Environmental Data Science (2021), 1–12
doi:10.1017/xxxx

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Skillful statistical prediction of sub-seasonal temperature by training on dynamical model data

Laurie Trenary[1]* and Timothy DelSole[1]

[1]Department of Atmospheric, Oceanic, and Earth Science and Center for Ocean-Land-Atmosphere Studies, George Mason University, 4400 University Drive, Virginia, Fairfax 22030, United States
*Corresponding author. E-mail: ltrenary@gmu.edu

(Received 30 June 2021)

Keywords: subseasonal prediction; machine learning; week 3-4; lasso regression; CMIP6

Abstract

This paper derives statistical models for predicting wintertime sub-seasonal temperature over the western US. The statistical models are based on Least Absolute Shrinkage and Selection Operator (lasso) and trained on two separate datasets, namely observations and dynamical model simulations. Surprisingly, statistical models trained on dynamical model simulations can predict observations better than observation-trained models. The reason for this is that dynamical models, though imperfect, can produce training sets that are orders of magnitude longer than observations.

Impact Statement

In this study we show that a statistical prediction model for observed western US wintertime temperature trained on long dynamical model simulations outperforms a statistical model trained on observations alone. This encouraging result suggests that statistical sub-seasonal prediction models can be further improved by training on both dynamical model simulations and observations.

1. Introduction

Medium range weather (up to 10 days) and long range climate forecasts (months-to-seasons) have been used operationally for decades. While the performance of forecast systems targeting these timescales have steadily improved, until recently, relatively little effort has been dedicated to the advancing prediction capabilities at the intermediate sub-seasonal (2-week-to-1 month) timescales (e.g., National Academy of Sciences, 2016). Nevertheless, there is evidence that forecasts are skillful on sub-seasonal timescales (Newman et al., 2003; Pegion and Sardeshmukh, 2011). In particular, state-of-the art numerical forecast models demonstrate skill in sub-seasonal prediction, including regional precipitation and temperature, extreme events (heat waves, cold waves, likelihood of hurricane formation), as well as tornado and hail activity (DelSole et al., 2017; Vitart and Robertson, 2018).

Tremendous societal need have driven improvements in sub-seasonal forecast capabilities. Warnings of weather-hazards such as drought or cold temperature extremes 2-to-4 weeks in advance have the potential to save lives and mitigate changing demands on energy supplies, water resources, the

agriculture sector, and fisheries (White et al., 2017). Given the far-reaching societal benefits, numerical modeling projects within the United States (SubX) and internationally (S2S) have been established with the goal of improving sub-seasonal forecast skill (Vitart et al., 2017; Pegion et al., 2019). Parallel to the establishment of numerical modeling initiatives, in 2016 the U.S. Bureau of Reclamation and the National Oceanic and Atmospheric Administration (NOAA) established a Sub-seasonal Climate Forecast Rodeo, a one-year forecast competition where participants were tasked with developing statistical models for realtime prediction of western United States (US) temperature and precipitation. The inaugural winners, Hwang et al. (2019), developed a forecast system of non-linear statistical models trained on a diverse set of observational predictors (i.e., soil moisture, geo-potential heights). Their statistical forecast system was found to be more accurate than operational US Climate Forecasting System (CFSv2). The success of the Hwang et al. (2019) forecast system, demonstrated unequivocally the utility of statistical methods for sub-seasonal prediction.

It is plausible that statistical models could be improved still further if they were trained on longer data sets. Unfortunately, observational data sets for sub-seasonal prediction are limited to 50 years or less. Moreover, the effective sample size is smaller than this because predictability mechanisms differ across seasons, suggesting that models need to be trained for each season separately, and daily temperature is serially correlated. One approach to obtaining longer training sets is to use dynamical models to generate them. Of course, dynamical models are imperfect and may be less skillful than some statistical models. Nevertheless, dynamical models are based on the laws of physics and simulate many of the complex physical processes that impact sub-seasonal predictability, so it makes sense to try to use both observational and dynamical models to constrain the statistical fit. In this paper, we train statistical models on dynamical model simulations and then use the resulting statistical models to predict observational data. To mitigate the impact of model errors, particularly those in the sub-grid parameterizations that often differ between dynamical models, we pursue a multi-model approach in which the output of several dynamical models are pooled together for training data. This leads to sample sizes orders of magnitude longer than observational data sets. Note that in this approach, observations are not used to estimate empirical coefficients, so any predictive skill arising from the resulting statistical models clearly comes from the dynamical models and demonstrates that the dynamical models simulate statistical relations relevant to sub-seasonal predictability. In this paper, we use dynamical models to estimate the regression coefficients of the statistical models, and then use observations to select the tuning parameter in the statistical model. The resulting prediction model is then compared to those trained on observations.

The design of our forecast problem is similar to that of the Forecast Rodeo. Specifically, we predict the week 3-4 local temperature at a set of grid points over the western US. Each forecast model is estimated from the Least Absolute Shrinkage and Selection Operator (lasso) method, which is a standard method in machine learning (Hastie et al., 2017). Our approach is similar to that of DelSole and Banerjee (2017) and Buchmann and DelSole (2021), except that here we predict many grid points, instead of just one local region (as in the former paper) or just one large-scale pattern (as in the latter paper). We find that statistical models trained on dynamical model data perform better than those trained on observations, suggesting that sample size is indeed a limiting factor in statistical prediction of sub-seasonal temperature.

We clarify that our goal is not to derive a statistical forecast system that outperforms that of Hwang et al. (2019). Rather, our goal is to show the potential of improving skill of statistical models by incorporating information from dynamical model simulations. Several studies suggest that the strongest source of sub-seasonal predictability comes from sea surface temperatures (SSTs) in the tropical Pacific (DelSole et al., 2017; McKinnon et al., 2016; Vitart, 2013). Accordingly, we use a set of predictors that capture variations in SSTs. It is quite likely that the performance could be improved further by increasing our predictor set to include other variables like precipitation, sea ice concentration, and indices of the Madden-Julian Oscillation, as in Hwang et al. (2019), but the dynamical models do not simulate these variables well. Moreover, SSTs are the dominant source of predictability (Vitart, 2013). For these reasons, only SSTs are included as predictors.

This paper is presented as follows. In Section 2, we describe the data sets and statistical models used to predict observed wintertime sub-seasonal temperature over the western US. The statistical models are trained either on observations or pre-industrial control runs from Climate Model Inter-comparison project phase 6 (CMIP6) archive. In Section 3, we present the results. The paper concludes with a summary of our major results.

2. Data and Methods

2.1. Observations

The target data for our study is observed 2-week mean temperature, obtained by averaging daily minimum and maximum 2-meter temperatures from the NOAA-Climate Prediction Center Global Gridded Temperature dataset (https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html). The data are available from 1979 to present, but we focus our analysis on the years 1982-2019 to avoid the large number of missing data at the start of the record. Daily sea surface temperature (SST) data are obtained from the NOAA Optimum Interpolation Sea Surface Temperature dataset for the period 1982-2019 (Reynolds et al., 2007). All data are re-gridded to a 1x1 degree resolution.

2.2. CMIP6

We also use climate model simulations for training. The particular model simulations we use are from the CMIP6 archive. To avoid the confounding effects of external forcings (i.e., greenhouse gases, anthropogenic aerosols, etc.) on predictability, we limit our selection to pre-industrial control simulations. A collection of 13 models with a total of 6889 years of daily data are selected for analysis (see Table 1). These models were selected because they provide daily surface temperature data. Consistent with observations, the target data are 2-week mean temperatures, estimated by averaging the minimum and maximum of daily 2-meter temperature. Model SST data are also used. All data are re-gridded to a 1x1 degree resolution. A description of the CMIP6 experiments can be found in Eyring et al. (2016).

