Algorithms for Estimating Trends in Global Temperature Volatility

Author 1 and Author 2
Address line
Address line

Abstract

Trends in terrestrial temperature variability are perhaps more relavant for species viability than trends in mean temperature. In this paper, we develop methodology for estimating such trends using multi-resolution climate data from polar orbiting weather satellites. We derive two novel algorithms for computation that are tailored for dense, gridded observations over both space and time. We evaluate our methods with a simulation constructed to mimic these data's features and on a large, publicly available, global temperature dataset with the goal of tracking trends in cloud reflectance temperature variability.

1 Introduction

How is this first part related to the rest of the paper?

The amount of sunlight reflected from clouds is among the largest sources of uncertainty in climate prediction (Boucher et al. 2013). But climate models fail to reproduce global cloud statistics and understanding the reasons for this failure is a grand challenge of the World Climate Research Programme (Bony et al. 2015). By understanding how cloudliness changes over time, examining changes in brightness temperature within various cloud types, and inferring changes in cloud radiative effects, atmospheric scientists can better model climate change. Modeling inter-satellite biases, satellite drift, and seasonal and long-term climate phenomena like El Niño will lead to a better understanding of climate change (Schreier et al. 2014; Baum et al. 1994; 1992; Frey, Ackerman, and Soden 1996).

Numerous studies have examined the overall impacts of clouds on climate variability (Myers, Mechoso, and DeFlorio 2018; Grise et al. 2013; Bender, Ramanathan, and Tselioudis 2012), but such investigations have been hampered by the lack of a suitable dataset. Ideal data would have global coverage at high spatial resolution, a long enough record to recover temporal trends, and be multispectral (Wielicki et al. 2013). The International Satellite Cloud Climatology Project (ISCCP) has been producing cloud property information from limited spectral channels for over three decades (Rossow and Schiffer 1991). These data have been used in various climate variability studies (Bender, Ramanathan, and Tselioudis 2012, e.g.), however, the utility of these data for long-term study of climate variability has been questioned (Evan, Heidinger, and Vimont 2007).

ISCCP provides a relatively long record, but it only incorporates radiances from limited visible and infrared channels from Advanced Very High Resolution Radiometer (AVHRR) instruments. Ongoing work seeks to create a more spectrally-detailed dataset which can avoid the above issues by combining radiance data from AVHRR imagers with readings from High-resolution Infrared Radiation Sounders on board legacy satellites (Staten et al. 2016; Schreier et al. 2010; Kahn et al. 2007). In anticipation of this new dataset, our work develops novel methodology for examining the trends in variability of climate data across space and time.

1.1 Variability Rather Than Average

Trends in terrestrial temperature variability are perhaps more relavant for species viability than trends in mean temperature (Huntingford et al. 2013), because an increase in temperature variability will increase the probability of extreme hot outliers (Vasseur et al. 2014). Recent climate literature suggests that it is more difficult for society to adapt to these extremes than to the gradual increase in the mean temperature (Hansen, Sato, and Ruedy 2012; Huntingford et al. 2013). Furthermore, the willingness of popular media to emphasize the prevalence extreme cold events coupled with a fundamental misunderstanding of the relationship between climate (the global distribution of weather over the long run) and weather (observed short-term, localized behavior) leads to public misunderstanding of climate change. In fact, a point of active debate is the extent to which the observed increased frequency of extreme cold events in the northern hemisphere can be attributed to increases in temperature variance rather than to increases in mean climate (Screen 2014; Fischer, Beyerle, and Knutti 2013; Trenberth et al. 2014).

Nevertheless, research examining trends in the volatility of spatio-temporal climate data is scarce. Hansen, Sato, and Ruedy (2012) studied the change in the standard deviation (SD) of the surface temperature in the NASA Goddard Institute for Space Studies gridded temperature data set by examining the empirical SD at each spatial location relative to that location's SD over a base period, and showed that these estimates are increasing. Huntingford et al. (2013) took a similar approach in analyzing the ERA-40 data set. Their

results showed that there still is an increase in the SDs from 1958-1970 to 1991-2001, but this is much less than what is obtained from the method used in (Hansen, Sato, and Ruedy 2012). The authors also computed the time-evolving global SD from the de-trended time-series at each position, which suggests that the global SD has been stable.

These and other related research, e.g., Rhines and Huybers (2013), have several shortcomings. First, no statistical analysis has been performed to examine if the changes in the SD are statistically significant. Second, the methodologies for computing the SDs are highly sensitive to the choice of base period. Third, and most importantly, temporal and spatial correlations between the observations are completely ignored.

In the present work, we examine variance (rather than the mean) for a number of reasons. First, instrument bias in the satellites increases over time so examining the mean over time conflates that bias with any actual change in mean (though the variance is unaffected). Second, extreme weather events (hurricanes, droughts, wildfires in California, heatwaves in Europe) may be driven more strongly by increases in variance than by increases in mean. Finally, even if the global mean temperature is constant, there may still be climate change. In fact, atmospheric physics suggests that, across space, average temperatures should not change (extreme cold in one location is offset by heat in another). But if swings across space are becoming more rapid, then, even with no change in mean global temperature over time, increasing variance can lead to increases in the prevalence of extreme events.

1.2 Main Contributions

The main contribution of this work is to develop a new methodology for detecting the trend in the volatility of spatio-temporal data. In this methodology, the variance at each position and time is considered as a hidden variable. The values of these hidden variables are then estimated by maximizing the likelihood of the observed data. Following (Tibshirani 2014), we penalize the differences between the estimated variances which are temporally and spatially "close", resulting in a generalized LASSO problem. However, in our application, the dimension of this optimization problem is massive, and so the standard solvers are inadequate. We develop two algorithms which are computationally feasible. In the first method, we adopt an optimization technique called alternating direction method of multipliers (ADMM, Boyd et al. 2011), to divide the total problem into several sub-problems of much lower dimension and show how the total problem can be solved by iteratively solving these sub-problems. The second method, called the *linearized ADMM algorithm* (Parikh and Boyd 2014) solves the main problem by iteratively solving a linearized version of it. We will compare the benefits of each method.

Our main contributions are as follows:

1. We propose a model for nonparametric variance estimation for a spatio-temporal process and discuss the relationship between our methods and those existing in the machine learning literature (Section 2).
2. We derive two alternating direction method of multiplier algorithms (ADMM) to fit our estimator when applied to very large data (Section 3). We give situations under which each algorithm is most likely to be useful. Open-source Python code is available on `github`.
3. We illustrate our methods on a large, publicly available, global temperature dataset with the goal of tracking worldwide trends in variance as well as a simulation constructed to mimic these data's features (Section 4).

