9th International Workshop on Climate Informatics
October 02-04, 2019

Hosted by École Normale Supérieure, Paris, France

GRAPH-GUIDED REGULARIZATION FOR IMPROVED SEASONAL FORECASTING

Abby Stevens[1], Rebecca Willett[1,1b], Antonios Mamalakis[2], Efi Foufoula-Georgiou[2,3], James Randerson[3], Padhraic Smyth[4], Stephen Wright[5] & Alejandro Tejedor[6]

***Abstract*—Understanding the factors that determine regional climate variability and change is a challenge with important implications for the economy, security, and environmental sustainability of many regions around the globe. Unprecedented quantities of high-resolution climate data provide an enormous opportunity to explore this question systematically and exhaustively. Simple, off-the-shelf machine learning and statistical analysis methods can yield misleading results when applied directly to such data. Standard model selection methods are fragile in the face of complex dependence structures in the climate system. This abstract describes a regression scheme that explicitly accounts for spatiotemporally correlated features via a regularization approach based on an underlying correlation graph. Using large ensemble climate outputs to estimate the strength of correlations among features, we form a graph with edge weights corresponding to pairwise correlations. This graph is used to define a graph total variation regularizer that promotes similar weights for highly correlated features. We apply our scheme to predicting winter precipitation totals in the southwestern US using sea surface temperatures (SST) over the entire Pacific basin at multiple time lags, and demonstrate that our method provides strong predictive performance.**

I. MOTIVATION

The growing quantities of high-resolution Earth observations and climate model output [1] provide an opportunity to discover previously unknown teleconnections (long-range connections among climate modes) with strong predictive potential to improve seasonal-to-subseasonal forecasting. However, many statistical prediction schemes which aim to exploit established climate teleconnections between large-scale modes of variability (e.g., the El Niño-Southern Oscillation, the Pacific North America pattern, the Madden-Julian Oscillation, etc.) and regional hydroclimate fail to capture the highly complex and nonstationary nature of the climate system. On the other hand, dynamical models show limited predictive skill at lead times longer than two weeks [2], due to imperfect physical conceptualizations and inaccurate initial conditions.

[1]Department of Statistics, University of Chicago, [1b] Department of Computer Science, University of Chicago, [2]Department of Civil and Environmental Engineering, University of California, Irvine, [3]Department of Earth System Science, University of California, Irvine, [4]Department of Computer Science, University of California, Irvine, [5]Computer Sciences Department, University of Wisconsin, Madison, [6]Max Planck Institute for the Physics of Complex Systems, Dresden Germany

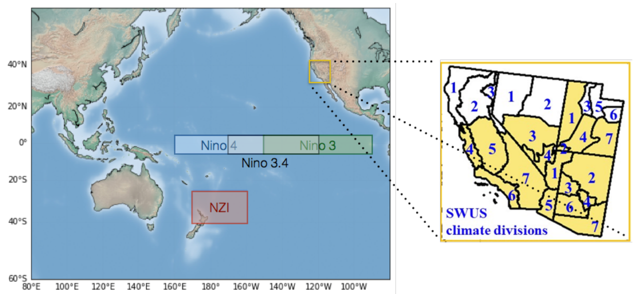

Fig. 1. The Pacific basin with known predictive regions highlighted and the specific climate regions over the expanded SWUS.

Recently, a new teleconnection between sub-tropical sea surface temperatures off the coast of New Zealand (NZI) and regional precipitation in the Southwestern US (SWUS) (Figure 1) was discovered, exhibiting stronger and earlier predictive potential than any other known mode of variability, including the El Niño Southern Oscillation (ENSO), which has long been used for SWUS seasonal precipitation forecasting [3]. A natural question to ask is whether there are additional, undiscovered teleconnections that, once identified, could improve seasonal forecasting.