2.3. Data processing

The statistical forecast models are fit at the 499 grid-points shown in fig. 1, using 2-week averages for target and predictor variables. The predictions target the winter months December to February and the forecast year for a given winter corresponds to December. To illustrate, a forecast for winter 2000, targets 2-week averages during December 2000 through February 2001. Averages for surface temperature are estimated for every start day in the winter months December-February. For instance, the 2-week average for December 1^{st} is given by the average over the period December 1^{st}-14^{th}, and the corresponding predictors are averaged for the dates November 4^{th}-17^{th}.

We estimate the climatology as a 5th order polynomial fit across all 2-week means between December-February (e.g., DelSole et al., 2017). This order of polynomial is selected to ensure that the climatological signals are accurately estimated at the different geographical locations. We tested whether the statistical models performance are sensitive to polynomial order and found no impact on skill (not shown). Also, Pegion et al. (2019) demonstrated that other methods for estimating daily climatology, such as local regression and polynomial or harmonic regression, produce nearly identical climatologies. Anomalies are estimated with respect to climatology. In a previous study by Johnson et al. (2014), it was shown that the climate change signal (trend in temperature) can inflate the skill of week 3-4 temperature predictions. For this reason, we de-trend the observed data.

We select predictors from the Pacific and Atlantic Oceans, where fluctuations in SSTs are known to impact global climate through teleconnections (Horel and Wallace, 1981). The respective domains of the predictors are shown in fig. 2 as the black and blue regions. Climatically relevant variations of SST are large-scale and characterized by distinct patterns (e.g., Deser et al., 2010). No single pattern of

SST variability drives all predictable variations in climate. For this reason, we use multiple patterns to represent large-scale variations of SSTs. Typically, these large-scale patterns of variability are estimated using empirical orthogonal function (EOF) analysis. A drawback to EOF analysis is that the patterns recovered are data dependent. This means that the leading patterns of variability that represents one data set may not be the same pattern recovered for a different data set. Ultimately, when comparisons between data sets are desirable, it is preferable to use a common basis set to describe all data. For this reason, we will isolate the large scale variations by projecting the daily SST data onto the eigenvectors of the Laplacian operator (DelSole and Tippett, 2015). The laplacian eigenvectors form a basis that depends only on the domain geometry and are orthogonal in space, with the first pattern associated with the spatial mean, the second a dipole, the third a tripole, and so forth. Projecting the data onto each Laplacian eigenvector yields a time series for each eigenvector. An example of the leading laplacians are shown in figure 3. In this study we will represent large-scale SST variations in the Pacific and Atlantic using 50 Laplacians for each basin. The inclusion of more Laplacians time series did not impact the performance of the lasso models.

2.4. Building the Statistical Forecast Systems

Our objective is to determine if training on dynamical models can produce skillful prediction models. There is no unique configuration for statistical models, so we consider a range of reasonable choices. Specifically, we construct five distinct statistical forecast systems to predict observed western US sub-seasonal winter temperatures. Each forecast system is comprised of a set of statistical models, that predict each grid-point in the target region (see fig. 1), yielding a total of 499-statistical models per forecast system. The forecast systems differ in terms of how they are fit and the data sets used for training. A description of each statistical forecast systems is provided below and summarized in Table 2.

In the description below, n and s denote the temporal and spatial indices, where $n = 1, \ldots, N$ and $s = 1, \ldots, S$. In addition, let Y_{ns} denote the target variable, z_n denotes the Nino3.4 index, and X_{np} denotes the predictors for $p = 1, \ldots, P$.

2.4.1. Benchmark

ENSO is the greatest contributor to seasonal predictability over the US, and hence it is anticipated to be a strong contributor to sub-seasonal predictability (National Academy of Sciences, 2016). Accordingly, we construct a benchmark forecast where 2-week average surface winter temperature anomalies are predicted at each grid-point from the 2-week lagged Nino3.4 index, a commonly used measure of ENSO variability. The Nino3.4 index is the spatial average of sea surface temperature anomalies between 5°S-5°N over longitudes 120-170°W.

The prediction based on Nino3.4 index is a linear regression model, where the regression coefficients $\beta_s = (\beta_{0,s}, \beta_{1,s})$ are obtained by minimizing the cost function

$$\hat{\beta}_s^{\text{Nino3.4}} = \arg\min_{\beta_s} \left\{ \sum_{n=1}^{N} \left(Y_{ns} - \beta_{0,s} - \beta_{1,s} z_n \right)^2 \right\}. \tag{1}$$

Note that the regression coefficients are chosen to minimize the prediction error at each spatial location separately. These regression models are fit using a leave-one-out approach, such that the regression coefficients $\beta_{0,s}$ and $\beta_{1,s}$ for a given winter are estimated from all other winters.

2.4.2. Lasso

The performance of the benchmark models will be compared to predictions made by a suite of statistical models based on Least Absolute Shrinkage and Selection Operator (lasso). The different lasso models have the same set of target and predictor variables. That is, the target variables are 2-week mean surface

temperatures anomalies and the predictors are large scale SST anomalies in the Pacific and Atlantic Oceans, which are represented by 50 laplacian time series for each basin, giving a total of 100 SST predictors. We estimate the lasso coefficients by either minimizing the cost function locally or across all grid-points. This distinction between lasso and OLS is discussed in more detail below.

Lasso is similar to OLS, except that the mean square error is minimized subject to a constraint on the norm of the regression coefficients (Tibshirani, 1996). More precisely, the lasso coefficients $\beta_{p,s}$ are obtained by minimizing the cost function

$$\hat{\beta}_s^{\text{single-task}} = \arg\min_{\beta_s} \left\{ \frac{1}{2N} \sum_{n=1}^{N} \left(Y_{ns} - \beta_{0,s} - \sum_{p=1}^{P} X_{n,p} \beta_{p,s} \right)^2 + \lambda \sum_{p=1}^{P} \left| \beta_{p,s} \right| \right\}, \tag{2}$$

where $|\cdot|$ denotes the absolute value and λ is a tuning parameter that determines the strength of the penalty term. As λ is increased, the lasso coefficients are shrunk toward zero. Conversely, as λ goes to zero, the penalty term has less weight and the cost function approaches the traditional OLS form.

There is no closed form solution to equation 2 and the minimization problem must be solved iteratively. We use the `glmnet` package to find lasso solutions as λ is varied (Friedman et al., 2010). Examples of how λ is selected for different lasso models are provided in Section 2.4.3.

The above formulation predicts each target variable separately. We call this formulation "single-task" lasso. We derive two single-task lasso models: one trained on observations and one trained on CMIP6 data. These lasso models are called OBS-single-task and CMIP6-single-task, respectively.

Temperature fields are spatially correlated, so making use of information between neighboring grid points during the training stage may yield a better prediction model. One approach to doing this is "multi-task lasso", which was used by Hwang et al. (2019). The cost function for multi-task lasso is

$$\hat{\beta}^{\text{multi-task}} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{s=1}^{S} \sum_{n=1}^{N} \left(Y_{ns} - \beta_{0,s} - \sum_{p=1}^{P} X_{n,p} \beta_{p,s} \right)^2 + \lambda \sum_{p=1}^{P} \sqrt{\sum_{s=1}^{S} \beta_{p,s}^2} \right\}. \tag{3}$$

In multi-task lasso, squared errors are summed over all targets and the penalty term now applies to the whole group of predictors and a given predictor is either included in the statistical model for all targets, or excluded for all targets.

We selected lasso regression because it has a good track record of producing models with out-of-sample skill, and because it sets some regression coefficients identically to zero, thus performing predictor selection and aiding in the interpretability of the statistical models.

2.4.3. Selecting the lasso Tuning Parameter

For lasso models trained on observations, the first 18 years (1982-1999) of observational data are used to fit the regression model and select λ using a 10-fold cross-validation. When CMIP6 data are used for training, the λ selected minimizes the Normalized Mean-Square-Error (NMSE) with respect to the same 18 years of observations; namely 1982-1999. A summary of how the statistical models are trained and tuning parameter λ selected is presented in Table 2.

