic evolution of initial climate states. A clarification of the limitations imposed by each aspect of these challenges guides efficient allocation of research resources toward improving the forecasting systems.

In this work, we have demonstrated the possibility of building a competitive statistical seasonal forecast model using the big data of climate simulations. By explicitly modeling the internal climate variability and GCM formulation differences, this statistical surrogate forecasting ideology allows us to partially bypass the aforementioned challenges in dynamical forecast, and provides a top-down viewpoint to examine how a complicated GCM encodes the seasonal predictability information. Table 2 summarizes the pros and cons of applying climate simulation supported statistical forecast, as compared to dynamical forecast.

To realize the advantages and overcome the difficulties of this statistical surrogate forecasting ideology, we propose the CGF methodology, which generates plausible samples of the considered predictand, conditioned on the predictor and GCM formulation information. Underpinning this CGF methodology is a generative deep neural network model that incorporates stochastic gradient variational inference (Kingma & Welling, 2013) for probabilistic representation of internal climate variability noise, and the entity embedding technique (Guo &

---

**Figure 8.** Similar as Figure 7 but for 2 m air temperature forecast.
that accounts for GCM formulation differences. Within this framework, we use convolutional neural nets (CNNs, LeCun & Bengio, 1995) to learn feature representations of the geospatial data.

We build a unique data set consisting 52,201 years of climate simulation data, in order to facilitate data-driven seasonal forecast. Based on this data set, we apply the CGF methodology to make a global seasonal forecast of precipitation and 2 m air temperature, using the previous season upper ocean thermal status as predictor. The CGF forecast is compared with dynamical forecast from the North American Multimodel Ensemble (NMME) project. The results show that, the CGF methodology can faithfully represent the seasonal predictability information encoded in a considered GCM. We can successfully apply this learned relationship in real-world seasonal forecast, achieving competitive performance compared to the same GCM supported dynamical forecast. Using this CGF as benchmark, we highlight the impact of insufficient forecast spread sampling that limits the forecast capability of the considered dynamical forecast system. Finally, we introduce two different strategies for composing ensembles using the CGF methodology, highlighting the potential for leveraging the strengths of multiple GCMs to achieve advantageous forecast. The contribution of this work is summarized as follows:

1. We clarify the rationale and benefits for building statistical seasonal forecast model using the climate simulation big data.
2. We develop the CGF methodology, which explicitly represents the internal climate variability noise and GCM formulation biases for probabilistic seasonal forecast.
3. We demonstrate the competitive forecasting skill of the CGF methodology. A detailed comparison between the CGF and dynamical forecast highlights the limitations in current dynamical forecast system, and pinpoints the directions for forecast improvement.

---

**Pros and Cons of Applying Climate Simulation Supported Statistical Forecast**

| Challenge                  | Pros                                                      | Cons                                                      |
|----------------------------|-----------------------------------------------------------|-----------------------------------------------------------|
| Initialization             | Avoid initialization shock by learning from uninterrupted runs of GCM simulations. | Exacerbate initialization uncertainty for neglecting processes other than the considered prediction source $X$. |
| Model formulation          | Combine power of multi GCMs by learning from multi-GCM ensemble simulations. | Inherit original GCM bias while introduce extra bias in building the statistical model. |
| Forecast spread sampling   | Efficiently infer forecast spread using a generative model. | Difficulty for approximating high dimensional distribution. |
5.2. Limitations and Future Work

The current work marks a primary attempt toward bridging dynamical forecast and empirical forecast to enhance both sides, therefore improve the accuracy and uncertainty quantification in seasonal forecast. Several deficiencies should be addressed in future work, which are highlighted below.

First, regarding the experimental setup, the problem considered here is to predict the extended boreal winter (October–March) mean precipitation and 2 m air temperature anomalies. To make such a long term averaging forecast is relatively an easier task, as compared to the conventional 3-month average forecasts that are routinely issued by operational forecasting centers. To scale up this methodology for monthly forecast, future works may explore three-dimensional (3D) convolutional neural network to explicitly sample month-by-month climate variabilities within the CGF framework, better ensuring the temporal and spatial consistency of the forecasts.