Rather than relying on ad hoc methods for discovering important teleconnections, we seek to cast the forecasting problem as a regression problem in which modes (influential climate patterns) are not specified in advance but rather are allowed to emerge from the data as sources of predictability. Such a method must account for small sample sizes and high dimensionality, strong spatiotemporal dependencies among the predictors, and the need for interpretability in the climate sciences. Regularized regression has shown promise in accomplishing this task ([4], [5], [6], [7]).

However, a key challenge is that the covariates or features of such models (i.e., SSTs over space and time) are highly correlated, violating key assumptions underlying many modern high-dimensional regression methods such as the Lasso. Simultaneously, we seek to leverage data generated via simulation of physics-based climate models in addition to observational data. Treating simulation data as additional samples of observational data fails to account for model errors and sampling bias associated with simulations [8]. This abstract describes an approach in which we perform regularization over a graph corresponding to the correlation structure underlying the data. This approach (a) is provably robust to strong correlations among features or covariates and (b) leverages climate model simulations by using them to set parameters of the regularizer.

II. Problem Formulation

The SWUS region is highly vulnerable to drought, which has severe economic and ecological implications. Accurate and early prediction in this region is therefore of particular interest. The largest amount of yearly precipitation occurs during winter (November-March), and there is high inter-annual variability. The SWUS is divided into 16 climate divisions of interest [3], and we are able to extract winter averages from each region for 1940-2015 [1].

Given known teleconnections that are the current state-of-the-art for seasonal forecasting [3], we use as our predictors sea-surface temperature (SST) at various time lags across the Pacific basin. Specifically, we consider mean SST on a $10° \times 10°$ grid in July, August, September, and October. The data are from the 20th Century Reanalysis project [2]. We consider 226 locations over the Pacific at 4 different time lags, for a total of $p = 904$ features and $n = 75$ years of observations.

For a given year i and climate division r, we seek to solve the regression problem

$$y_r^{(i)} = \sum_{j=1}^{p} X_j^{(i)} \beta_j + \epsilon^{(i)}$$

where $X_j^{(i)}$ is the summer SST measurement at (location, time lag) j preceding $y_r^{(i)}$, the winter precipitation observation for region r. We also assume $\epsilon^{(i)} \sim N(0, \sigma^2)$. Although precipitation is non-negative and tends to be skewed in distribution, we find experimentally that we can reasonably approximate monthly winter totals over the SWUS regions with Gaussian noise. X and y are both centered and normalized to have unit variance and mean zero. We know that the columns of X are highly correlated in both space and time. Our goal is to estimate coefficients β that yield low prediction error for the seasonal forecasting problem and are physically interpretable from a climate science perspective.

[1] https://www.ncdc.noaa.gov/cag/time-series/us

[2] https://www.esrl.noaa.gov/psd/data/20thC_Rean/

III. Methods

A. Graph total variation

For a response $y \in \mathbb{R}^n$ and covariates $X \in \mathbb{R}^{n \times p}$, $p \gg n$, we seek to estimate β^* such that $y = X\beta^* + \epsilon$ where β^* is well-aligned with the correlation structure of X. For a zero-centered X, let $\Sigma := E(X^T X)$ be the covariance matrix of X and $\hat{\Sigma}$ be an estimate of Σ. Let $\hat{s}_{j,k} := \text{sign}(\hat{\Sigma}_{j,k})$. Our estimator, which we call *graph total variation* (GTV) [9], is given by

$$\begin{aligned} \hat{\beta} \quad &= \quad \arg\min_{\beta} \|y - X\beta\|_2^2 \\ &+ \quad \lambda_{TV} \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k}\beta_k| \\ &+ \quad \lambda_1 \|\beta\|_1, \end{aligned} \tag{1}$$