To illustrate the λ selection process, figs. 4a-c. show curves of NMSE versus λ at three different locations for predictions made by the CMIP6-single-task model. The regression coefficients are estimated from CMIP6 simulations, yielding $\beta_s(\lambda)$, then, based on these coefficients, $\text{NMSE}_s(\lambda)$ is evaluated at each location s using observations for predictors and target variable. The λ that minimizes NMSE is denoted by a red asterisk. Two extremes cases are shown in figs. 4a and c, where the λ that minimizes NMSE is small (near zero) and large, respectively. For the case where $\lambda \approx 0$ (fig. 4a), the cost function approaches the traditional OLS form and all the predictors are included. Alternatively, when λ is large (fig. 4c), the regression coefficients are set to zero. The NMSE-λ curve shown in fig. 4b represents an intermediate scenario, where only some regression coefficients are set to zero.

2.5. Skill Metrics

Statistical model performance will be evaluated in terms of temporal correlation, spatial correlation and Mean Square Error (MSE), (Coelho et al., 2019; Jolliffee and Stephenson, 2012).

Temporal correlation is estimated at each grid-point as:

$$\rho(s) = \frac{\sum_{t=1}^{T} (F'(s,t) \cdot V'(s,t))^2}{\sqrt{\left(\sum_{t=1}^{T} F'(s,t)^2\right) \cdot \left(\sum_{t=1}^{T} V'(s,t)^2\right)}}, \tag{4}$$

where $F'(s,t)$ and $V'(s,t)$ are the matched pairs of centered forecast and verification data at location s.

Spatial similarity in the predicted and observed spatial patterns will be measured using the anomaly correlation or cosine similarity (Jolliffee and Stephenson, 2012). To avoid confusion, we will refer to this metric as the spatial correlation. Note, this is the only metric used in evaluating the machine learning forecast models of Hwang et al. (2019). Formally, the spatial correlation is

$$\gamma(t) = \frac{\sum_{s=1}^{S} (F'(s,t) \cdot V'(s,t))^2}{\sqrt{\left(\sum_{s=1}^{S} F'(s,t)^2\right) \cdot \left(\sum_{s=1}^{S} V'(s,t)^2\right)}}. \tag{5}$$

Note that this expression is similar in for to the standard correlation in the spatial domain, with the exception that the spatial mean is not removed. Importantly, the spatial correlation can be computed for each two-week period; i.e., it is a time series.

Forecast accuracy is often measured by MSE and is formally expressed as:

$$MSE(s) = \frac{\sum_{t=1}^{T} (F'(s,t) - V'(s,t))^2}{T}, \tag{6}$$

A standard approach is to compare MSE to some reference forecast (F_{ref}), typically the climatological mean corresponding to $\bar{F}_{ref}(s) = 0$, which yields the $NMSE(s)$

$$NMSE(s) = \frac{\sum_{t=1}^{T} (F'(s,t) - V'(s,t))^2}{\sum_{t=1}^{T} (V'(s,t))^2}. \tag{7}$$

A forecast with NMSE > 1 has no skill, since its MSE is greater than that of the reference forecast. The NMSE can be decomposed into its constituent parts when expressed as the Mean-Square-Error Skill Score (MSESS). Following Murphy (1988), MSESS can be expanded as follows:

$$MSESS(s) = 1 - NMSE(s) = \rho(s)^2 - \left(\rho(s) - \frac{\sigma_{F'}(s)}{\sigma_{V'}(s)}\right)^2 - \left(\frac{\bar{F}'(s) - \bar{V}'(s)}{\sigma_{V'}(s)}\right)^2, \tag{8}$$

where $\rho(s)$ is the temporal correlation (see eqn. 4), and $\sigma_{F'}(s)$ and $\sigma_{V'}(s)$ are the standard deviations for the forecast and verfication anomalies at location s, respectively. The first term on the far righthand side is the square of the temporal correlation ($\rho(s)^2$) and gives the maximum value of MSESSs. The next two terms reduce the skill score and represent the amplitude and the mean biases, respectively. For this study, the forecast and verification data are centered, consequently the mean bias term is zero. We perform a decomposition of MSESS to evaluate the trade-off between correlation and amplitude bias impacts on statistical model performance.

The above measures are either time dependent or spatially dependent. To identify a single best forecast, we use either the average NMSE

$$[NMSE] = \frac{\sum_{s=1}^{S} \sum_{t=1}^{N} (F'(s,t) - V'(s,t))^2}{\sum_{s=1}^{S} \sum_{t=1}^{N} V'(s,t)^2}, \tag{9}$$

or the spatial average temporal correlation

$$[\rho] = \frac{\sum_{s=1}^{S} \rho(s)}{S}. \tag{10}$$

The averaging procedure in eqn. 9 is standard practice and follows that of Wilks (2011).

2.6. Statistical Significance of Metrics

The data in our study are serially correlated, so standard methods of estimating uncertainty and statistical significance for performance metrics are not applicable. Accordingly, resampling methods are used to evaluate the uncertainty and statistical significance of the performance metrics. Details are given in appendix A.

3. Results

3.1. Statistical Model Selection

First, we evaluate the overall performance of the statistical models in terms of the spatially averaged NMSE and correlation coefficient to help isolate the best performing statistical models. The confidence intervals of the two metrics are represented by the vertical bars in figs. 5a and b, where the bars capture the fluctuations of each metric when different segments of the observations are used in statistical model evaluation. Confidence intervals are computed as described in sec. A.1.

Referring to fig. 5a, we see that the confidence intervals for the spatial averaged NMSE is near or above 1 for all candidate models. This illustrates the challenge of sub-seasonal forecasting: spatial average measures of local skill tend to be indistinguishable from values expected from no-skill models. Overall, the statistical models trained on CMIP6 data have lower NMSE than statistical models trained on observations, including the benchmark Nino3.4 model. In particular, the spatially averaged temporal correlation shown in fig. 5b, is low and generally negative for the statistical models trained using observation data. Only the CMIP6-single-task model consistently produces forecasts that are characterized by a positive spatially averaged temporal correlation. This result shows that training statistical models on dynamical model simulations can yield better forecasts than models trained on observations only. The OBS-multi-task model shows no improvement over the OBS-single-task. Differences in skill for the two best performing statistical models, namely CMIP6- and OBS-single-task models, are compared in terms of MSESS and its decomposition (see eqn. 8) in sec. 3.3.

3.2. Training Set Size and Statistical Model Performance

A plausible explanation for why statistical models trained on CMIP6 simulations predict observations better than observation-trained models is that the size of the training data is much larger in the former than in the latter. To test this hypothesis, we re-train the CMIP6-single-task model using a training data set that varies in length from 50 to 3000 years and evaluate model performance with respect to spatially averaged NMSE for the verification period 2000-2018. For instance, if the length of the training set is specified to be 50 years, a training set of that length is found by sampling CMIP6 data with replacement, where each year is a set of consecutive 2-week forecasts for the target months December-February. The process of selecting a training set, model re-training, and evaluation is repeated 60 times for each specified training set size. The range of uncertainty is reported as the 5^{th}-95^{th} percentiles of the 60 estimates of spatially averaged NMSE for each sample size and shown as the vertical bars in fig. 6. The threshold for a skillful forecast is denoted by the horizontal black line which shows a NMSE of 1. When the training set size is 50 years, which is more than double the number of years available for training with observations, the CMIP6-single-task model has no skill. As the sample size is increased up to 3000 years, the spatially averaged NMSE systematically drops, yielding reliably skillful forecasts (NMSE <1) when the training set size is 2000 years or greater. These results are consistent with those of DelSole and Banerjee (2017), who perform a similar sensitivity analysis on training set size.

3.3. Statistical Model Comparison

The map of MSESS estimates from the CMIP6-single-task model, shown in fig. 7a, is characterized by regions of positive values along the west-coast, from northern California up through Washington

and extending eastward into Idaho. Forecasts in the continental interior are characterized by negative MSESS. In contrast, we see MSESS values for the OBS-single-task model, shown in fig. 7d, are negative over much of forecast region.