While the motivation for our methodology is its application to large, gridded climate data, we note that our algorithms are also applicable to neuroimaging, image denoising, or examining large collections of financial instruments.

2 Smooth Spatio-temporal Variance Estimation

Kim et al. (2009) proposed ℓ_1-trend filtering as a method for estimating a smooth, time-varying trend. It is formulated as the optimization problem $\min_\beta \frac{1}{2}\sum_{t=1}^{T}(y_t-\beta_t)^2 + \lambda\sum_{t=1}^{T-2}|\beta_t - 2\beta_{t+1}+\beta_{t+2}|$ or equivalently:

$$\min_{\beta} \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|D_t\beta\|_1 \tag{1}$$

where $y=\{y_t\}_{t=1}^{T}$ is an observed time-series, $\beta\in\mathbb{R}^T$ is the smooth trend, D_t is a $(T-2)\times T$ matrix, and λ is a tuning parameter which balances fidelity to the data (small errors in the first term) with a desire for smoothness. In Appendix A, we specify the form of the matrix D_t which results in piecewise linear estimated β. Kim et al. (2009) proposed a specialized primal-dual interior point (PDIP) algorithm for solving (1). From a statistical perspective, (1) can be viewed as a constrained maximum likelihood problem with independent observations from a normal distribution with common variance, $y_t\sim \mathrm{N}(\beta_t,\sigma^2)$, subject to a piecewise linear constraint on β. Alternatively, solutions to (1) are equivalent to maximum a posteriori Bayesian estimators based on Gaussian likelihood with a special Laplace prior distribution on β. Note that the structure of the estimator is determined by the penalty function $\lambda\|D_t\beta\|_1$ rather than any parametric trend assumptions—autoregressive, moving average, sinusoidal seasonal component, etc. The resulting trend is therefore essentially nonparametric in the same way that splines are. In fact, using squared ℓ_2 norm as the penalty instead of ℓ_1 results exactly in regression splines.

2.1 Modifications for Variance

Inspired by the ℓ_1-trend filtering algorithm, we propose a non-parametric model for estimating the variance of a time-series. To this end, we assume that at each time step t, there is a parameter h_t such that the observations y_t are independent normal variables with zero mean and variance $\exp(h_t)$. The negative log-likelihood of the observed data in this model is $l(y\mid h)\propto -\sum_{t=1}^{T} h_t - y_t^2 e^{-h_t}$. Crucially, we assume that the parameters h_t vary smoothly. To impose this assumption, we estimate h_t by solving the penalized, negative log-likelihood:

$$\min_{h} -l(y\mid h) + \lambda\|D_t h\|_1 \tag{2}$$

where D_t has the same structure as above.

As with (1), one can solve (2) using the PDIP algorithm (as in, e.g., `cvxopt`, Andersen, Dahl, and Vandenberghe 2013). In each iteration of PDIP we need to compute a search direction by taking a Newton step on a system of nonlinear equations. Due to space limitations, we defer details to Appendix A in the Supplement, where we show how to derive the dual of this optimization problem and compute the first and second derivatives of the dual objective function.

2.2 Adding Spatial Constraints

The method in the previous section can be used to estimate the variance of a single time-series. In this section, we extend this method to the estimation of the variance of spatio-temporal data.

At a specific time t, the data are measured on a grid of points with n_r rows and n_c columns for a total of $S = n_r \times n_c$ spatial locations. Let y_{ijt} denote the value of the observation at time t on the i^{th} row and j^{th} column of the grid, and h_{ijt} denote the corresponding hidden variable. We seek to impose both temporal and spatial smoothness constraints on the hidden variables. Specifically, we seek a solution for h which is piecewise linear in time and piecewise constant in space (although higher-order smoothness can be imposed with minimal alterations to the methodology). We achieve this goal by solving the following optimization problem:

$$\min_h \sum_{i,j,t} h_{ijt} + y_{ijt}^2 e^{-h_{ijt}} + \lambda_t \sum_{i,j} \sum_{t=1}^{T-2} |h_{ijt} - 2h_{ij(t+1)} + h_{ij(t+2)}| + \lambda_s \sum_{t,j} \sum_{i=1}^{n_r-1} |h_{ijt} - h_{(i+1)jt}| + \lambda_s \sum_{t,i} \sum_{j=1}^{n_c-1} |h_{ijt} - h_{i(j+1)t}| \tag{3}$$

The first term in the objective is proportional to the negative log-likelihood, the second is the temporal penalty for the time-series at each location (i, j), while the third and fourth, penalize the difference between the estimated variance of two vertically and horizontally adjacent points, respectively. The spatial component of this penalty is a special case of trend filtering on graphs (Wang et al. 2016) which penalizes the difference between the estimated values of the signal on the connected nodes (though the likelihood is different). As before, we can write (3) in matrix form where h is an $T \times S$ vector and D_t is replaced by $D \in \mathbb{R}^{(N_t+N_s)\times(T\cdot S)}$, where $N_t = S \cdot (T-2)$ and $N_s = T \cdot (2n_r n_c - n_r)$ are the number of temporal and spatial constraints, respectively[1]. The exact form of this matrix in clarified in Appendix A in the Supplement. Then, as we have two different tuning parameters for the temporal and spatial components, we write $\Lambda = \left[\lambda_1 \mathbf{1}_{N_t}^\top,\ \lambda_2 \mathbf{1}_{N_s}^\top\right]^\top$ leading to:[2]

$$\min_h -l(y \mid h) + \Lambda^\top |Dh|. \tag{4}$$

[1] N_s is obtained by counting the number of unique constraints at each location and at all times.

[2] Throughout the paper, we use $|x|$ for both scalars and vectors. For vectors we use this to denote a vector obtained by taking the absolute value of each entry of x.

2.3 Related Work

TODO: I've moved this here rather than above to focus the intro more on the data.

Variance estimation for financial time series has a lengthy history, focused especially on parametric models like the generalized autoregressive conditional heteroskedasticity (GARCH) process (Engle 2002) and stochastic volatility models (Harvey, Ruiz, and Shephard 1994). These models (and related AR processes) are specifically for parametric modelling of short "bursts" of high volatility, behavior typical of financial instruments. Parametric models for spatial data go back at least to (Besag 1974) who proposed a conditional probability model on the lattice for examining plant ecology.

More recently, nonparametric models for both spatial and temporal data have focused on using ℓ_1-regularization for trend estimation. Kim et al. (2009) proposed ℓ_1-trend filtering for univariate time series, which forms the basis of our methods. These methods have been generalized to higher order temporal smoothness (Tibshirani 2014), graph dependencies (Wang et al. 2016), and, most recently, small, time-varying graphs (Hallac et al. 2017).

