

RESEARCH ARTICLE

Advancing interpretability of machine learning prediction models

Laurie Trenary¹* and Timothy DelSole¹

¹Department 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 March 2022)

Keywords: subseasonal prediction; machine learning; model interpretation

Abstract

This paper proposes an approach to diagnosing the skill of a machine learning prediction model based on finding combinations of variables that minimize the normalized mean square error. This technique is attractive because it compresses the positive skill of a forecast model into the smallest number of components. The resulting components can then be analyzed much like principal components, including the construction of regression maps for investigating sources of skill. The technique is illustrated with a machine learning model of week 3–4 predictions of western US wintertime surface temperatures. The technique reveals at least two patterns of large-scale temperature variations that are skillfully predicted. The predictability of these patterns are generally consistent between climate model simulations and observations. Not surprisingly, the predictability is determined largely by sea surface temperature variations in the Pacific, particularly the region associated with the El Niño–Southern Oscillation. This technique provides a tool for understanding what processes contribute to predictability and how these processes differ between dynamical models and observations.

Impact Statement

Machine learning has emerged as a powerful tool for making climate predictions, but the associated models often are too complex to interpret. Methods for extracting useful scientific knowledge from machine learning models have been developed (e.g., explainable artificial intelligence), but most of these methods apply only to relatively simple outputs. In contrast, many climate applications require predicting a spatial field of variables. This paper proposes an approach to reducing the dimension of the prediction field by finding components that are predicted with the most skill. This technique is illustrated by training separate machine learning models at hundreds of spatial locations, and then using this technique to show that only a few patterns are predicted with significant skill. Individual patterns can then be analyzed using regression techniques to diagnose the source of the skill.

1. Introduction

Machine-learning techniques can produce climate forecasts that outperform predictions made by state-of-the-art numerical forecast models (Hwang et al., 2019). Nevertheless, machine learning models are criticized because they are not based on physics and are often difficult to interpret. Regardless of these criticisms, if a machine learning model consistently outperforms the best physics-based model, then it

has the potential to lead to deeper scientific understanding of processes driving predictable variations in the climate system. The question arises as to how to extract scientifically meaningful information from machine learning models. Methods for extracting such information are often called explainable Artificial Intelligence, or *explainable AI*.

A variety of approaches have been proposed in explainable AI. One approach is to examine the sensitivity of a machine learning model to changes in the training set. For instance, if certain predictors are critical to the skill of a model, then removing them ought to reduce the skill of the model. Various methods for ranking predictor importance involve sequential forward and backward selection methods and permutation methods (McGovern et al., 2019). Unfortunately, such methods are computationally ineffective to implement in deep learning models (McGovern et al., 2019; Toms et al., 2020). Also, it is well known that forward and backward selection lead to misleading interpretations in linear regression (see Harrell, 2001, page 56), and these problems will certainly be amplified in machine learning where thousands of predictors are common. Moreover, predictor importance typically varies with predictand, which complicates interpretability in multivariate prediction.

Another approach to explainable AI is backward propagation. This method determines the input pattern that most closely reproduces a given output from a trained neural network (e.g., McGovern et al., 2019; Toms et al., 2020). This approach may provide clues as to which features in the input of a neural network are important for producing a given output. Though, the result may be difficult to interpret if multiple modes of variability contribute to the output, in which case the input pattern represents a mixture of modes. A related technique is layer-wise relevance propagation, which produces a heat map in the dimensions of the original input that identifies the input features most relevant for the network output (Toms et al., 2020). Both methods are often used in classification problems with few output categories (Gange et al., 2019; McGovern et al., 2019; Toms et al., 2020). However, for climate predictions targeting large geographical area, there are a large number of outputs, in which case it is unclear how effective these methods would be in aiding in interpretation.