where λ_1 and λ_{TV} are regularization parameters chosen through cross validation. We can interpret this estimator from a graph perspective by defining a *covariance graph* based on $\hat{\Sigma}$. Let $G = (V, E, W)$ be an undirected weighted graph with vertices $V = \{1, 2, \ldots, p\}$, edges $E := \{(j,k) : |\hat{\Sigma}_{j,k}| > 0, j \neq k\}$, and weight matrix W with $W_{j,k} = w_{j,k} = |\hat{\Sigma}_{j,k}|^{1/2}$. Let $\Gamma \in \mathbb{R}^{|E| \times p}$ be the *weighted edge incidence matrix*, of G, where each row ℓ represents a pair of connected vertices (j_ℓ, k_ℓ):

$$\begin{aligned} \Gamma_{\ell, j_\ell} &= |\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} \\ \Gamma_{\ell, k_\ell} &= -\text{sign}(\hat{\Sigma}_{j_\ell, k_\ell}) |\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} \end{aligned} \tag{2}$$

We can thus simplify the total variation term in (1) as

$$|\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k}\beta_k| = \|\Gamma\beta\|_1$$

It is worth highlighting the differences between GTV and similar well-studied structured estimators, such as the fused Lasso [10] and the generalized Lasso [11]. The theoretical guarantees of these estimators assume that X satisfies the restricted eigenvalue condition [12], which is often violated when the columns of X are highly correlated. GTV, on the other hand, performs well in the presence of strong correlations.

GTV promotes estimates of β that contain sparse clusters of coefficients corresponding to highly correlated variables. That is, the stronger the correlation between X_j and X_k, the more similar $\hat{\beta}_j$ and $\hat{\beta}_k$. In the case of perfect correlation (i.e. $X_j = X_k$), Lasso will assign weight to either X_j or X_k, while GTV will distribute the weight evenly across X_j and X_k. This results in more interpretable model selection, which is of interest in climate and other application areas. This estimator adaptively selects clusters of features aligned with an estimated underlying graphical structure.

B. Covariance matrix estimation

The GTV regularization term depends on a reliable estimate of Σ. It is well documented ([13], [14]), that, in high-dimensional settings where $p \gg n$, the sample covariance $\hat{\Sigma}_S = \frac{1}{n} X X^T$ is not a consistent or accurate estimate of Σ. In some application areas, including climate science, there exists side information that is not based on X with which we can estimate Σ.

Climate models are physical mathematical models that simulate how energy and matter interact. In climate and related domains, there is hope that leveraging climate models alongside observations can help reduce uncertainties in predictive schemes [15]. One way to do this is data augmentation, which treats the simulation data as independent and identically distributed draws from the distribution of the observed data and combines all data to feed to the model, but this can be problematic in light of model biases and sensitivities to initialization [8].

We propose treating these climate simulations as side information we can use to estimate Σ. We use simulations from the CESM Large Ensemble Project, known as LENS[1]. LENS is a 40-member ensemble of Community Earth System Model V1 (CESM1) simulations, each of which is subject to the same radiative forcing scenario but with slightly perturbed initial temperature conditions. We linearly interpolate the LENS simulations of summer SST onto the same spatial grid as our observations.

Letting $X_L \in \mathbb{R}^{40n \times p}$ be the centered matrix of stacked features from all LENS ensemble members, we let $\hat{\Sigma}_L$ be the sample covariance matrix of X_L. We are assuming that both X and X_L are draws from a distribution with the same covariance matrix, Σ, and since X_L has a much higher sample size than X, we believe $\hat{\Sigma}_L$ is a better estimate of Σ than $\hat{\Sigma}_S$.