Our methodology is similar in flavor to (Hallac et al. 2017) or related work in (Gibberd and Nelson 2017; Monti et al. 2014), but with several fundamental differences. These papers aim to discover the time-varying structure of a network. To achieve this goal, they use Gaussian likelihood with unknown precision matrix and introduce penalty terms which (1) encourage sparsity among the off-diagonal elements and (2) discourage changes in the estimated inverse covariance matrix from one time-step to the next. Our goal in the present work is to detect the temporal trend in the variance of each point in the network, but the network is known (corresponding to the grid over the earth). The variance of each point, however changes. To modify the objective function in (Hallac et al. 2017, Eq. 2), we would enforce complete sparsity on the off-diagonal elements (since they are not estimated) and add a new penalty to enforce spatial behavior across the diagonal elements. Thus, (4) is not simply a special case of these existing methods.

3 Optimization Methods

For a spatial grid of size S and T time steps, D in equation 4 will have $3Tn_rn_c - 2n_rn_c - Tn_r$ rows and $S \cdot T$ columns. For a $1^\circ \times 1^\circ$ grid over the entire northern hemisphere and daily data over 10 years, we have $S = 90 \times 360 \approx 32,000$ spatial locations and $T = 3650$ time points, and so D has approximately 10^8 columns and 10^8 rows. In each step of the PDIP algorithm, we need to solve a linear system of equations which depends on $D^\top D$ (see appendix A and B). Therefore, applying the PDIP directly is infeasible for our data.[3]

In the next section, we develop two ADMM algorithms for solving this problem efficiently. The first casts the problem as a so-called consensus optimization problem (Boyd

[3] We note that D is a highly structured, sparse matrix, but, unlike trend filtering alone, it is not banded. We are unaware of general linear algebra techniques for inverting such matrix, despite our best efforts.

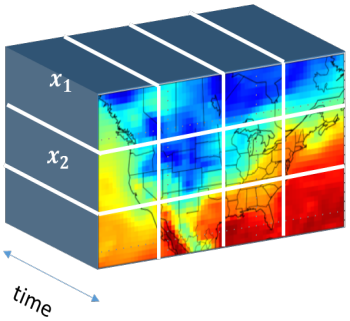

Figure 1: The cube represents the global variable h in space and time. The sub-cubes specified by the white lines are x_i.

et al. 2011) which solves smaller sub-problems using PDIP and then recombines the results. The second uses proximal methods to avoid matrix inversions.

3.1 Consensus Optimization

Consider an optimization problem of the form $\min_h f(h)$, where $h \in \mathbb{R}^n$ is the *global variable* and $f(h) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex. Consensus optimization breaks this problem into several smaller sub-problems that can be solved independently in each iteration.

Assume it is possible to define a set of local variables $x_i \in \mathbb{R}^{n_i}$ such that $f(h) = \sum_i f_i(x_i)$, where each x_i is a subset of the global variable h. More specifically, each entry of the local variables corresponds to an entry of the global variable. Therefore we can define a mapping $\mathcal{G}(i,j)$ from the local variable indices into the global variable indices: $k = \mathcal{G}(i,j)$ means that the j^{th} entry of x_i is h_k (or $(x_i)_j = h_k$). For ease of notation, define $\tilde{h}_i \in \mathbb{R}^{n_i}$ as $(\tilde{h}_i)_j = h_{\mathcal{G}(i,j)}$. Then, the original optimization problem is equivalent to the following problem:

$$\min_{\{x_1,\ldots,x_N\}} \sum_i f_i(x_i) \quad s.t. \quad \tilde{h}_i = x_i. \tag{5}$$

It is important to note that each entry of the global variable may correspond to several entries of the local variables and so the constraints $\tilde{h}_i = x_i$ enforce the consensus between the local variables corresponding to the same global variable. The *augmented Lagrangian* corresponding to (5) is $L_\rho(x,h,y) = \sum_i \left(f_i(x_i) + u_i^\top(x_i - \tilde{h}_i) + (\rho/2)\|x_i - \tilde{h}_i\|_2^2\right)$. Now, we can apply ADMM to L_ρ. This results in solving N independent optimization problems followed by a step to achieve consensus among the solutions in each iteration. To solve the optimization problem (4) using this method, we need to address two questions: first, how to choose the local variables x_i, and second, how to the update them.

In Figure 1, the global variable h is represented as a cube. We decompose h into sub-cubes as shown by white lines. Each global variable inside the sub-cubes corresponds to only one local variable. The global variables on the border (white lines), however, correspond to more than one local variable. With this definition of x_i, the objective (4) decomposes as

Algorithm 1 Consensus ADMM

1: **Input:** data y, penalty matrix D, $\epsilon, \rho, \lambda_t, \lambda_s > 0$.
2: **Set:** $h \leftarrow 0$, $z \leftarrow 0$, $u \leftarrow 0$. ▷ Initialization
3: **repeat**
4: $x_i \leftarrow \underset{x_i}{\operatorname{argmin}} -l(y_i \mid x_i) + \Lambda_{(i)}^\top |D_{(i)} x_i| + (u_i)^\top x_i + (\rho/2)\|x_i - \tilde{h}_i\|_2^2$. ▷ Update local vars using PDIP
5: $h_k \leftarrow (1/S_k) \sum_{\mathcal{G}(i,j)=k} (x_i)_j$. ▷ Global update.
6: $u_i \leftarrow u_i + \rho(x_i - \tilde{h}_i)$. ▷ Dual update
7: **until** $\max\left\{\left\|h^{m+1} - h^m\right\|, \|h^m - x^m\|\right\} < \epsilon$
8: **Return:** h.

$\sum_i f_i(x_i)$ where $f_i(x_i) = -l(y_i \mid x_i) + \Lambda_{(i)}^\top |D_{(i)} x_i|$, and $\Lambda_{(i)}$ and $D_{(i)}$ contain the temporal and spatial penalties corresponding to x_i only in one sub-cube along with its boundary. Thus, we now need to use PDIP to solve N problems each of size n_i, which is feasible for small enough n_i. Algorithm 1 gives the general version of this algorithm. A more detailed discussion of this is in Appendix B of the Supplement where we show how to compute the dual and the derivatives of the augmented Lagrangian.

Because consensus ADMM breaks the large optimization into sub-problems that can be solved independently, it is amenable to a split-gather parallelization strategy via, e.g., the map reduce framework. In each iteration, the computation time will be equal to the time to solve each sub-problem plus the time to communicate the solutions on the master processor and perform the consensus step. Since each sub-problem is small, with parallelization, the computation time in each iteration will be small. In addition, our experiments with several values of λ_t and λ_s showed that the algorithm converges in few hundreds iterations. This algorithm is most useful if we can parallelize the computation over several machines with low communication cost between machines. In the next section, we describe another algorithm which makes the computation feasible on a single machine.