In this paper, we are concerned with diagnosing and interpreting the skill of a model that predicts an entire spatial field, such as surface temperature over a geographic region. We are particularly interested in sub-seasonal predictions, for instance, predicting week 3-4 temperature, where the skill is low when measured with respect to local measures of normalized mean square error (NMSE) or correlation. The low skill locally does not preclude the existence of predictable large-scale patterns, since a significant source of sub-seasonal predictability comes from large-scale atmospheric teleconnections (National Academy of Sciences, 2016). Despite the low skill as measured by these metrics, it is possible that a large-scale pattern is predictable, but this predictability is obscured locally by unpredictable weather variability that dominates at each grid point. The question arises as to whether this predictable large-scale pattern, if it exists, can be extracted from the forecast data. Renwick and Wallace (1995) review various approaches to extracting such patterns. Here, we focus on a method due to Déqué (1988), which we call Skill Component Analysis (SCA), following DelSole and Tippett (2022), who review this method. SCA finds linear combinations of data that minimizes the NMSE. This methodology is analogous to Predictable Component Analysis (PrCA), except PrCA yields eigenvectors that maximize predictability, which is distinct from skill (DelSole and Chang, 2003; DelSole and Tippett, 2022).

One of the limitations of SCA is that it requires a relatively large amount of data. Recently, Trenary and DelSole (2022) derived a machine learning model by training it on thousands of years of daily data from a multi-model set of physics-based simulations. The resulting machine learning model could skillfully predict observed week 3-4 temperature over the western US during winter, despite never being trained on observations. The large training set provides an opportunity to test the ability of SCA to extract low-dimensional predictable components of a machine learning model.

The purpose of this paper is to apply SCA to a machine learning model for sub-seasonal prediction. The SCA is used to diagnose the components of temperature that are most skillfully predicted by a machine learning model and to infer the source(s) of predictability associated with these patterns. In Section 2, we describe the data sets and statistical model for predicting observed wintertime sub-seasonal temperature over the western US. In Section 3 we describe SCA, and in Section 4 we use

the technique to identify the most skillfully predicted large-scale temperature patterns in the statistical forecast system of [Trenary and DelSole \(2022\)](#). The paper concludes with a summary of our major results.

2. Data and Methods

2.1. Statistical Forecast System

We examine predictability in the statistical forecast system developed in [Trenary and DelSole \(2022\)](#). Their forecast system targets week 3-4 wintertime temperature anomalies over the western United States and is comprised of 499 grid-point lasso regression models. The predictors for this forecast system 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. Each grid-point model is trained on pre-industrial control simulations from 13 models from the Climate Model Inter-comparison Project phase 6 (CMIP6) archive (see for model details, [Eyring et al., 2016](#)), comprising a total of 6889 years of daily data. The target and predictors are 2-week means, and predictions target December-February. Anomalies are defined with respect to a climatology estimated as a 5th order polynomial fit in time across all 2-week means between December and February. Further details of the models and observations, as well as justifications for the particular choices in the model and analysis, can be found in [Trenary and DelSole \(2022\)](#).

3. Skill Component Analysis

Among the statistical models examined in [Trenary and DelSole \(2022\)](#), here we examine only the CMIP6-single-task model, which was the best performing model in that study. Since this is the only model examined in the present paper, this model will be referred to as simply the statistical model. The skill of this statistical model was measured by spatially averaged correlation, which was less than 0.1, and spatially averaged NMSE, which was indistinguishable from 1. This low skill for sub-seasonal prediction is consistent with previous studies (e.g., [He. et al., 2021](#); [Hwang et al., 2019](#); [DelSole et al., 2017](#); [Pegion et al., 2019](#)). As discussed in the introduction, the apparent low skill might be an artifact of the choice of skill measure. Here, we attempt to diagnose skill in this forecast by finding the linear combination of variables that minimize the NMSE. Let t and s denote the temporal and spatial indices, where $t = 1, \dots, T$ and $s = 1, \dots, S$. Let $F(s, t)$ and $V(s, t)$ denote the anomaly forecast and target variables, respectively. Then, the linear combinations are

$$r_V(t) = \sum_{s=1}^S q(s)V(s, t) \quad \text{and} \quad r_E(t) = \sum_{s=1}^S q(s) (F(s, t) - V(s, t)), \quad (1)$$

where $q(s)$ contains the linear coefficients. The *NMSE* associated with this component can be written as

$$NMSE = \frac{\mathbf{q}^T \Sigma_E \mathbf{q}}{\mathbf{q}^T \Sigma_V \mathbf{q}}, \quad (2)$$