[1] http://www.cesm.ucar.edu/projects/community-projects/LENS/

C. Multitask GTV

We know that precipitation patterns across the entire SWUS are tied to similar summer atmospheric events, but possibly to a different degree for each region. Because of this, we seek to simultaneously solve m regression problems, a technique known as *multitask learning*. We assume that there is an unknown subset of covariates that are relevant for prediction, and this subset is preserved across the m regions. Let $Y = [y^{(1)}, y^{(2)}, \dots y^{(m)}] \in \mathbb{R}^{n \times m}$ be the matrix of the m response vectors corresponding to each climate region shown in Figure 1 and $B = [\beta^{(1)}, \beta^{(2)}, \dots \beta^{(m)}] \in \mathbb{R}^{p \times m}$ be the matrix of the corresponding m coefficient vectors. We wish to solve the following objective function, which we refer to as *MultiGTV*:

$$\begin{aligned} \hat{B} = \arg\min_B & \|Y - XB\|_F^2 \\ & + \lambda_1 \sum_{r=1}^{m} \left(\|\Gamma \beta^{(r)}\|_1 + \|\beta^{(r)}\|_1 \right) + \lambda_2 \sum_{j=1}^{p} \|B_{j,:}\|_2 \end{aligned} \tag{3}$$

where Γ is the edge-incidence matrix from (2) and $B_{j,:}$ is the j^{th} row of B. The first regularization term encourages coefficient estimates for each region that are sparse and well-aligned with the covariance of X, while the second promotes similarity in the support of the coefficient vectors across regions. We use a variant of an ADMM algorithm (Alternating Direction Method of Multipliers) [16] to solve this objective function.

IV. EVALUATION

For all experiments, the models are trained on the first 50 years of data and the remaining 25 years are held out for testing. Regularization parameters are chosen through 5-fold cross validation on the training data, and all reported errors are computed on the test data.

We compare the performance of GTV and MultiGTV with other well-known structured regression methods. Because GTV estimates coefficients for a single response and MultiGTV estimates many responses simultaneously, in order to compare performance in a meaningful way we report errors as follows. For a given region, we compute the mean squared error (MSE) on the corresponding test data. Then, we compute and report the area-weighted means and standard errors (SE) of the MSE across all regions in the SWUS.

We benchmark our methods against ordinary least squares (OLS) and Lasso (along with the multitask version of Lasso) in order to set a baseline. Then, we consider the following graph-based methods, which all solve a form of the GTV objective function with the edge-incidence matrix Γ defined by a variety of graphs:

1) Fused Lasso: The graph includes only immediate spatial and temporal neighbors
2) GTV with $\hat{\Sigma}_S$: The graph includes edges and weights based on the covariance of the SST observations, $\hat{\Sigma}_S$
3) GTV with $\hat{\Sigma}_L$: The graph includes edges and weights according to the covariance of the LENS SST simulations, $\hat{\Sigma}_L$

Each of these graphs can be used in the MultiGTV method as well. The results of all methods under consideration are shown in Table I.

Method	**MSE**	**SE**
OLS	1.018	0.038
Lasso	1.014	0.037
Fused Lasso	1.069	0.062
GTV with $\hat{\Sigma}_S$	0.989	0.052
GTV with $\hat{\Sigma}_L$	**0.949**	0.041
Multitask Lasso	0.964	0.029
Multitask Fused Lasso	0.929	0.036
MultiGTV with $\hat{\Sigma}_S$	0.921	0.035
MultiGTV with $\hat{\Sigma}_L$	**0.919**	0.027

TABLE I
AREA-WEIGHTED MEAN AND STANDARD ERRORS OF THE MSE FOR EACH OF THE 16 CLIMATE REGIONS UNDER CONSIDERATION. THE TOP SECTION INCLUDES METHODS THAT ESTIMATE COEFFICIENTS FOR EACH REGION SEPARATELY AND THE BOTTOM INCLUDES METHODS THAT ESTIMATE ALL REGIONS SIMULTANEOUSLY. THE BEST PERFORMING METHOD IN EACH SECTION IS IN BOLD.

We see that GTV outperforms the other methods in both the multitask and regular cases. Additionally, we see that estimating the covariance graph using the LENS data ($\hat{\Sigma}_L$) provides stronger predictive performance than using just the observations ($\hat{\Sigma}_S$). The multitask methods outperform their single-response counterpart, but it is worth noting that MultiGTV with both $\hat{\Sigma}_S$ and $\hat{\Sigma}_L$ result in nearly identical predictive performances, suggesting that the improvement in performance when using the LENS simulations seen in the ordinary GTV setting is not as influential in the multitask setting.