3.2 Linearized ADMM

Consider the generic optimization problem $\min_x f(x) + g(Dx)$ where $x \in \mathbb{R}^n$ and $D \in \mathbb{R}^{m \times n}$. Each iteration of the linearized ADMM algorithm (Parikh and Boyd 2014) for solving this problem has the form

$$\begin{aligned} x &\leftarrow \underset{\mu f}{\mathbf{prox}} \left(x - (\mu/\rho) D^\top (Dx - z + u)\right) \\ z &\leftarrow \underset{\rho g}{\mathbf{prox}} \left(z + u\right) \\ u &\leftarrow u + Dx - z \end{aligned}$$

where the algorithm parameters μ and ρ satisfy $0 < \mu < \rho / \|D\|_2^2$, $z, u \in \mathbb{R}^m$ and the proximal operator is defined as $\mathbf{prox}_{\alpha\varphi}(u) = \min_x \ \alpha \cdot \varphi(x) + \frac{1}{2}\|x - u\|_2^2$. Proximal algorithms are feasible when these proximal operators can be evaluated efficiently which, as we show next, is the case for our problem.

Lemma 1. *Let* $f(x) = \sum_k x_k + y_k^2 e^{-x_k}$ *and* $g(x) = \|x\|_1$.

Algorithm 2 Linearized ADMM

Input: data y, penalty matrix D, $\epsilon, \rho, \lambda_t, \lambda_s > 0$.
Set: $h \leftarrow 0$, $z \leftarrow 0$, $u \leftarrow 0$. ▷ Initialization
repeat
  $h_k \leftarrow \mathscr{W}\left(\frac{y_k^2}{\mu}\exp\left(\frac{1-\mu u_k}{\mu}\right)\right) + \frac{1-\mu u_k}{\mu}$ for all $k = 1, \ldots TS$. ▷ Primal update
  $z \leftarrow S_{\rho\lambda}(u)$. ▷ Elementwise soft thresholding
  $u \leftarrow u + Dh - z$. ▷ Dual update
until $\max\{\|Dh - z\|, \|z^{m+1} - z^m\|\} < \epsilon$
Return: z.

Then,

$$[\underset{\mu f}{\mathbf{prox}}(u)]_k = \mathscr{W}\left(\frac{y_k^2}{\mu}\exp\left(\frac{1-\mu u_k}{\mu}\right)\right) + \frac{1-\mu u_k}{\mu},$$
$$\underset{\rho g}{\mathbf{prox}}(u) = S_{\rho\lambda}(u)$$

where $\mathscr{W}(\cdot)$ is the Lambert W function (Corless et al. 1996), $[S_\alpha(u)]_k = \text{sign}(u_k)(|u_k| - \alpha_k)_+$ *and* $(v)_+ = v \vee 0$.

Therefore, Algorithm 2 gives a different method for solving the same problem. For this algorithm, both the primal update and the soft thresholding step are performed elementwise at each point of the spatio-temporal grid. It can therefore be extremely fast to perform these steps. However, because there are now many more dual variables, this algorithm will require more outer iterations to achieve consensus. It therefore is highly problem and architecture dependent whether Algorithm 1 or Algorithm 2 will be more useful in any particular context.

4 Empirical Evaluation

In this section, we examine both simulated and real spatio-temporal climate data. All the computations were performed on a Linux machine with four 3.20GHz Intel i5-3470 cores.

4.1 Simulations

Before examining real data, we apply our model to some synthetic data. This example was constructed to mimic the types of spatial and temporal phenomena observable in typical climate data. We generate a complete spatio temporal field wherein observations at all time steps and all locations are independent Gaussian random variables with zero mean. However, the variance of these random variables follows a smoothly varying function in time and space given by the following parametric model:

$$\sigma^2(t, r, c) = \sum_{k=1}^{K} W_k(t) \cdot \exp\left(\frac{(r - r_k)^2 + (c - c_k)^2}{2\sigma_k^2}\right)$$
$$W_k(t) = \alpha_k \cdot t + \exp(\sin(2\pi\omega_k t + \phi_k)).$$

The variance at each time and location is computed as the weighted sum of K bell-shaped functions where the weights are time-varying, consist of a linear trend $\alpha_k \cdot t$ and a periodic term $\beta_k \cdot \sin(2\pi\omega_k t + \phi_k)$. The bell-shaped functions impose spatial smoothness while the linear trend and the periodic terms enforce the temporal smoothness similar to the seasonal component in the real climate data. We simulated the data on a 5 by 7 grid and for 780 time steps with $K = 4$. This yields a small enough problem so that it can be evaluated many times while still mimicking important properties of climate data. Specific parameter choices of the variance function are shown in Table 1. For illustration, we also plot the variance function for all locations at $t = 25$ and $t = 45$ (Figure 2, top panel) in as well as the variance across time at $(0, 0)$ (bottom panel, orange).

Table 1: Parameters used to simulate data.

s	r_s	c_s	σ_s	α_s	ω_s	ϕ_s
1	0	0	5	0.5	0.121	0
2	0	5	5	0.1	0.121	0
3	3	0	5	-0.5	0.121	$\pi/2$
4	3	5	5	-0.1	0.121	$\pi/2$

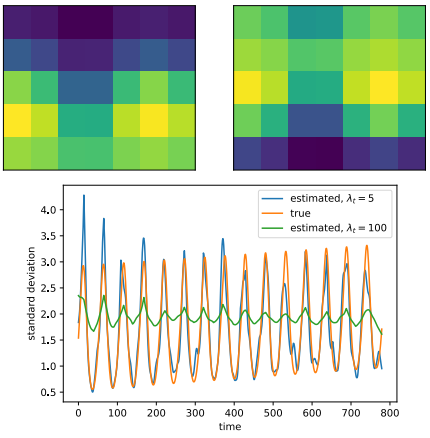

Figure 2: Top: Variance function at $t = 25$ (left) and $t = 45$ (right). Bottom: The true (orange) and estimated standard deviation function at the location (0,0). The estimated values are obtained using linearized ADMM with $\lambda_s = 0.1$ and two values of λ_t: $\lambda_t = 5$ (blue) and $\lambda_t = 100$ (green).