where Σ_E and Σ_V are the sample covariance matrices of $(F(s, t) - V(s, t))$ and $V(s, t)$, respectively. In SCA, we seek the $\mathbf{q}$ that minimizes the *NMSE* in eqn. 2. Following [DelSole and Tippett \(2022\)](#), this minimization problem leads to the generalized eigenvalue problem

$$\Sigma_E \mathbf{q} = \lambda \Sigma_V \mathbf{q}. \quad (3)$$

Typically, this eigenvalue problem has S distinct solutions, where the eigenvalue λ gives the value of *NMSE* corresponding to a given eigenvector $\mathbf{q}$. Accordingly, the eigenvalues are ordered from *smallest*

to largest, $\lambda_1 < \dots < \lambda_S$, and the corresponding eigenvectors are denoted $\mathbf{q}_1, \dots, \mathbf{q}_S$. The first eigenvector has the smallest possible $NMSE$ and is therefore the most skillful component. The associated time series for this component is $\mathbf{q}_1^T \mathbf{V}$, as in eqn. 1. The second eigenvector gives the smallest $NMSE$ out of all combinations whose time series are uncorrelated with the first, and is therefore the second most skillfully predicted pattern, and so on. This methodology is analogous to Predictable Component Analysis, except Predictable Component Analysis, yields eigenvectors that maximize predictability, which is distinct from skill (DelSole and Chang, 2003; DelSole and Tippet, 2022). Note that unlike EOF analysis where the eigenvectors and principal component time series are separately orthogonal, in SCA only the time series are uncorrelated.

Typically, dimension reduction is performed prior to SCA to avoid singular covariance matrices. For our problem, the CMIP6 data set is sufficiently large that the covariance matrices are non-singular. Nevertheless, applying SCA to CMIP6 data yields components that are skillful in CMIP6, but have no skill in observational data. We interpreted this result to mean that the SCA overfit toward the sample. To mitigate overfitting, we project the western US data onto the leading 50 Laplacian eigenvectors prior to performing SCA. Our main conclusions showed little sensitivity for truncations between 20-50.

The significance of a given value of $NMSE$ was evaluated with respect to the sampling distribution of the eigenvalues under the null hypothesis of no skill. To do so, we randomly sample pairs of forecasts and verifications in the CMIP6 data set. The forecast-verification pairs are selected for a complete winter, to preserve weekly serial correlations (if any), and for a complete set of laplacian series for predictors and predictands, to preserve the spatial correlations. We then randomly shuffle the years for the forecast data to misalign the forecast and verification data, and then perform SCA on this data set. This process is repeated 5000 times to build up an empirical distribution for the individual eigenvalues. An SCA component is considered significant if its $NMSE$ falls below the 5th percentile from the distribution of randomly shuffled data.

4. Results

The rminimized NMSEs from SCA are shown in fig. 1, where the black asterisks denote the eigenvalues (or optimized NMSE) and the red curve is the 5th percentile from the randomized forecasts. There are notably a number of modes that are below the significance curve. However, only the first two modes are well separated. In the event of no predictability, we would expect the $NMSE$ to be equal to or greater than one. Indeed, this is evident in figure 1, where we see that the eigenvalues for the first two modes are below 1 and the $NMSE$ continues to increase for larger truncations. The 49th Skill Component deviates from the significance curve for reasons that are unknown to us, but it is an isolated component at the most extreme no-skill limit (its NMSE is 1.4), so it has no bearing to understanding sub-seasonal skill. The spatial patterns of the two leading Skill Components are shown in figure 2. These two patterns are estimated for CMIP6 data by regressing the 6889-year time series of the leading two Skill Component time series onto the multi-model CMIP6 temperature anomaly data. The pattern of the most skillfully predicted pattern is shown in fig. 2a and is similar to the canonical El Nino-Southern Oscillation (ENSO) teleconnection patterns (e.g., Trenberth et al., 1998). The second most skillfully predicted pattern is shown in fig. 2b and projects strongly onto the leading mode and the ENSO teleconnection pattern.