To diagnose the sources of these errors, we decompose the MSESS according to eqn. 8 into contributions from the squared temporal correlation coefficient and amplitude bias. To aid in interpretation, all subsequent analyses are based on the square-root of these two terms. Recall that since the forecasts and verification data are centered, the mean bias term is zero. Maps of the temporal correlation for the CMIP6-single-task and OBS-single-task models are shown in fig. 7b and e, respectively. For the CMIP6-single-task predictions, positive correlations are found over much of the western US, with the largest correlations along the west-coast, and weaker and sometimes negative correlation in the continental interior. These regions of positive correlation determine the positive MSESS values along the west coast shown in fig. 7a. Positive temporal correlation coefficients recovered from OBS-single-task model are similarly concentrated in the pacific northwest. However, the OBS-single-task model have large areas of negative correlations in the southern portions of the forecast region. The positive correlations estimated for the OBS-single-task model correspond to regions of near zero MSESS values shown in fig. 7d. For both set of statistical models, the contributions of temporal correlation to forecast skill are reduced by the amplitude bias, shown in fig. 7c and f. This analysis indicates that both statistical forecast systems underestimate the amplitude of the predicted temperature anomalies. However, this amplitude bias is markedly larger for predictions made by the OBS-single-task model.

Since the upper bound of the MSESS is determined by the temporal correlation, we quantify statistical significance of this metric as discussed in Section A.2. The percentage of grid-point forecasts that predict the correct sign of temperature anomalies and the corresponding p-values are listed in the title of figure 7b and e. For CMIP6-single-task, 78% of the grid-points are characterized by positive correlation, which has a p-value of 0.021(i.e., it is field significant). In contrast, the OBS-single-task model has positive correlations only over 59% of the grid points, which has p-value 0.26 (i.e., not field significant).

The maps of temporal correlation indicate that forecasts produced by statistical models trained on CMIP6 data are generally skillful over large portions of the western US. This metric says nothing about how well the model performs for any given 2-week average. Figure 8a and b shows the distribution of the spatial correlation as a function of time for the CMIP6-single-task and OBS-single-task models, respectively. For a given winter, the vertical bar represents the 25th-75th percentile of the spatial correlation and the black asterisk gives the median value. The number of forecasts for a given winter varies between 90 and 91, depending upon leap year. The number listed next to each median gives the percentage of forecasts that predict the correctly signed temperature anomalies. We can assess the significance of the spatial correlation scores by counting the winters within which the number of positive scores exceeds the negative scores. This count is the total number of skillful forecast winters. Under the null hypothesis of no skill, the forecast has equal chances of producing positive or negative scores and the count of skillful winters should follow a binomial distribution with $p = 0.5$. Although forecasts within a winter are serially correlated, forecasts between winters are assumed independent and therefore the binomial distribution can be applied to the count of skillful winters. The CMIP6-single-task model produces forecasts where 14 of the 19 forecast winters predict the correctly signed temperature anomalies, while 13 of the 19 winter forecasts produce correctly signed anomalies in the OBS-single-task model. The overall skill of the CMIP6-single task and OBS-single-task models in predicting the spatial pattern of temperature anomalies is indistinguishable from a no-skill forecast for both statistical models. These results suggests that while CMIP6-single-task model performs well overall in terms of individual grid points, the individual statistical forecast models do not consistently predict the large-scale pattern of temperature anomalies. The study by Hwang et al. (2019), where forecasts are evaluated only with respect to spatial correlation, also found statistical prediction are characterized by a wide range in spatial correlation scores. It is worth pointing out that the performance of the statistical forecast system derived in Hwang et al. (2019), cannot be directly compared with the statistical model evaluated here, because the statistical models are trained on different datasets, target a different range of dates, and use a slightly different metric.

In terms of aggregate metrics, the CMIP6-single-task model provides the most skillful forecasts. Here we examine a pair of high- and low- skill forecasts made by this statistical forecast system. The high skill forecast, shown in fig. 10b, tends to capture the overall spatial structure of the observed temperature anomalies shown fig. 10a. Notably, the amplitude of the predicted anomalies are reduced compared to observations. Consistent with analysis presented in fig. 7, the forecast is skillful in terms of predicting the correct sign of temperature anomalies, but underestimates their amplitude. The low skill forecast, shown in fig. 10d, similarly predicts temperature anomalies that vary with latitude. However, the sign of the anomalies is completely opposite of the observed temperature anomalies shown in fig. 10c. Notably, the individual CMIP6-single-task models are capturing coherent patterns, which may suggest a common forcing mechanism is driving predictable variations over the target region.

3.4. Predictor Selection

Lastly, we examine differences in the frequency of predictor selection between the CMIP6-single-task and OBS-single-task-models. Lasso assigns zero values to selected regression coefficients, clearly indicating that the associated predictor is less important than the other predictors with non-zero coefficients. Thus, we can use non-zero coefficients as a kind of predictor selection. Since lasso is fitted at each grid point, we can collect statistics of predictor selection across grid points. The frequency with which each predictor is selected using CMIP6 and observational training data is shown in fig. 9. Regardless of the forecast model, no predictor is selected across all grid-points. That said, a notable distinction between figs. 9a and b, is the larger percentage of predictor selection for lasso models fit using CMIP6 data. Generally, laplacians time series from the Pacific are selected by a large percentage of the individual CMIP6-single-task models. In contrast, the OBS-single-task model shows less agreement in predictor selection across location. The robust selection of key predictors for the CMIP6-trained models can likely be attributed to improved estimates of regression coefficients given the vast amount of data used in statistical model training.

4. Conclusion

This paper derives statistical models for predicting wintertime sub-seasonal temperature over the western US. Our goal was to show that statistical models trained on dynamical model data can be skillful, thereby demonstrating that dynamical models provide information relevant to sub-seasonal prediction. As a reference benchmark, we use simple linear regression to predict 2-week mean temperature at each grid point based on the Nino3.4 index. This benchmark is compared to models with tens more predictors derived from lasso. The lasso coefficients are estimated in two different ways, namely under a single-task or multi-task formulation. In all cases, the forecasts are validated on observational data that was excluded from the statistical model construction.

With respect to spatial averages of NMSE and temporal correlation, the statistical models trained on CMIP6 data are more skillful than statistical models trained on observation data. Performance of the most skillful statistical model, CMIP6-single-task, is characterized by a spatially averaged NMSE <1 and a regionally average temporal correlation that is positive. This is not the case for the OBS-single-task model, a similarly configured set of statistical models trained on observation data, where the spatially averaged NMSE is greater than 1 and the spatially averaged temporal correlation is negative. A direct comparison of the MSESS between the CMIP6-single-task and OBS-single-task models, shows a greater portion of the forecast region with positive MSESS values for the CMIP6-single-task model. The MSESS for OBS-single-task is characterized by mostly negative values. The positive MSESS identified for the CMIP6-single-task model can be attributed to the statistical model's ability to correctly predict the sign of the temperature anomalies (i.e., positive and field significant temporal correlation coefficients) for large portions of the western US. In contrast, the OBS-single-task model skill in predicting the correct sign of the temperature anomalies is largely limited to the pacific Northwest. The

greatest source of error between the two statistical models is the amplitude bias, where forecasts from the OBS-single-task model are characterized by negative amplitude bias over the much of the target region, which in turn, accounts for negative MSESS values. We demonstrate that size of the training set impacts skill of the statistical models and conclude from this that the increase in sample size from using simulated data more than compensates for the limitations due to imperfections in dynamical models. These results are encouraging and suggest that skill of statistical sub-seasonal prediction models can be further improved by using both dynamical model simulations and observations in statistical model training. Presumably, the best statistical models are those that combine both dynamical model simulations and observational data. Such approaches are part of an active field of research in *transfer learning* (Zhuang et al., 2021).