We estimated the linearized ADMM for all combinations of values of λ_t and λ_s from the sets $\lambda_t \in \{0, 1, 5, 10, 50, 100\}$ and $\lambda_s \in \{0, 0.05, 0.1, 0.2, 0.3\}$. For each pair, we then compute the mean absolute error (MAE) between the estimated variance and the true variance at all locations and all time steps. For $\lambda_t = 5$ and $\lambda_s = 0.1$, the MAE was minimized. The bottom panel of Figure 2 shows the true and the estimated standard deviation at location (0,0) and $\lambda_t = 5$ (blue) and $\lambda_t = 100$ (green) ($\lambda_s = 0.1$). Larger values of λ_t lead to estimated values which are "too smooth". The left panel of Figure 3 shows the convergence of both methods as a function of iteration. It is important to note that each iteration of the linearized algorithm takes 0.01 seconds on average while each iteration of the consensus ADMM takes about 20 seconds. Thus, where the lines meet at 400 iterations requires about 4 seconds for the linearized method and 2 hours for the consensus method.

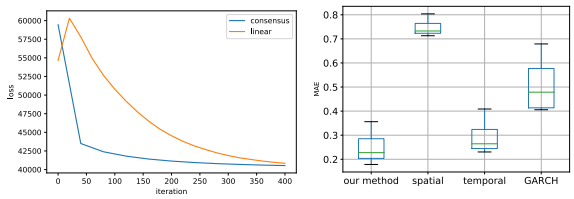

Figure 3: Left: Value of the objective function for linearized (orange) and consensus (blue) ADMM against iteration. Right: MAE for (1) our method with optimal values of λ_t and λ_s (2) temporal penalty only (3) spatial penalty only and (4) a GARCH(1,1).

To further examine the performance of the proposed model, we next compare it to three alternatives: a model which does not consider the spatial smoothness (equivalent to fitting the model in Section 2.1 to each time-series separately), a model which only imposes spatial smoothness, and a GARCH(1,1) model. We simulated 100 datasets using the method explained above with $\sigma_s \sim \text{uniform}(4, 7)$. The right panel of Figure 3 shows the boxplot of the MAE for these models. We note that, using an algorithm akin to (Hallac et al. 2017) ignores the spatial component and

4.2 Data Analysis

Consensus ADMM in Section 3.1 is appropriate when we can easily parallelize it over multiple machines. Otherwise, it is significantly slower, so all the results reported in this section are obtained using Algorithm 2. We applied this algorithm to the Northern hemisphere of the ERA-20C dataset available from the European Center for Medium-Range Weather Forecasts[4]. We use the 2 meter temperature measured daily at 12 p.m from January 1, 1960 to December 24, 2010.

Preprocessing Examination of the time-series alone demonstrates strong differences between trend and cyclic behavior across spatial locations. One might try to model the cycles by the summation of sinusoidal terms with different frequencies. However, for some time-series this may need many terms to be included in the summation to achieve a reasonable level of accuracy. In addition, such a model cannot capture the non-stationarity in the cycles.

Figure 8 shows the time-series of the temperature of three cities: Indianapolis (USA), San Diego (USA) and Manaus (Brazil). The time-series of Indianapolis and San Diego show clear cyclic behavior, though the amplitude of the cycles changes. The time-series of Manaus does not show any regular cyclic behavior. For this reason, we first apply trend filtering to remove seasonal terms and de-trend every time series. For each time-series, we found the optimal value of the penalty parameter using *k-fold cross-validation* with $k = 5$. We used the R package **genlasso** to perform these computations (Arnold and Tibshirani 2016). The blue curve in the left

[4] https://www.ecmwf.int

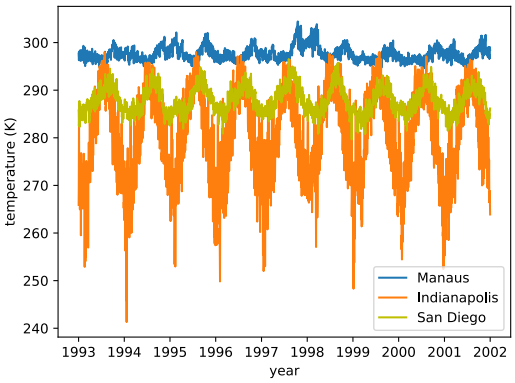

Figure 4: Time-series of the temperature (in Kelvin) of three cities.

panel of Figure Figure 9 shows the time-series of the temperature of Indianapolis after detrending using this method.

The red curve in the left panel of Figure Figure 9 shows the estimated SD (which is $\exp(h_t/2)$) of the residuals of the time-series of Indianapolis obtained from out proposed model. For ease of analysis, we compute the average of the estimated SD for each year. Both are shown in the middle panel of Figure 9. To smooth the annual trend, we add a long horizon penalty to (2). The estimated, smoothed SDs are shown the right panel of Figure 9. The annual average of the estimated SDs shows a linear trend with a positive slope.

Convergence As shown in Algorithm 2, we evaluated convergence using $\epsilon = 0.001\%$ of the MSE of the data. Our simulation experiments showed that the convergence speed depends on the value of λ_t and λ_s. Furthermore, using the solution obtained for smaller values of these parameters as a warm start for the larger values, the converges speed improves.

Model Selection One common method for choosing the penalty parameters in the Lasso problems is to find the solution for a range of the values of these parameters and then choose the values which minimize a model selection criterion. However, such analysis needs the computation of the degrees of freedom. Several previous work have investigated the df in generalized lasso problems (Tibshirani and Taylor 2012; Hu, Zeng, and Lin 2015; Zeng, Hu, and Li 2017). However, all these studies have considered the linear regression problem and, to the best of our knowledge, the problem of computing the df for generalized lasso with general objective function has not been considered yet. In this paper, we use a heuristic method for choosing λ_t and λ_s: we compute the optimal solution for a range of values of these parameters and choose the values which minimize $\mathscr{L}(\lambda_t, \lambda_s) = -l(y|h) + \sum \|D_{total} h\|$. This objective is a compromise between the negative log likelihood ($-l(y|h)$) and the complexity of the solution ($\sum \|D_{total} h\|$). For smoother solutions the value of $\sum \|D_{total} h\|$ will be smaller but with the cost of larger $-l(y|h)$. We computed the optimal solution for

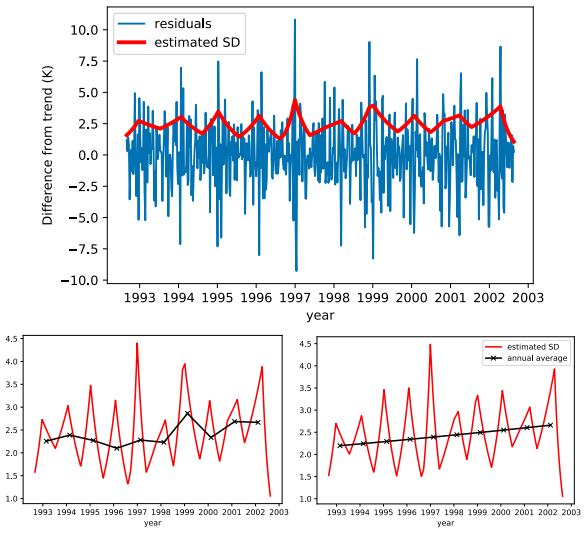

Figure 5: Left: The residuals of the time-series of Indianapolis (averaged weekly) and the estimated SD obtained from the method of Section 2.1 (red). Middle: the estimated SDs (red) and their annual average (black) without imposing the long horizon penalty. Right: the same as middle panel but here the long horizon penalty is imposed. See the text for more details. TODO: Fix the caption (top-bottom or left right)