We next perform a similar calculation using observations. Specifically, we first compute a linear combination as in (1) using observational data for $V(s, t)$, which results in a time series $r(t)$, and then compute regression coefficients between this time series and the observed grid point temperature data over the western US. The resulting pattern for the leading SCA component is shown in fig. 3a. Unlike the model results that are characterized by a meridional temperature dipole (fig. 2a), the most skillfully predicted pattern recovered from observations is a zonally oriented temperature dipole. Why are the patterns different? Only 19 years of observed data are used to estimate this pattern and it is possible that the sample size may impact the recovered pattern. To test this, a climate model is randomly selected,

in this case GFDL-ESM4, and the pattern of the leading Skill Component is estimated for different 19 year periods. Some representative results, shown in figs. 3b and c, indicate that the most skillfully predicted pattern is sensitive to sampling. As such, we cannot conclude that the dynamical model and observed patterns are different, rather the difference is likely an artifact of sample size.

With the leading two components identified (figs. 3a and b), we now quantify how well the grid-point based models predict these large scale patterns for observations and each CMIP6 model. The predictions and associated verification data are both projected onto the SCA eigenvectors and the correlation between the two time series is computed. The CMIP6 model data are sampled to have the equivalent number of years as observations (i.e., 19 years) when estimating the correlation.

The results for the 1st and 2nd skill components are shown in figs. 4a and b, respectively. The distribution of correlation coefficients found for predictions in each CMIP6 model are denoted by the vertical bars, which represent the 5th and 95th percentiles, and the mean correlation is denoted by the black asterisks. The dashed lines in figs. 4a and b, denotes the correlation for predictions of observations. For the leading skill component, shown in fig. 4a, the distribution of correlations overlap for the different dynamical models, indicating that there is consistency in the predictability of this large-scale pattern across the dynamical models. For all but three of the dynamical models, the distribution of correlations include observations, indicating consistency in predictive skill of this pattern in observations and within 10 CMIP6 models. The predictive skill of this pattern differ from observations in the two CNRM models and the MRI model, suggesting that these particular dynamical models may be deficient in simulating the physical processes contributing to the predictability of this pattern. That said, it is worth noting that the skill in predicting this pattern is significantly larger than the skill based on the average of local correlations (see fig. 7, Trenary and DelSole, 2022). Moreover, the distributions of correlations in all but one of the dynamical models are distinct from zero, indicating that a robust source of predictable variations of western US surface temperature exists in a majority of the dynamical models and is linked to the same pattern. This analysis confirms that the grid-point lasso models trained on CMIP6 data are skillfully predicting a large scale pattern. Moreover, this skill associated with prediction of this pattern is generally consistent across dynamical models and with observations.

The range of correlations for the 2nd skill component, shown in fig. 4b, are generally reduced relative to the 1st and the skill in predicting this mode remains consistent across climate model. However, there are several dynamical models that include zero, indicating that the skill in predicting this pattern is not significant.

Lastly, to determine the source of predictability, the time series of the leading skill component recovered from multi-model CMIP6 data and observations is regressed onto their respective SST anomalies. The resulting regression maps are shown in figs. 5a and b. The most prominent feature of these maps is the Pacific ENSO pattern which is identified in both datasets. Outside of the Pacific, the SST regressions differs. The prominence of the ENSO-SST pattern is not surprising given that the SCA patterns resembles the traditional canonical teleconnection patterns. The regression pattern recovered for the 2nd skill component (not shown) also projects strongly onto ENSO. The above analysis was repeated by regressing the leading skill component associated with the predicted western US surface temperatures onto SST and similar patterns were recovered. This analysis suggests that the skillful prediction is associated with ENSO. A previous study by DelSole et al. (2017), similarly found that sub-seasonal predictability of wintertime temperatures over the US can largely be attributed to ENSO. It is perhaps not surprising that ENSO is a major source of predictability for sub-seasonal forecasts, when it is the dominate source of predictability on seasonal timescales (National Academy of Sciences, 2016). That said, it is worth noting that the statistical model analyzed here outperforms a benchmark forecast where the Nino3.4 index is the sole predictor (see fig. 5, Trenary and DelSole, 2022). This indicates that the ENSO related SST variations impacting predictability are not entirely captured by the Nino3.4 index. To test this, the CMIP6-single-task predictions for CMIP6 and observation data are projected onto the SCA eigenvectors and then correlated with the associated Nino3.4 index. The correlation is 0.86 for observations and 0.67 for CMIP6. This confirms that the skill in predictability is linked to ENSO in

both observations and CMIP6, but the Nino3.4 index only captures some fraction of the relevant SST variations.