In general, the single-task models performed better than multi-task models, when evaluated in terms of spatial averages of NMSE and temporal correlation. Even still, the skill of the single-task models is nearly indistinguishable from no skill, with regards to these two performance metrics. This low skill on a local basis does not necessarily mean there is no significant skill for large-scale patterns. Notably, the individual CMIP6-single-task models predict coherent large-scale temperature anomalies over the target region. This suggests that statistical model skill might be further improved by isolating predictable large-scale patterns. In future work, we develop a statistical technique that is able to identify predictable large-scale patterns despite the limited local predictability.

Acknowledgments. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1) for producing and making available their model output.

Funding Statement. This research was supported primarily by the National Science Foundation (AGS 1822221). The views expressed herein are those of the authors and do not necessarily reflect the views of this agency.

Competing Interests. None.

Data Availability Statement. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals and data can be found here https://esgf-node.llnl.gov/projects/cmip6/. The observational data are provided by the National Oceanic and Atmospheric Administration Climate Prediction Center and be downloaded directly from https://psl.noaa.gov/data/gridded/index.html.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the United States.

Author Contributions. L.T. is responsible for writing the original draft and all formal analysis and investigation. T.D. is responsible for conceptualization of the project and project supervision. Both authors contributed to revising and editing of the manuscript. All authors approved the final submitted draft.

A. Statistical Significance of Metrics

This appendix describes our methods for quantifying uncertainty in the skill metrics.

A.1. Uncertainty of Spatial Averaged Metrics

To quantify uncertainty of forecasts trained on CMIP6 data, we do not re-estimate the regression coefficients because these are very robust due to the large training set. However, the regression coefficients depend on λ. To quantify uncertainty associated with λ, we randomly select 18 distinct years from the observational record, use the corresponding matched-pairs of target and predictors to determine λ (see sec. 2.4.3), and then use the remaining 19-year record of observation data to evaluate the performance of the CMIP-single and -multi-task models. To preserve the serial correlation in the data, the paired target and predictors are sampled such that all sequential data for a given winter are drawn simultaneously. This process is repeated 1000 times to build up distributions of the spatially averaged NMSE and temporal correlation. The uncertainty is given as the 5^{th} and 95^{th} percentiles of these distributions.

For forecasts trained on observations, both λ and the regression coefficients have uncertainties that need to be quantified. To do this, a bootstrap sample of 18 distinct years from the observational record is used to train the statistical models and select λ through 10-fold cross-validation. The remaining 19 years of paired target variables and predictors are then used to evaluate the newly trained models. As before, the serial correlation in the data is preserved by selecting all sequential data for a given winter.

This process is repeated 1000 times for the OBS-single-task model to build distributions of the performance metrics. We use only 100 permutations for the OBS-multi-task because of the excessive computational requirements needed for the re-training and evaluation. The uncertainty is given as the 5^{th} and 95^{th} percentiles of these distributions.

Lastly, to estimate uncertainties for the benchmark Nino3.4 statistical model, we randomly select 37 years of paired forecasts and verification data and estimate the performance metrics for the bootstrapped samples. Again, we preserve serial correlation by selecting sequential data for each winter. This process is repeated 1000 times to build up distributions of the two metrics. The uncertainty for each metric is given as the 5^{th} and 95^{th} percentiles of the respective distributions.

A.2. Statistical Significance of Similarity Metrics

To quantify uncertainty in the grid-point estimate temporal correlation (eqn. 4) we use a permutation method (DelSole et al., 2017). Under the null hypothesis of no predictability, the forecasts and observations are independent. Thus, a permutation sample can be derived by separately permuting the year labels for forecast and observed data. The permutation sample preserves the mean and variance of the forecast and observations, but temporally misaligns the forecast-observation pairing. For this particular problem the data are permuted for winter forecasts targeting the months December-February. We permute the winter forecasts and verification data 5000 times to create 5000 realizations of correlation maps from the null hypothesis of no skill (or more precisely, the null hypothesis of exchangeability). The temporal correlation between the local forecast and verification is computed for each grid point separately and statistical significance is assessed locally by comparing the correlation value to the local 95^{th} percentile from the permutation sample. In addition, the field significance is assessed based on the counts of positive correlations. Negative correlations are not considered skillful and therefore not included in the count. This process is repeated 10,000. The resulting cumulative distribution function is then used to determine the local p-value of the number of positive correlations.

References

Buchmann, P. and DelSole, T. (2021). Week 3-4 prediction of wintertime conus temperature using machine learning techniques. *Frontiers, 3, https://www.frontiersin.org/article/10.3389/fclim.2021.697423*.

Coelho, C. A. S., Brown, B., Wilson, L., Mittermaier, M., and Casati, B. (2019). Chapter 16 - Forecast Verification for S2S Timescales. Editor(s): Andrew W. Robertson, Frédéric Vitart, Sub-Seasonal to Seasonal Prediction, Elsevier, *J. Atmos. Sci., 60:409–416. 2019, Pages 337-361, ISBN 9780128117149, https://doi.org/10.1016/B978-0-12-811714-9.00016-4*.

DelSole, T. and Tippett, M. K. (2015). Laplacian eigenfunctions for climate analysis. *J. Clim., 28(18):7420–7436*.

DelSole, T., Trenary, L., Tippett, M. K., and Pegion, K. (2017). Predictability of week-3–4 average temperature and precipitation over the contiguous united states. *J. Clim., 30(10):3499–3512*.

DelSole, T. and Banerjee, A. (2017). Statistical Seasonal Prediction Based on Regularized Regression. *J. Clim., 30(4):1345–1361*.

Deser, C., Alexander, M. A., Xie, S. P.m and Phillips, A. S. (2010). "Sea Surface Temperature Variability: Patterns and Mechanisms" *Ann. Rev. Mar. Sci., 2(1):115–143, https://doi.org/10.1146/annurev-marine-120408-151453*.

Déqué, M. (1988). 10-day predictability of the northern hemisphere winter 500-mb height by the ECMWF operational model. *Tellus, 40A:26–36*.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development, 9(5):1937–1958*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33(1):1–22*.

Hastie, .T, Tibshirani, R., and Friedman, J. (2017). The elements of statistical learning: data mining, inference, and prediction. 2nd Edition,*Springer, New York*.

Horel, J. D. and Wallace., J., M., (1981). "Planetary-Scale Atmospheric Phenomena Associated with the Southern Oscillation". *Mon. Weather Rev., 109(4):813 – 829, https://journals.ametsoc.org/view/journals/mwre/109/4/1520-0493_1981_109_0813_psapaw_2_0_co_2.xml*.

Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., and Mackey, L. (2019). Improving Subseasonal Forecasting in the Western U.S. with Machine Learning. In ACM, editor, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2325–2335*.

Jolliffe, I. T. and Stephenson, D. B. (2012). Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition,*Wiley-Blackwell, Oxford*.

Johnson, N. C., Collins, D. C., Feldstein, S. B., L'Heureux, M. L., and Riddle, E. E., (2014). "Skillful Wintertime North American Temperature Forecasts out to 4 Weeks Based on the State of ENSO and the MJO". *Weather Forecast., 29(1):23 – 38, https://journals.ametsoc.org/view/journals/wefo/29/1/waf-d-13-00102_1.xml*.

McKinnon, K. A., Rhines, A., Tingley, M. P., and Huybers, P. (2016). Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nature Geoscience, 9(5):389–394*.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weather Rev., 16:2417–2424, doi:10.1175/1520-0493*.

National Academy of Sciences (2016). *Strategies for Subseasonl to Seasonal Forecasts*. National Academy Press.

Newman, M., Sardeshmukh, P. D., Winkler, C. R., and Whitaker, J. S. (2003). A study of subseasonal predictability. *Mon. Wea. Rev., 131(8):1715–1732*.

Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., Li, W., Sinsky, E., Guan, H., Gottschalck, J., Metzger, E. J., Barton, N. P., Achuthavarier, D., Marshak, J., Koster, R. D., Lin, H., Gagnon, N., Bell, M., Tippett, M. K., Robertson, A. W., Sun, S., Benjamin, S. G., Green, B. W., Bleck, R., and Kim, H. (2019). The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment. *Bulletin of the American Meteorological Society, 100(10):2043 – 2060*.

Pegion, K. and Sardeshmukh, P. D. (2011). Prospects for improving subseasonal predictions. *Mon. Wea. Rev., 139(11):3648–3666*.

Reynolds, R., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *J. Climate, 20:5473–5496*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288*.

Vitart, F. (2013). Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc., page in press*.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L. (2017). The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society, 98(1):163 – 173*.

Vitart, F. and Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science, 1(1):3, https://doi.org/10.1038/s41612-018-0013-0*.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E. (2017). Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications, 24(3):315–325*.

Wilks, D. (2011). "Chapter 8 - Ensemble Forecasting, section 8.6.3". Statistical Methods in the Atmospheric Sciences. 3rd Edition,*Elsevier, San Diego. p359–365*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109(1):43–76*.

Table 1. List of CMIP6 models with pre-industrial control runs analyzed in this study. Models were selected if daily data was available.

CMIP6 I.D.	Experiment	Label	Number of years
AWI-CM-1-1-MR	r1i1p1f1-gn	AWI	499
CNRM-CM6-1	r1i1p1f1-gr	CNRM-CM6	499
CNRM-ESM2-1	r1i1p1f1-gr	CNRM-CM6	499
CanESM5	r1i1p1f1-gn	CanESM5	799
EC-Earth3	r1i1p1f1-gr	EC	500
EC-Earth3 -Veg	r1i1p1f1-gr	EC-Veg	499
GFDL-CM4	r1i1p1f1-gr1	GFDL-CM4	499
GFDL-ESM4	r1i1p1f1-gr1	GFDL-ESM4	499
HadGEM3-GC31-LL	r1i1p1f1-gn	HadGEM3	499
IPSL-CM6A-LR	r1i1p1f1-gr	IPSL	499
MRI-ESM2	r1i1p1f1-gn	MRI	199
NorEMS2-LM	r1i1p1f1-gn	NorESM2	300
UKESM1-0-LL	r1i1p1f2-gn	UKESM2	1099

Table 2. Summary of Statistical Forecast Models

Statistical Model Name	β_s estimation	λ selection
Nino3.4 (benchmark)	OLS: observation using leave-one-out (37 years)	N/A
OBS-single-task	Lasso: observations (18 years)	10-fold cross-validation
OBS-multi-task	Lasso: observations (18 years)	10-fold cross-validation
CMIP6-single-task	Lasso: CMIP6 (6889 years)	observations (18 years)
CMIP6-multi-task	Lasso: CMIP6 (6889 years)	observations (18 years)

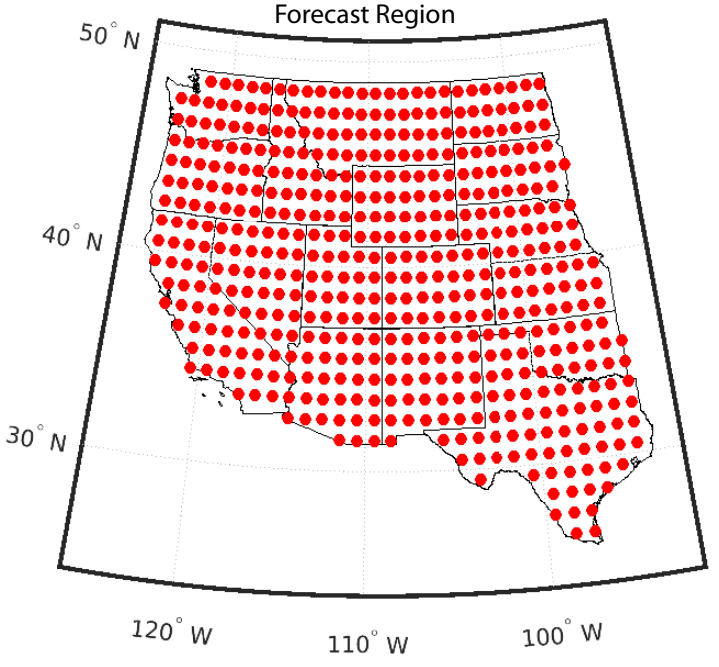

Figure 1. *Map of the forecast target region. Each red dot denotes a forecast location on a 1x1 degree grid.*

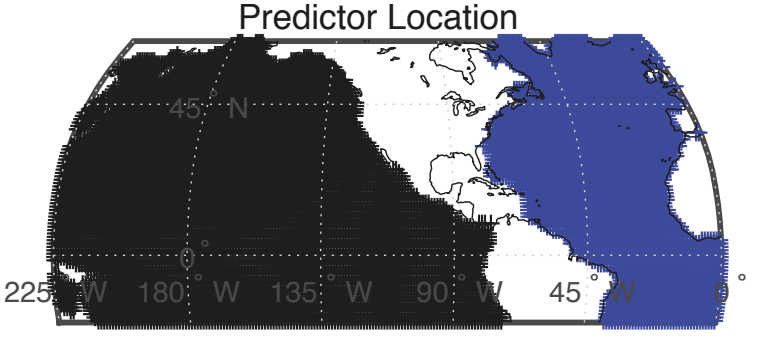

Figure 2. *Map depicting predictor locations. Black and blue denote the domains of the Pacific and Atlantic basins, respectively. Large-scale variations in both basins are represented by the leading 50 Laplacian eigenvectors.*

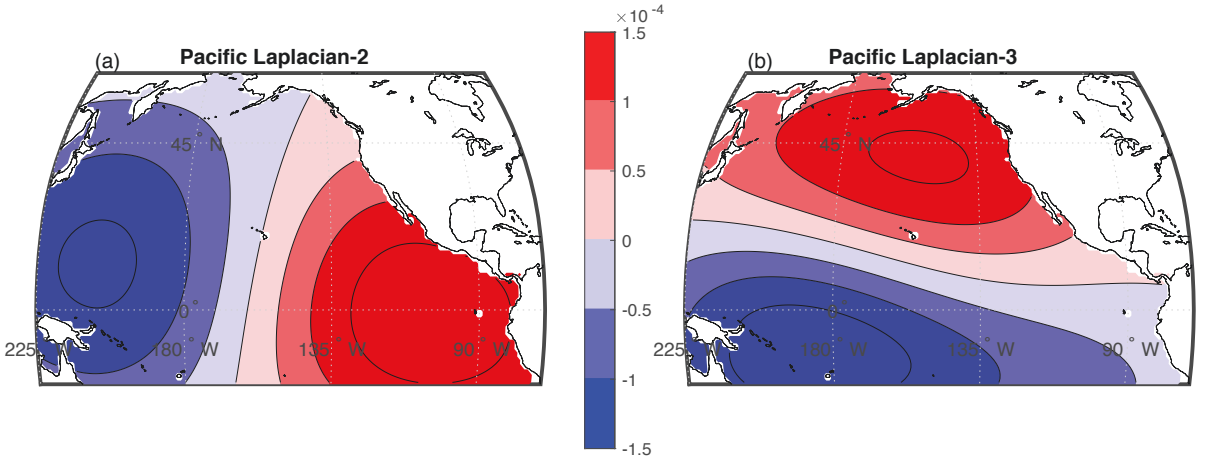

Figure 3. *The 2nd (a) and 3rd (b) eigenvectors of the Laplacian operator for the Pacific basin.*