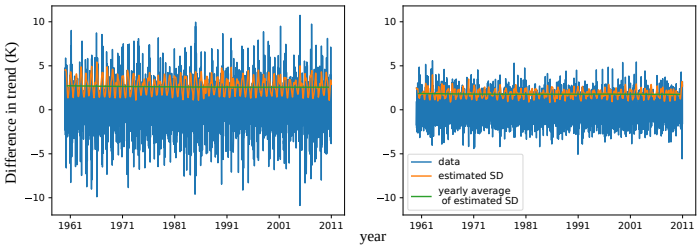

Figure 6: Detrended data and the estimated SD for a small midwestern city (left) and San Diego (right).TODO: Fix these

all the combinations of the following sets of values: $\lambda_t \in \{0, 2, 4, 8, 10, 15, 200, 1000\}$, $\lambda_s \in \{0, .1, .5, 2, 5, 10\}$. The best combination was $\lambda_t = 4$ and $\lambda_s = 2$. All the analyses in the next section are performed on the solution for these values.

Analysis of Trends in Temperature Volatility The top row of Figure 6 shows the detrended data, the estimated standard deviation and the yearly average of these estimates for two cities in the US: a small midwestern city (left) and San Diego (right). The estimated SD captures the periodic behavior in the variance of the time-series. In addition, the number of linear segments changes adaptively in each time window depending on how fast the variance is changing.

The yearly average of the estimated SD captures the trend in the temperature volatility. For example, we can see that the variance in the midwestern city displays a small positive trend. To determine how the volatility has changed in each

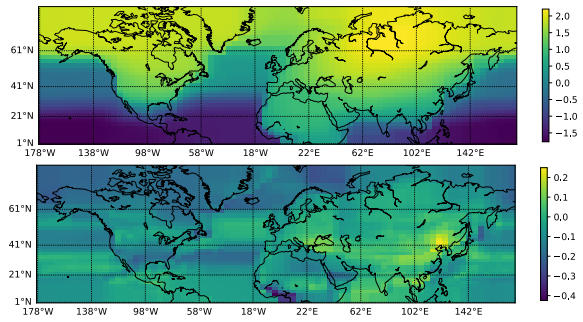

Figure 7: The average of the detrended estimated variance over the northern hemisphere (top) and the change in the variance from 1961 to 2011 (bottom).

location, we subtract the average of the estimated variance in 1961 from the average in the following years and compute their sum. The value of this change in the variance in each location is depicted in the right panel of Figure 7. The left panel of this shows the average estimated variance in each location. Since the optimal value of the spatial penalty is rather large ($\lambda_s = 2$) the estimated variance is spatially very smooth.

The SD in most locations on the northern hemisphere had a negative trend in this time period, though spatially, this decreasing pattern is localized mainly toward the extreme northern latitudes and over oceans. In many ways, this is consistent with climate change predictions: oceans tend to operate as a local thermostat, regulating deviations in local temperature, while warming polar regions display fewer days of extreme cold. The most positive trend can be observed in Asia and particularly in south-east Asia.

5 Discussion

In this paper, we proposed a new method for estimating the variance of spatio-temporal data. The main idea is to cast this problem as a constrained optimization problem where the constraints enforce smooth changes in the variance for neighboring points in time and space. In particular, the solution is piecewise linear in time and piecewise constant in space. The resulting optimization is in the form of a generalized LASSO problem with high-dimension, and so applying the PDIP method directly is infeasible. We therefore developed two ADMM-based algorithms to solve this problem: the consensus ADMM and linearized ADMM.

The consensus ADMM algorithm converges in a few hundreds of iterations but each iteration takes much longer than the linearized ADMM algorithm. The appealing feature of the consensus ADMM algorithm is that if it is parallelized on enough machines the computation time per iteration remains constant as the problem size increases. The linearized ADMM algorithm on the other hand converges in a few thousand iterations but each iteration is performed in a split second. However, since the algorithm converges in many iterations it is not very appropriate for parallelization. The reason is that after each iteration the solution computed in each machine should be broadcast to the master machine and this

operation takes some time which depends on the speed of the network connecting the slave machines to the master. A direction for future research would be to combine these two algorithms in the following way: the problem should be split into the sub-problems (as in the consensus ADMM) but each sub-problem can be solved using linearized ADMM.

References

Andersen, M. S.; Dahl, J.; and Vandenberghe, L. 2013. CVXOPT: A Python package for convex optimization, version 1.1. 6. *Available at cvxopt. org* 54.

Arnold, T. B., and Tibshirani, R. J. 2016. Efficient Implementations of the Generalized Lasso Dual Path Algorithm. *Journal of Computational and Graphical Statistics* 25(1):1–27.

Baum, B. A.; Wielicki, B. A.; Minnis, P.; and Parker, L. 1992. Cloud-property retrieval using merged HIRS and AVHRR data. *Journal of Applied Meteorology* 31(4):351–369.

Baum, B. A.; Wielicki, B. A.; Minnis, P.; Arduini, R. F.; and Tsay, S.-C. 1994. Multilevel cloud retrieval using multispectral HIRS and AVHRR data: Nighttime oceanic analysis. *Journal of Geophysical Research* 99:5499–5514.

Bender, F. A.; Ramanathan, V.; and Tselioudis, G. 2012. Changes in extratropical storm track cloudiness 1983–2008: Observational support for a poleward shift. *Climate Dynamics* 38(9-10):2037–2053.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 192–236.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31(3):307–327.

Bony, S.; Stevens, B.; Frierson, D. M.; Jakob, C.; Kageyama, M.; Pincus, R.; Shepherd, T. G.; Sherwood, S. C.; Siebesma, A. P.; Sobel, A. H.; et al. 2015. Clouds, circulation and climate sensitivity. *Nature Geoscience* 8(4):261.

Boucher, O.; Randall, D.; Artaxo, P.; Bretherton, C.; Feingold, G.; Forster, P.; Kerminen, V.-M.; Kondo, Y.; Liao, H.; Lohmann, U.; et al. 2013. Clouds and aerosols. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. 571–657.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* 3(1):1–122.

Corless, R. M.; Gonnet, G. H.; Hare, D. E. G.; Jeffrey, D. J.; and Knuth, D. E. 1996. On the LambertW function. *Advances in Computational Mathematics* 5(1):329–359.