Second, regarding the model configuration, while the CGF methodology is designed to explicitly represent the impact of internal climate variability and GCM formulation differences, the results and analysis here might be compromised, due to either over-simplifying the seasonal predictability information, or failure to disentangle different uncertainty sources in building the CGF model. Regarding the former, the current CGF methodology does not take into account prediction sources beyond upper-ocean thermal status. As the CGF model neglects the impact of other potentially informative predictors, such as ocean salinity (Li et al., 2016), soil moisture (MacLeod et al., 2016), and sea ice (Cvijanovic et al., 2017), the CGF model may not have fully exploited the predictability information encoded in the GCMs, therefore failing to clearly reveal the underlying GCMs' capabilities and deficiencies. In our future work, we plan to include extra predictors to enhance the CGF methodology. Regarding the latter, although we develop distinct modules to explicitly represent the impact of internal climate variability and GCM formulation differences, these two aspects might not be well disentangled in model training. Here disentanglement refers to the idea that the change in one factor should not affect other factors, therefore providing better explainability and robust prediction. Novel machine learning techniques that learn disentangled representations have achieved considerable progress (Chen et al., 2016; Locatello et al., 2019; Siddharth et al., 2017), and may support further enhancement of the CGF methodology.

Third, regarding the model training, we have not carried out extensive experiments to test the training strategies and hyperparameter options. In particular, we adopt a random shuffling strategy to split the training and validation datasets. This may introduce auto-correlation between the training and validation sets, due to the persistent impact of low frequency climate signals. As a result, we may overfit the validation set, ending up finding a less optimal model for the test set (last 30 years of simulation, and most recent 30 years of observation). To avoid this problem, in future work, we may consider applying a continuous chunk of simulations for validation, and the remainder for training.

Fourth, regarding the connection with conventional empirical forecasting methodologies, the current CGF model applies CNN to learn feature representations from the geospatial data, which differs from the feature engineering approach (i.e., circulation patterns defined by climate indexes) in conventional empirical forecast practices. It remains contested whether this feature learning strategy offers novel insights besides the well-understood and well-defined seasonal co-variability relationships.

Finally, the CGF model allows us to directly model the probabilistic dependency between the considered predictor and predictand variable. However, this black box model does not provide human-understandable justifications for its result. Future works might explore neural network interpretability techniques to verify if the CGF methodology applies physically reasonable prediction information, and explore novel teleconnection patterns from the big data of climate simulations.

Data Availability Statement

The CMIP5 and CMIP6 climate simulation data are available through https://esgf-node.llnl.gov/projects/cmip5/ and https://esgf-node.llnl.gov/projects/cmip6/. The NMME reforecast data are available through https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/index.html?Set-Language=en. The GPCP precipitation data are available through https://psl.noaa.gov/data/gridded/data.gpcp.html. The ERA5 2°m air temperature data are available through https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-month-
Acknowledgments
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC. The views expressed here do not necessarily reflect the opinion of the United States Government, the United States Department of Energy, or Lawrence Livermore National Laboratory. This work was supported by LLNL Laboratory Directed Research and Development project 19-ER-032. This document is released with IM tracking number LLNL-RNRL-810385. The valuable comments from the two reviewers have helped us to significantly improve the quality of this work.