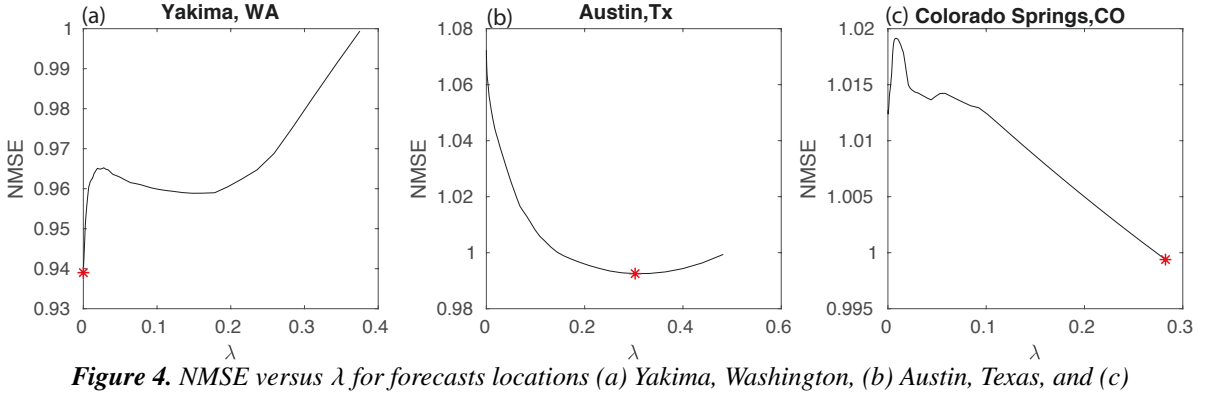

Figure 4. *NMSE versus λ for forecasts locations (a) Yakima, Washington, (b) Austin, Texas, and (c) Colorado Springs, Colorado. The NMSE curves are estimated from predictions made with the CMIP6-single-task model and validated with respect to observations for winters (DJF) during 1982-1999. A red asterisk denotes the λ that minimizes the NMSE.*

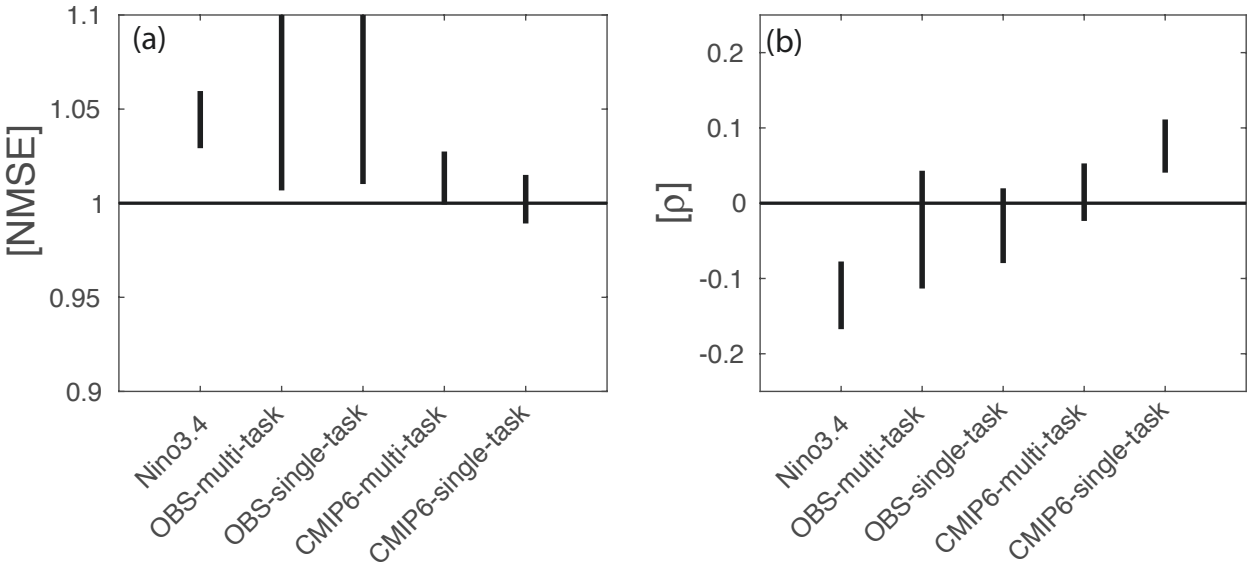

Figure 5. *Spatial averages of the (a) NMSE and (b) temporal correlation. The horizontal black line denotes a NMSE of 1 in (a) and the zero correlation in (b). Five statistical models are compared: the benchmark Nino3.4 regression model, two observation-trained and two CMIP6-trained lasso models. The vertical black bars denote the 95% confidence intervals obtained from the bootstrap method applied to observation data in the period 1982-2018. The precise bootstrap method varies across statistical models (see appendix for details).*

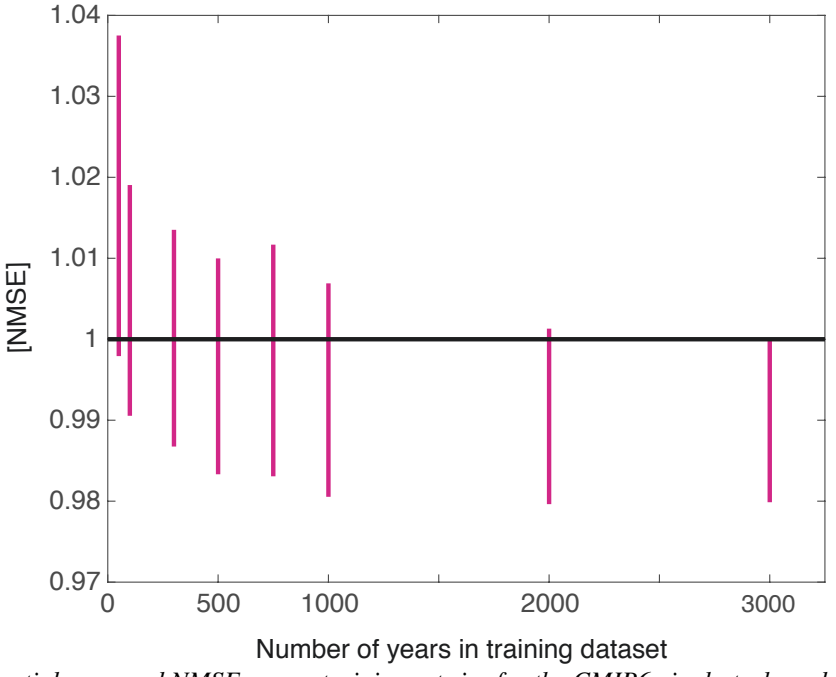

Figure 6. *Spatial averaged NMSE versus training set size for the CMIP6-single-task model. The bars give the 5th-95th percentiles from randomly selecting CMIP6 data, training the single-task lasso on this data set, then verifying the lasso model on observed winter (DJF) temperatures for the period 2000-2018. The number of years included in the training dataset is varied from 50, 100, 300, 500, 750,1000, 2000, 3000 years. The percentiles are computed from 60 repetitions.*

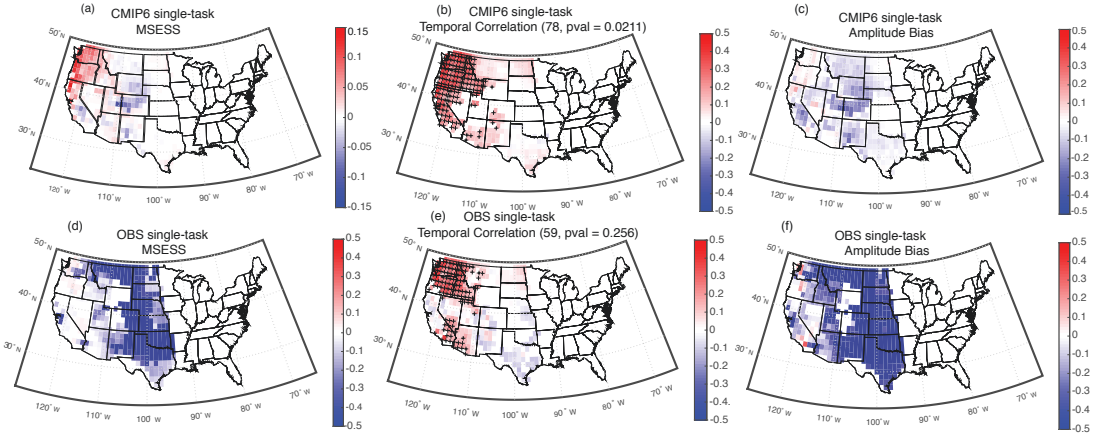

Figure 7. *Skill performance of CMIP6 single-task (top row) and OBS-single-task (bottom row) forecasts, using MSESS (first column) and its decomposition into temporal correlation (middle column) and amplitude bias (last column). Each metric is evaluated with respect to observed and forecast winter (DJF) temperature anomalies for the period 2000-2018. The percentage of forecasts that positively correlate with verification data is listed in parentheses. Statistical significance of the correlation maps is estimated with respect to the procedure discussed in the appendix, with the corresponding p-value listed in the title. The + sign denotes grid-points where the correlation is locally significant.*