5. Conclusion

This paper proposes an approach to diagnosing the skill of a machine learning model based on finding combinations of variables that minimize the NMSE. This approach was proposed by Déqué (1988) in the context of diagnosing weather prediction models and recently reviewed by DelSole and Tippett (2022). We apply the method to statistical forecasts for week 3-4 prediction of western US wintertime temperatures. This is an instructive example because the spatially averaged NMSE of these forecasts is indistinguishable from one, suggesting no skill. Despite this, the optimization technique identifies at least two large-scale temperature variations that are skillfully predicted by the machine learning model. The apparent low skill is an artifact of the skill measure, which is computed first by evaluating skill at each grid point and then averaging this measure across grid points. Unfortunately, unpredictable weather noise dominates each grid point and thereby obscures whatever predictability may exist from large-scale teleconnection patterns. The leading pattern resembles the canonical ENSO teleconnection pattern and the skill in predicting this pattern is consistent across a majority of the different CMIP6 models and observations. Predictability of this pattern is inconsistent between three CMIP6 models and observations, suggesting that these dynamical models are deficient in simulating key physical processes that contribute to predictable variations in western US surface temperature anomalies. We further show that the source of predictability for this pattern is largely related to Pacific SST anomalies and ENSO in particular. The second most skillfully predicted component is predicted with far less skill in both observations and dynamical models, and some dynamical models demonstrate no skill in predicting this mode. As is true for the leading mode, the second mode appears to be forced by ENSO. Though these results confirm our expectations about the source of predictability in this particular case, the technique is sufficiently general that it may provide new insights into prediction problems in which the source of predictability is less well understood.

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.

References

- DelSole, T. and Chang, P. (2003). Predictable component analysis, canonical correlation analysis, and autoregressive models. *J. Atmos. Sci.*, 60:409–416.
- 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.
- DelSole, T. and Tippet, M. (2022). Statistical Methods for Climate Scientists. 1st Edition, Cambridge University Press, Cambridge. doi:10.1017/9781108659055.
- 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.
- Gagne II, D. J., Haupt, S. E., Nychka, D. W., and Thompson, G.. Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. *Mon. Weather Rev.*, 147(8):2827–2845, <https://journals.ametsoc.org/view/journals/mwre/147/8/mwr-d-18-0316.1.xml>.
- Harrell, F. E. (2001). Regression Modeling Strategies. 1st Edition, Springer, New York.
- He, S., Li, X., Trenary, L., Cash, B. A., DelSole, T., Banerjee, A. (2021). Learning and Dynamical Models for Sub-seasonal Climate Forecasting: Comparison and Collaboration.. *arXiv*, <https://arxiv.org/abs/2110.05196>.
- 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.
- McGovern, A., Lagerquist, R., Gagne, J. D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, <https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml>.
- National Academy of Sciences (2016). *Strategies for Subseasonal to Seasonal Forecasts*. National Academy Press.
- 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., Tippet, 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.
- Renwick, J. A. and Wallace, J. M. (1995). Predictable anomaly patterns and the forecast skill of northern hemisphere wintertime 500-mb height fields. *Mon. Wea. Rev.*, 123:2114–2131.
- Toms, B. A., Barnes, E. A., Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, 12(9): e2019MS002002, <https://doi.org/10.1029/2019MS002002>.
- Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., and Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, 103:14291–14324.
- Trenary, L. and DelSole, T. (2022). Skillful statistical prediction of sub-seasonal temperature by training on dynamical model data. *Environ. Data. Sci.*, under review.