Engle, R. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20(3):339–350.

Evan, A. T.; Heidinger, A. K.; and Vimont, D. J. 2007. Arguments against a physical long-term trend in global isccp cloud amounts. *Geophysical Research Letters* 34(4).

Fischer, E. M.; Beyerle, U.; and Knutti, R. 2013. Robust spatially aggregated projections of climate extremes. *Nature Climate Change* 3:1033—1038.

Frey, R. A.; Ackerman, S.; and Soden, B. J. 1996. Climate parameters from satellite spectral measurements. Part 1: Collocated AVHRR and HIRS/2 observations of spectral greenhouse parameter. *Journal of Climate* 9(2):327–344.

Gibberd, A. J., and Nelson, J. D. 2017. Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics* 26(3):623–634.

Grise, K. M.; Polvani, L. M.; Tselioudis, G.; Wu, Y.; and Zelinka, M. D. 2013. The ozone hole indirect effect: Cloud-radiative anomalies accompanying the poleward shift of the eddy-driven jet in the southern hemisphere. *Geophysical Research Letters* 40(14):3688–3692.

Hallac, D.; Park, Y.; Boyd, S.; and Leskovec, J. 2017. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, 205–213. New York, NY, USA: ACM.

Hansen, J.; Sato, M.; and Ruedy, R. 2012. Perception of climate change. *Proceedings of the National Academy of Sciences* 109(37).

Harvey, A.; Ruiz, E.; and Shephard, N. 1994. Multivariate stochastic variance models. *The Review of Economic Studies* 61(2):247–264.

Hu, Q.; Zeng, P.; and Lin, L. 2015. The dual and degrees of freedom of linearly constrained generalized lasso. *Computational Statistics & Data Analysis* 86:13–26.

Huntingford, C.; Jones, P. D.; Livina, V. N.; Lenton, T. M.; and Cox, P. M. 2013. No increase in global temperature variability despite changing regional patterns. *Nature* 500(7462):327–330.

Kahn, B. H.; Fishbein, E.; Nasiri, S. L.; Eldering, A.; Fetzer, E. J.; Garay, M. J.; and Lee, S.-Y. 2007. The radiative consistency of atmospheric infrared sounder and moderate resolution imaging spectroradiometer cloud retrievals. *Journal of Geophysical Research: Atmospheres* 112(D9).

Kim, S.-J.; Koh, K.; Boyd, S.; and Gorinevsky, D. 2009. ℓ_1 trend filtering. *SIAM Review* 51(2):339–360.

Monti, R. P.; Hellyer, P.; Sharp, D.; Leech, R.; Anagnostopoulos, C.; and Montana, G. 2014. Estimating time-varying brain connectivity networks from functional mri time series. *NeuroImage* 103:427–443.

Myers, T. A.; Mechoso, C. R.; and DeFlorio, M. J. 2018. Importance of positive cloud feedback for tropical atlantic interhemispheric climate variability. *Climate Dynamics* 51(5-6):1707–1717.

Parikh, N., and Boyd, S. 2014. Proximal Algorithms. *Foundations and Trends® in Optimization* 1(3):127–239.

Rhines, A., and Huybers, P. 2013. Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proceedings of the National Academy of Sciences* 110(7):E546–E546.

Rossow, W. B., and Schiffer, R. A. 1991. Isccp cloud data products. *Bulletin of the American Meteorological Society* 72(1):2–20.

Schreier, M.; Kahn, B.; Eldering, A.; Elliott, D.; Fishbein, E.; Irion, F.; and Pagano, T. 2010. Radiance comparisons of modis and airs using spatial response information. *Journal of Atmospheric and Oceanic Technology* 27(8):1331–1342.

Schreier, M.; Kahn, B.; Sušelj, K.; Karlsson, J.; Ou, S.; Yue, Q.; and Nasiri, S. 2014. Atmospheric parameters in a subtropical cloud regime transition derived by AIRS and MODIS: Observed statistical variability compared to era-interim. *Atmospheric Chemistry and Physics* 14(7):3573–3587.

Screen, J. A. 2014. Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nature Climate Change* 4:577—582.

Staten, P. W.; Kahn, B. H.; Schreier, M. M.; and Heidinger, A. K. 2016. Subpixel characterization of HIRS spectral radiances using cloud properties from AVHRR. *Journal of Atmospheric and Oceanic Technology* 33(7):1519–1538.

Tibshirani, R. J., and Taylor, J. 2011. The solution path of the generalized lasso. *Annals of Statistics* 39(3):1335–1371.

Tibshirani, R. J., and Taylor, J. 2012. Degrees of freedom in lasso problems. *The Annals of Statistics* 40(2):1198–1232.

Tibshirani, R. J. 2014. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics* 42:285–323.

Trenberth, K. E.; Zhang, Y.; Fasullo, J. T.; and Taguchi, S. 2014. Climate variability and relationships between top-of-atmosphere radiation and temperatures on earth. *Journal of Geophysical Research: Atmospheres* 120(9):3642–3659.

Vasseur, D. A.; DeLong, J. P.; Gilbert, B.; Greig, H. S.; Harley, C. D. G.; McCann, K. S.; Savage, V.; Tunney, T. D.; and O'Connor, M. I. 2014. Increased temperature variation poses a greater risk to species than climate warming. *Proceedings of the Royal Society of London B: Biological Sciences* 281(1779).

Wang, Y.-X.; Sharpnack, J.; Smola, A. J.; and Tibshirani, R. J. 2016. Trend filtering on graphs. *Journal of Machine Learning Research* 17(105):1–41.

Wielicki, B. A.; Young, D.; Mlynczak, M.; Thome, K.; Leroy, S.; Corliss, J.; Anderson, J.; Ao, C.; Bantges, R.; Best, F.; et al. 2013. Achieving climate change absolute accuracy in orbit. *Bulletin of the American Meteorological Society* 94(10):1519–1539.

Zeng, P.; Hu, Q.; and Li, X. 2017. Geometry and Degrees of Freedom of Linearly Constrained Generalized Lasso. *Scandinavian Journal of Statistics* 44(4):989–1008.