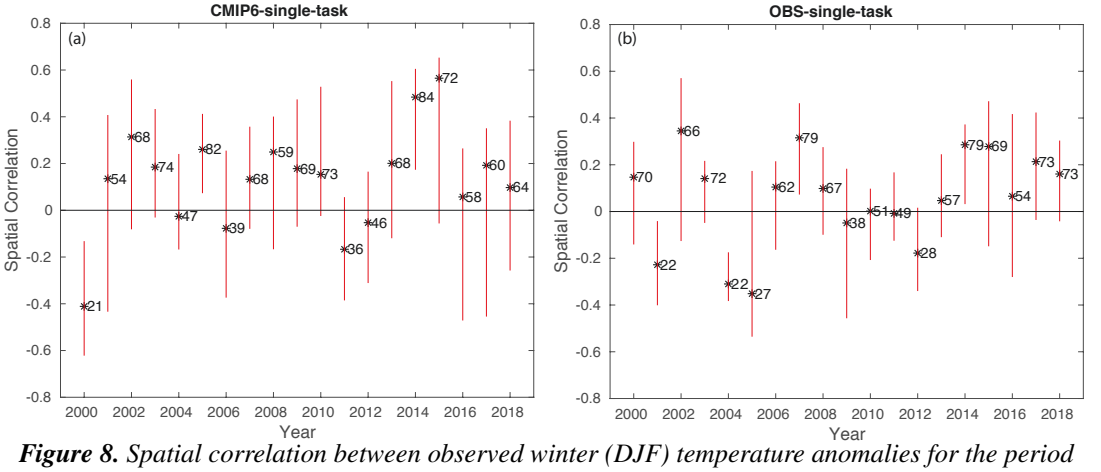

Figure 8. *Spatial correlation between observed winter (DJF) temperature anomalies for the period 2000-2018 and predictions made with (a) CMIP6-single-task and (b) OBS-single-task models. The vertical lines represent the 25^{th}-75^{th} percentiles of the 90 (or 91 if there is a leap year) spatial correlations during the winter year (where the year corresponds to the December). The median is denoted by the black asterisk. The percentage of forecast within a given winter that have a positive spatial correlation score with observations is listed next to median.*

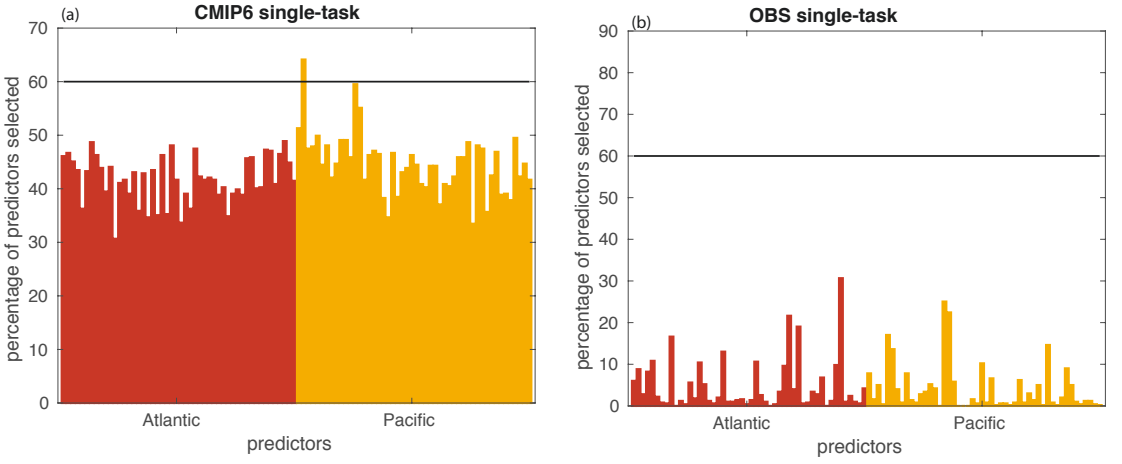

Figure 9. *Percentage of predictors selected across all 499 grid-points for the (a) CMIP6-single-task and (b) OBS-single-task models. The horizontal black line denotes the 60% selection level.*

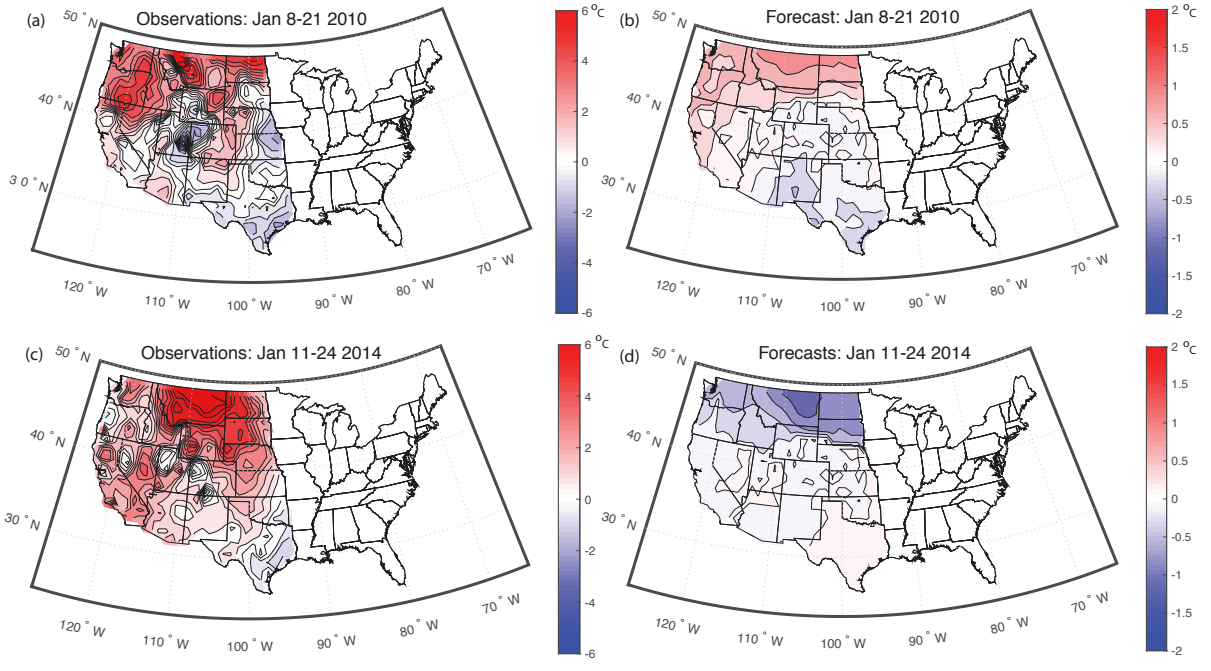

Figure 10. *Observed 2-week average temperature anomalies (left) and CMIP6-single-task forecasts (right) for a high-skill event on January 8^{th}-21^{st} 2010 (top) and low-skill event on January 11^{th}-24^{th} 2014 (bottom).*