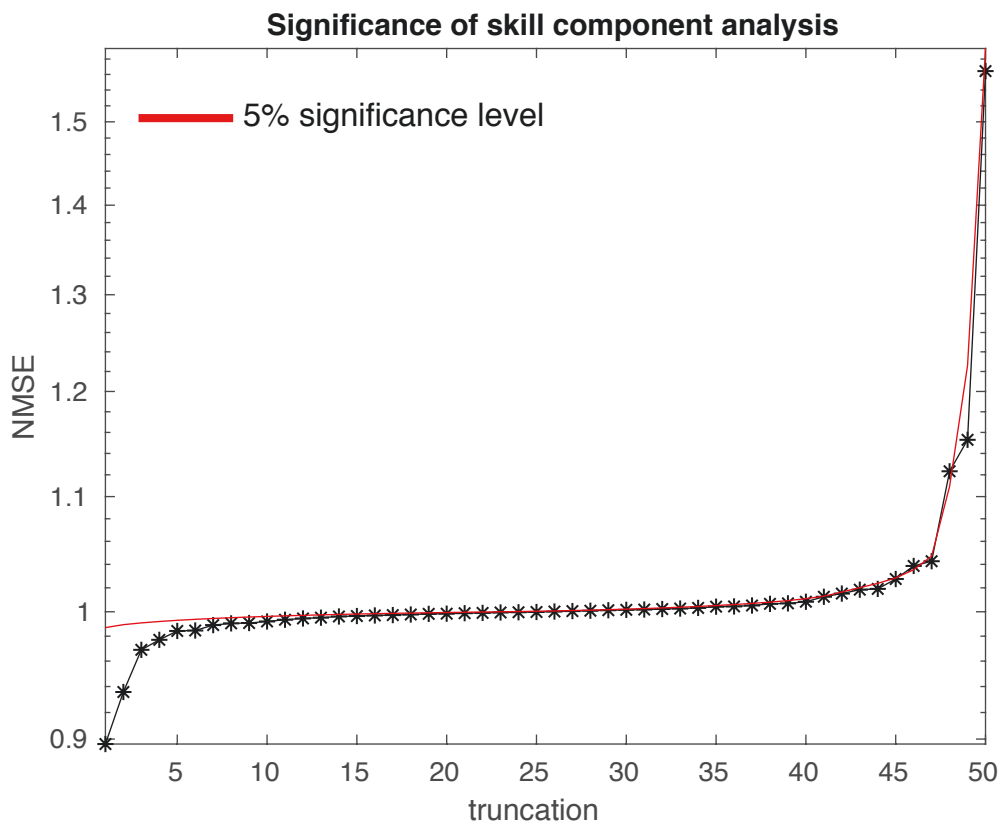


Figure 1. Multi-model NMSE (black) recovered from Skill Component Analysis and multi-model 5% significance level (red). Significance is estimated by the Monte Carlo method using 5000 iterations. Analysis is performed using independently sampled data (once per winter) over the entire multi-model record. A mode is considered significant if it is less than 1.

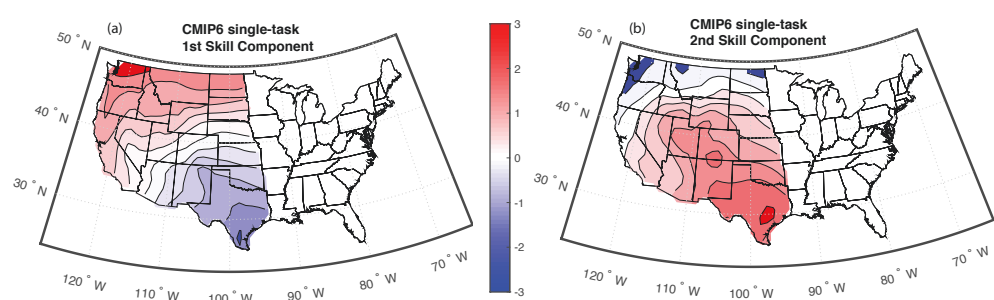


Figure 2. Patterns for the (a) 1st (a) and (b) 2nd leading skill components recovered from multi-model CMIP6 data. Predictions are made by the same CMIP6-single-task model.

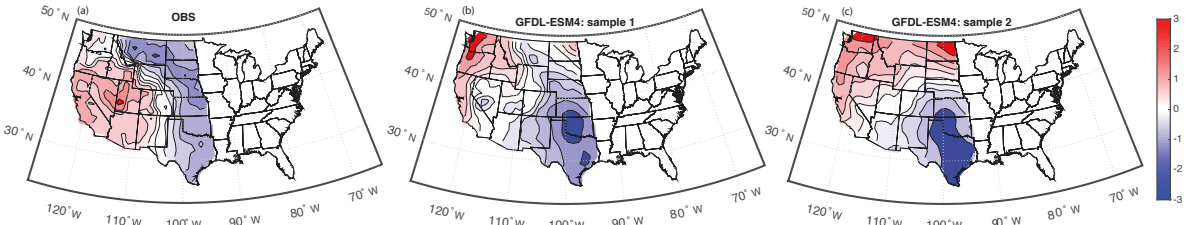


Figure 3. Patterns for the 1st skill component for (a) observations and two different randomly selected 19 year segments of data from the GFDL-ESM4 model (b) and (c).