References
Anderson, G. J., & Lucas, D. D. (2018). Machine learning predictions of a multiresolution climate model ensemble. Geophysical Research Letters, 45(9), 4273–4280. https://doi.org/10.1029/2018GL077049
Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214–223).
Balmaseda, M., & Anderson, D. (2009). Impact of initialization strategies and observations on seasonal forecast skill. Geophysical Research Letters, 36(1), L01701. https://doi.org/10.1029/2008gl035561
Becker, E., den Dool, H. v., & Zhang, Q. (2014). Predictability and forecast skill in NMME. Journal of Climate, 27(15), 5891–5906. https://doi.org/10.1175/jcli-d-13-00597.1
Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. Geophysical Research Letters, 47(9), e2020GL087408. https://doi.org/10.1029/2020gl087408
Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., & Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. Geophysical Research Letters, 41(5), 1803–1808. https://doi.org/10.1002/2014gl063882
Chen, X., Duan, Y., Houhouli, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of the 30th international conference on neural information processing systems (pp. 2180–2188).
Clark, A. J., Kain, J. S., Stensrud, D. J., Xue, M., Kong, F., Coniglio, M. C., et al. (2011). Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. Monthly Weather Review, 139(5), 1410–1418. https://doi.org/10.1175/2011mwr3624.1
Cvijanovic, I., Santer, B. D., Bondil, C., Lucas, D. D., Chiang, J. C., & Zimmerman, S. (2017). Future loss of arctic sea-ice cover could drive a substantial decrease in California’s rainfall. Nature Communications, 8(1), 1–10. https://doi.org/10.1038/s41467-017-01907-4
Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2018). Skillful climate forecasts of the tropical Indo-Pacific ocean using model-analogs. Journal of Climate, 31(14), 5437–5459. https://doi.org/10.1175/jcli-d-17-0661.1
Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2019). Diagnosing secular variations in retrospective enso seasonal forecast skill using CMIP5 model-analogs. Geophysical Research Letters, 46(3), 1721–1730. https://doi.org/10.1029/2018gl0808598
Dohyunn, M., Domeisen, D. I., Müller, W. A., Bell, L., Brune, S., Bunzel, F., et al. (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. Geophysical Research Letters, 45(8), 3605–3614. https://doi.org/10.1029/2018gl077209
Eyring, V., Bony, S., Mézeil, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization (p. 9). Geoscientific Model Development (Online). (LLNL-JRNL-736881).
Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452–459. https://doi.org/10.1038/nature14541
Ghil, M., Chekrour, M. D., & Stepan, G. (2015). A collection on "climate dynamics: Multiple scales and memory effects". The Royal Society Publishing.
Gibson, P. B., Waliser, D. E., & DeFlorio, M. J. (2019). A critical examination of a newly proposed interhemispheric teleconnection to southwestern US winter precipitation. Nature Communications, 10(1), 1–4. https://doi.org/10.1038/s41467-019-10528-y
Guo, C., & Berkahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.
Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. Nature, 573(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7
Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:doctcr>2.0.co;2
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803
Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., et al. (1997). The global precipitation climatology project (GPCP) combined precipitation dataset. Bulletin of the American Meteorological Society, 78(1), 5–20. https://doi.org/10.1175/1520-0477(1997)078<0005:gpccpp>2.0.co;2
Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The community earth system model: A framework for collaborative research. Bulletin of the American Meteorological Society, 94(9), 1339–1360. https://doi.org/10.1175/bams-d-12-00121.1
Jahn, A., Sterling, K., Holland, M. M., Kay, J. E., Maslanki, J. A., Bitz, C. M., et al. (2012). Late-twentieth-century simulation of Arctic sea ice and ocean properties in the CCSM4. Journal of Climate, 25(5), 1431–1452. https://doi.org/10.1175/jcli-d-11-00201.1
Khairin, V. V., & Zwiers, F. W. (2003). On the roc score of probability forecasts. Journal of Climate, 16(24), 4145–4150. https://doi.org/10.1175/1520-0442(2003)016<4145:otrocs>2.0.co;2
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., III, Paolino, D. A., Zhang, Q., et al. (2014). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society, 95(4), 585–601. https://doi.org/10.1175/bams-d-12-00050.1
LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks, 3(10).
Li, L., Schmitt, R. W., Ummenhofer, C. C., & Karnaúskaus, K. B. (2016). North Atlantic salinity as a predictor of Sahel rainfall. Science Advances, 2(5), e1501588. https://doi.org/10.1126/sciadv.1501588
Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schönkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In International conference on machine learning (pp. 4114–4124).
Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. Tellus, 21(3), 289–307. https://doi.org/10.3402/tellusa.v21i3.10086
