Supplementary-information for the paper: "S2S Reboot: An Argument for Greater Inclusion of Machine Learning in Subseasonal to Seasonal (S2S) Forecasts"

Judah Cohen, Dim Coumou Jessica Hwang, Lester Mackey,
Paulo Orenstein, Sonja Totz and Eli Tziperman

1 Cluster-based prediction

1.1 Introduction

The supplemental information contains a description of the clustering-based prediction approach, which is an improved version of that presented in S. Totz, E. Tziperman, D. Coumou, K. Pfeiffer, and J. Cohen, "Winter precipitation forecast in the european and mediterranean regions using cluster analysis." Geophys. Res. Lett., 44 (doi:10.1002/2017GL075674):12,41812,426, 2017. The corresponding code is available at https://www.seas.harvard.edu/climate/eli/Downloads/Clustering-based-prediction/European-temperature-2018b/

1.2 Cluster-based prediction methodology

Given the time series of the quantity to be predicted (predictand, e.g., anomaly winter (DJF) precipitation) and precursors (predictors, e.g., autumn (SON) sea ice cover and snow cover extent), we calculate the clusters of the predictand, and then use them to construct the prediction as described below. In order to obtain a cross-validated forecast, we choose one year to be predicted and then use all other years in order to build the prediction model. This is repeated for all years and the skill presented below is the average over all of these prediction calculations. For each predicted year, we first remove the mean of the precipitation using data from all years except for the predicted year.

Consider a forecast of precipitation anomaly time series at several locations, given by the predictand vector $\mathbf{prcp}(t)$. These precipitation data will be predicted using given precursors, e.g., time series of snow cover extent anomalies at several spatial locations given by the time-dependent vector $\mathbf{sce}(t)$, and time series of sea ice extent at several spatial locations, $\mathbf{sic}(t)$.

We assume that there are $N_{clusters}$ significant precipitation clusters. We use bold upper case variable names to denote clusters and composites, and lower case bold variable names to denote time series data. The prediction procedure requires the winter (DJF) precipitation clusters $\mathbf{PRCP}_i$, $i = 1, \ldots, N_{clusters}$ and the corresponding precursor composites (e.g., sea ice cover and snow cover extent anomalies from the autumn SON mean), $\mathbf{COMPOSITE}_i$. The clusters are calculated using hierarchical clustering of the winter precipitation anomaly data, while the composites for a given cluster i are calculated by averaging the predictors over all times in which the precipitation anomaly is assigned to its cluster i.

We also need a time series of the autumn-mean (averaged over SON) precursor anomaly (predictors) $\mathbf{precursor}_{SON}(t)$, for each spatial location. The time t denotes the year, where the precursors are evaluated during the fall (SON) and the precipitation of that year refers to the following DJF. For example, if the precursors are sea ice and snow cover, the vector of precursors (predictors) time series, and the vector of composites are calculated as follows.

First, we remove the mean of each precursor using all precursors data except the predicted year. Next, we normalize each precursor by the standard deviation. Finally, we combine different precursors into a single vector,

$$\begin{aligned}
\mathbf{sic}'_{SON}(t) &= \mathbf{sic}_{SON}(t) - \overline{\mathbf{sic}}_{SON} \\
\widehat{\mathbf{sic}}_{SON}(t) &= \mathbf{sic}'_{SON}(t)/\sigma_{\mathbf{sic}} \\
\mathbf{precursor}_{SON}(t) &= (\widehat{\mathbf{sic}}_{SON}(t), \widehat{\mathbf{sce}}_{SON}(t))^T
\end{aligned}$$

The variable $\overline{\mathbf{sic}}_{SON}$ is the time mean of the sea ice concentration using all times except the predicted year. The variable $\sigma_{\mathbf{sic}}$ is the standard deviation over all times and all grid points.

Then, we find the composites of the different autumn predictors by averaging the normalized predictors ($\widehat{\mathbf{sce}}(t)$,$\widehat{\mathbf{sic}}(t)$) over all autumn seasons (SON) for which the following winter precipitation anomaly is assigned to a given cluster. The predictors composites of the same cluster are combined into one composite

$$\mathbf{COMPOSITE}_{1,2} \;=\; (\mathbf{SIC}_{1,2}, \mathbf{SCE}_{1,2})^T$$

To obtain the prediction for the precipitation, we first find the projection of the current state of the predictors (snow cover and sea ice) on the $N_{clusters}$ predictor composites corresponding to the precipitation clusters.

Each predictor composite is associated with a precipitation cluster and provides information about the amplitude and spatial structure of winter precipitation expected given the autumn predictor composite. This allows us to calculate the expected precipitation pattern due to the projection of the current state of predictors on each cluster. Finally, we sum the contributions to the precipitation due to all clusters, to obtain the predicted total precipitation anomaly.

Mathematically, this proceeds as follows. To calculate the projection of $\mathbf{precursor}_{SON}(t)$ on the composite $\mathbf{COMPOSITE}_i$, we expand the current precursor state in terms of the precursor composites, to find the expansion coefficients, noting that the composites are not necessarily orthogonal. The expansion takes the form,

$$\mathbf{precursor}_{SON}(t) \approx \sum_{i=1}^{N_{clusters}} a_i(t)\ \mathbf{COMPOSITE}_i.$$

The expansion may only be approximate because the composites are not necessarily a complete set of vectors. To find the expansion coefficients $a_i(t)$, multiply by precursor composite $\mathbf{COMPOSITE}_j$, remembering that they are not necessarily orthogonal,

$$\mathbf{precursor}_{SON}(t) \cdot \mathbf{COMPOSITE}_j \;=\; \sum_{i=1}^{N_{clusters}} a_i(t)\ \mathbf{COMPOSITE}_i \cdot \mathbf{COMPOSITE}_j.$$