V. DISCUSSION

In this abstract we argue that the seasonal forecasting problem is improved by the use of graph-based regularization methods that explicitly account for spatial and temporal correlations among the features. We also present a novel method of leveraging large ensemble climate models to estimate the covariance graph for use in GTV, which results in the highest predictive performance of the methods considered. The intuition behind this discovery is that there are long-range teleconnections among climate variables that extend beyond nearest-neighbors approaches like fused Lasso, and accounting for these relationships is important for the forecasting problem.

This work lays the methodological foundation for a data-driven approach to seasonal forecasting that is grounded in physics. There are many potential next steps for this research. First, we note that the method as presented does not account for the nonlinearity of the system dynamics. We acknowledge that external forcing like anthropogenic climate change and/or natural multidecadal oscillations in the Pacific can affect the predictive skill of the algorithm. One of the next steps of this research is to account for changes in the weights of the predictors as a function of time.

Next steps also include a rigorous investigation of the locations and time lags at which Pacific SSTs are selected by our model to be predictive of SWUS precipitation. This analysis will address the robustness of the selected variables to help determine whether or not they correspond to true teleconnections from both a statistical and physical standpoint.

Finally, we hope to further validate our methods on different seasonal forecasting problems. Our methods are flexible and can easily adapt to different climate variables and prediction settings.

ACKNOWLEDGMENTS

Thanks to the following grants for funding this work: NSF DMS-1930049, NSF OAC-1934637, NSF DMS-18393366, ECCS-1839441 and DOE DE-AC02-06CH11357.

REFERENCES

[1] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," *Science*, vol. 331, pp. 700–702, 2011.
[2] N. A. of Sciences, "Strategies for subseasonal to seasonal forecasts," *The National Academies Press*, 2016.
[3] A. Mamalakis, J.-Y. Yu, J. T. Randerson, A. AghaKouchak, and E. Foufoula-Georgiou, "A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern us," *Nature Communications*, vol. 9, 2018.
[4] A. R. Goncalves, A. Banerjee, V. Sivakumar, , and S. Chatterjee, "Structured estimation in high dimensions: applications in climate," in *Large-scale machine learning in the earth sciences*, ch. 2, pp. 13–32, CRC Press, 2017.

[5] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, "Sparse group lasso: Consistency and climate applications," *Proceedings of the 2012 SIAM International Conference on Data Mining*, 2012.

[6] S. He, X. Li, V. Sivakumar, and A. Banerjee, "Interpretable predictive modeling for climate variables with weighted lasso," *AAAI Conference on Artificial Intelligence*, 2019.

[7] T. DelSole and A. Banerjee, "Statistical seasonal prediction based on regularized regression," *Journal of Climate*, vol. 30, no. 4, 2017.

[8] K. A. Mckinnon, A. Poppick, E. Dunn-Sigouin, and C. Deser, "An observational large ensemble to compare observed and modeled temperature trend uncertainty due to internal variability," *Journal of Climate*, vol. 52, no. 19, 2017.

[9] Y. Li, B. Mark, G. Rasutti, and R. Willett, "Graph-based regularization for regression problems with highly-correlated designs," *arXiv preprint arXiv:1410.5093*, 2018.

[10] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 91–108, 2005.

[11] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[12] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[13] T. T. Cai, Z. Ren, and H. H. Zhou, "Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1–59, 2014.

[14] A. Maurya, "A well-conditioned and sparse estimation of covariance and inverse covariance matrices using a joint penalty," *Journal of Machine Learning Research*, vol. 17, pp. 1–28, 2016.

[15] T. Schneider, S. Lan, A. Stuart, and J. Teixeira, "Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations," *Geophysical Research Letters*, vol. 44, pp. 12,396–12,417, 2017.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, 2011.