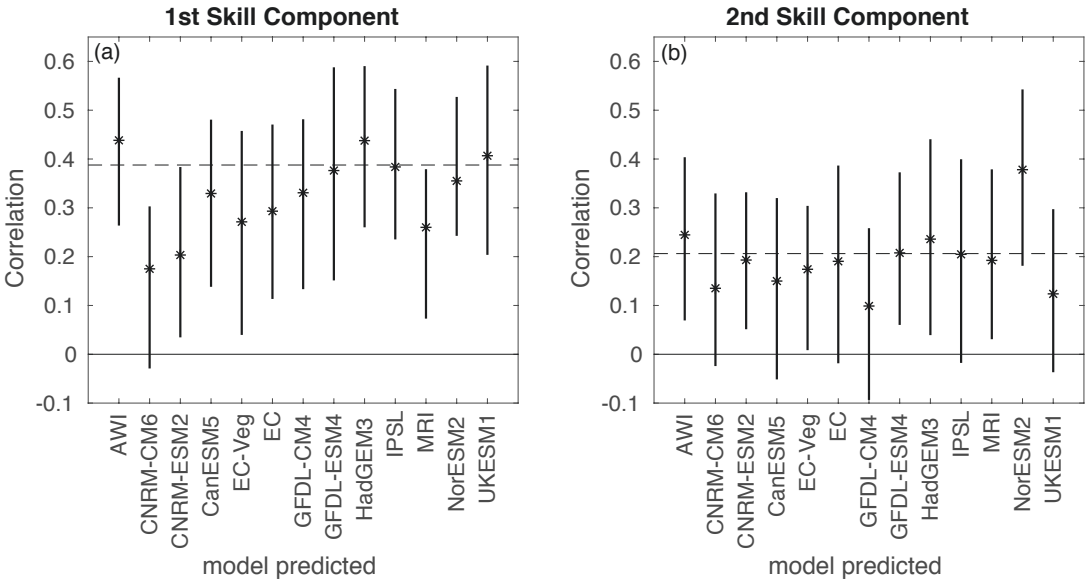


Figure 4. Correlation between the prediction and verification data for the (a) 1st and (b) 2nd leading skill components. These correlations are found by projecting both prediction and verification data onto the leading eigenvectors recovered from SCA and correlating the resulting time series. All predictions are made by the same CMIP6-single-task model. The black vertical bars show the 5th-95th percentile range of correlations for predictions within the specified CMIP6 model. The individual CMIP6 models are sampled to have the same number of years as observation (19 years). The black asterisk denotes the mean correlation. The correlation for predictions using observational data for the 2000-2018 are shown as the dashed line. The skill component analysis was performed using 50 Laplacian time series.

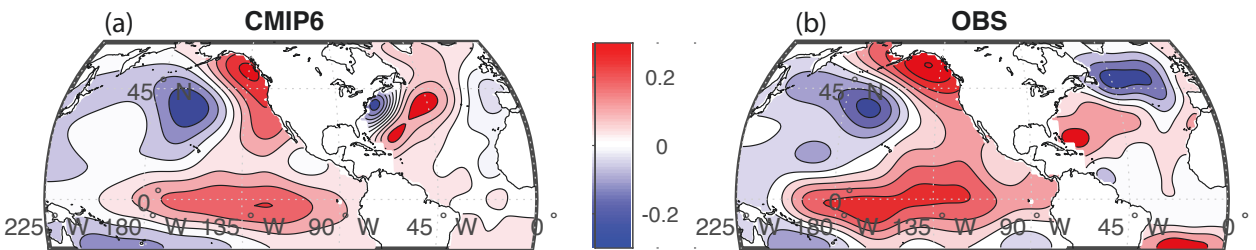


Figure 5. Regression of the leading skill component derived from the CMIP6 single-task models onto sea surface temperature from (a) multi-model CMIP6 data and (b) observations.