Next, we write this as a matrix equation for the unknown vector $\mathbf{a}(t)$ of coefficients $a_i(t)$. Define a matrix, $B_{ij} = \mathbf{COMPOSITE}_i \cdot \mathbf{COMPOSITE}_j$, and the right-hand side $\mathbf{\Gamma}_j(t) = \mathbf{precursor}_{SON}(t) \cdot \mathbf{COMPOSITE}_j$. This leads to the linear equations,

$$B\,\mathbf{a}(t) \;= \mathbf{\Gamma}(t),$$

that may be solved for the coefficients $a_i(t)$ at every time step (year t) in the data. Given that the matrix B may be ill conditioned, there may be many solutions for $\mathbf{a}(t)$. We choose the one with the smallest norm, using the SVD-based pseudo inverse such that singular values that are smaller than 1% of the largest singular value are set to zero (using python's pinv-function with the threshold set to 0.01).

The final expression for the predicted precipitation anomaly is obtained by summing the contribution of all clusters, each multiplied by the projection of the current state of precursors, $a(i)$,

$$\mathbf{prcp}(t) = \sum_{i=1}^{N_{clusters}} a_i(t)\ \mathbf{PRCP}_i. \tag{1}$$

2 Alternative Statistical Learning Approaches

kNN

Given the vector of features (e.g., lagged measurements, model forecasts, temporal characteristics, and geographic characteristics) associated with a target date and forecast region, a k-nearest neighbor (kNN) method would search for the historical dates and regions (neighbors) with features most similar to the target. The predicted weather pattern would then be a weighted average of the realized weather patterns associated with all neighbors. Such kNN approaches are especially popular in recommender systems (Bobadilla et al. 2013), where the algorithm is used to recommend items similar to items previously enjoyed by a customer or to recommend items enjoyed by customers similar to target customer. See Chapter 13 of Hastie et al. (2001) for more details.

Random forests

A decision tree is a prediction method that hierarchically partitions forecasting targets into homogeneous groups based on associated features (e.g., lagged measurements, location, and model forecasts) and forecasts the average historical weather pattern in each group. A random forest is an ensemble method that aggregates many different trees by averaging their predictions. To make the individual trees more diverse, the method uses only a randomly selected subset of features to create each partition. Random forests (Breiman 2001) and the closely related Bayesian additive regression tree method (Chipman et al. 2010) have led to state-of-the-art performance in a wide variety of prediction tasks including predicting disease progression in Lou Gehrigs disease patients (Kffner et al. 2015) and identifying breast lesions at high risk of cancer (Bahl et al. 2017). For more details and examples, see Chapter 8 of James et al. (2013).

Boosted decision trees

Boosting (Freund & Schapire 1997) is a learning method which sequentially combines lower-accuracy prediction rules, like decision trees, into a final higher accuracy ensemble. For boosted decision trees, we train a sequence of decision trees sequentially, each to correct the errors of the last: start by growing a tree to predict a target variable (e.g., future temperature on a given location), build a second tree to predict the mistakes made by the first tree, and then keep building trees to predict on the errors of the previous one. The aim is to improve prediction performance with the addition of each new tree. Boosting has delivered state-of-the-art performance for a variety of prediction tasks including Higgs Boson classification in high-energy physics and insurance claim classification (Chen & Guestrin 2016). More details can be found in Chapter 8 of James et al. (2013).

Gaussian processes

An extremely popular model in spatial statistics, Gaussian process regression, views the response variable (temperature, precipitation, etc.) as a smooth spatial surface. The smoothness of the surface is controlled by the covariance function of the Gaussian process, and spatial trends in the response variable are controlled by the mean function. The mean function can depend on additional features such as lagged measurements or other model forecasts. The result is a regression model that takes into account spatial dependencies. Gaussian processes have been used to forecast wind speed (Chen et al. 2014) and predict forest biomass (Banerjee et al. 2008). For an overview of the methodology, see Rasmussen & Williams (2006).

Neural networks

Neural networks are a highly flexible model class for relating a collection of inputs (e.g., lagged measurements or model forecasts at a set of locations) to a collection of outputs (e.g., temperature measurements at a set of locations). The inputs undergo a series of nonlinear transformations in the neural network's hidden layers; as the neural network is trained, the weights associated with these nonlinear transformations are learned in order to minimize prediction error. Neural networks, particularly deep networks with many hidden layers, have dramatically improved performance on a variety of learning tasks, including image recognition and machine translation (see, e.g., Deng & Yu 2014).

Causal effect networks

Causal discovery algorithms allow for interpretation of causal links between variables by determining whether we can say that, statistically, x provides more information about future values of y than past values of y alone. The causal effect network (CEN) aims to detect causal relationships amongst a set of time-series by iteratively testing the partial correlations conditioning on combinations of other time-series at different lags (Kretschmer et al. 2016). Thus, causal links in the CEN are those for which the linear relationship cannot be explained by the (combined) influence of other included indices or by auto-correlation. CEN is related to Granger-causality but allows for much stronger causal statements beyond, for example, the bi-variate only concept (Kretschmer et al. 2016).

Classically, one of the major limitations of statistical forecast models has been overfitting, which results in very high correlations to R-squared values on training data but the forecast fails on independent test data. Recent studies have introduced causal discovery algorithms to identify the causal precursors and remove those that arise from spurious correlations. This is an effective way to avoid overfitting problems and has resulted in robust statistical forecasts of polar vortex (PV) strength (Kretschmer et al. 2017) and Indian summer monsoon rainfall (Di Capua & Coumou 2017).