A PDIP for ℓ_1 Trend Filtering of variance

In this appendix we provide more details on how to solve the optimization problem with the objective specified in equation 2 using PDIP. The objective function is convex but not differentiable. Therefore, to be able to use PDIP we first need to derive the dual of this problem. We note that this is a generalized LASSO problem (Tibshirani and Taylor 2011). The dual of a generalized LASSO with the objective $f(x) + \lambda \|Dx\|_1$ is:

$$\min_{\nu} \quad f^*(-D^\top \nu) \qquad\qquad \text{s.t.} \quad \|\nu\|_\infty \leq \lambda$$

where $f^*(\cdot)$ is the Fenchel conjugate of f: $f^*(u) = \max_x u^\top x - f(x)$. It is simple to show that for the objective function of Equation 2

$$f^*(u) = \sum_t (u_t - 1) \log \frac{y_t^2}{1 - u_t} + u_t - 1.$$

Each iteration of PDIP involves computing a search direction by taking a Newton step for the system of nonlinear equations $r_w(v, \mu_1, \mu_2) = 0$, where $w > 0$ is a parameter and

$$r_w(v, \mu_1, \mu_2) := \begin{bmatrix} r_{dual} \\ r_{cent} \end{bmatrix} = \begin{bmatrix} \nabla f^*(-D^\top v) + \mu_1 - \mu_2 \\ -\mu_1(v - \lambda \mathbf{1}) + \mu_2(v + \lambda \mathbf{1}) - w^{-1}\mathbf{1} \end{bmatrix}$$

for $w > 0$, where μ_1 and μ_2 are dual variables for the ℓ_∞ constraint. Let $A = [\nabla r_{dual}^\top, \nabla r_{cent}^\top]^\top$. The newton step takes the following form

$$r_w(v, \mu_1, \mu_2) + A \begin{bmatrix} \nabla v \\ \nabla \mu_1 \\ \nabla \mu_2 \end{bmatrix} = 0$$

We have:

$$A = \begin{bmatrix} \nabla^2 f^*(-D^\top v) & I & -I \\ -\mathbf{diag}(\mu_1)\mathbf{1} & -v + \lambda \mathbf{1} & \mathbf{0} \\ \mathbf{diag}(\mu_2)\mathbf{1} & v + \lambda \mathbf{1} & \mathbf{0} \end{bmatrix}$$

Therefore, to perform the Newton step we need to compute $\nabla f^*(-D^\top v)$ and $\nabla^2 f^*(-D^\top v)$. It is straightforward to show that

$$\nabla f^*(-D^\top v) = -\nabla_u f^*(u) D^\top,$$
$$u = -D^\top v, \quad (\nabla_u f^*(u))_j = \log\left(\frac{y_j^2}{1 - u_j}\right)$$
$$\nabla^2 f^*(-D^\top v) = D \nabla_u^2 f^*(u) D^\top, \quad (\nabla_u^2 f^*(u))_j = \mathbf{diag}\left(\frac{1}{1 - u_j}\right)$$

Having computed the conjugate function and its gradient and Jacobian, now we can use a number of convex optimization software packages which have an implementation of PDIP to solve the optimization problem with the objective function 2. We chose the python API of the `cvxopt` package (Andersen, Dahl, and Vandenberghe 2013).

B PDIP Update in Algorithm 1

In this section we give more details on performing the x-update step in Algorithm 1. We need to solve the following optimization problem:

$$\hat{x} := \operatorname*{argmin}_x \left(\sum_{j=1}^{n_b} (x_j + y_j^2 e^{-x_j}) + (\rho/2)\|x - \tilde{z} + u\|_2^2 + \Lambda^\top |Dx| \right)$$

where n_b is the number of local variables in each sub-cube in Figure 1, and for ease of notation we have dropped the subscript i and superscript m.

The matrix D has the following form: $D^t = [D_{temp}^t | D_{spat}^t]$. The matrix D_{temp} is the following block-diagonal matrix and corresponds to the temporal penalty:

$$D_{temp} = \begin{bmatrix} D_t & & \\ & \ddots & \\ & & D_t \end{bmatrix}$$

where D_t was first introduced in Section 2 of the main text and has the following form:

$$D_t = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix}$$

The number of the diagonal blocks in D_{temp} is equal to the grid size $n_r \times n_c$. Each row of the matrix D_{spat} corresponds to one spatial constraint in Equation (3) in the text. For example, the first T rows correspond to $|h_{11t} - h_{21t}|$ for $t = 1, ..., T$, the next T rows correspond to $|h_{11t} - h_{12t}|$, and so on.

This optimization problem, is again a generalized LASSO problem with $f(x) = \sum_{j=1}^{n_b}(x_j + y_j^2 e^{-x_j}) + (\rho/2)\|x - \tilde{z} + u\|_2^2$.

As it was explained in Appendix A, the dual of this optimization problem is: $\min_\nu f^*(-D^\top \nu)$ with the constraints $|\nu_k| \leq \Lambda_k$. To use PDIP we first need to compute the conjugate function $f^*(\cdot)$. We have:

$$\begin{aligned} f^*(\xi) &= \max_x \quad \xi^\top x - f(x) \\ &= \max_x \quad \sum_{j=1}^{n_b} (\xi_j x_j - x_j - y_j^2 e^{-x_j} - (\rho/2)(x_j - \tilde{z}_j + u_j)) \end{aligned}$$

Setting the derivative of the terms inside the summation to 0, we obtain:

$$\xi_j - y_j^2 e^{-x_j^*} - \rho x_j^* + \rho(\tilde{z}_j - u_j) = 0 \tag{6}$$

where x^* is the maximizer in B. Then, it can be shown that x_j^* which satisfies (6) can be obtained as follows:

$$x_j^* = \mathscr{W}\left(\frac{y_j^2}{\rho} e^{\phi_j}\right) - \phi_j$$
$$\phi_j = \frac{1 - \xi_j - \rho(\tilde{z}_j - u_j)}{\rho}$$

In this equation, $\mathscr{W}(\cdot)$ is the *Lambert function* (Corless et al. 1996). Finally, the conjugate function is: $f^*(\xi) = \sum_{j=1}^{n_b}(\xi_j x_j^* - x_j^* - y_j^2 e^{-x_j^*} - (\rho/2)(x_j^* - \tilde{z}_j + u_j))$.

To use PDIP, we also need to evaluate ∇f^* and $\nabla^2 f^*$. First note that $\frac{\partial \mathscr{W}(q)}{\partial q} = \frac{\mathscr{W}(q)}{q(1+\mathscr{W}(q))}$ and $\frac{\partial^2 \mathscr{W}(q)}{\partial q^2} = -\frac{\mathscr{W}^2(q)(\mathscr{W}(q)+q)}{q^2(1+\mathscr{W}(q))^3}$. Using the chain rule we get:

$$\frac{\partial f^*(\xi)}{\partial \xi_j} = x_j^* + \frac{\partial x_j^*}{\partial \xi_j}\left[\xi_j - 1 + y_j^2 e^{-x_j^*} + \rho(\tilde{z}_j - u_j - x_j^*)\right]$$

where we have:

$$\frac{\partial x_j^*}{\partial \xi_j} = \frac{1}{\rho(1 + \mathscr{W}((y_j^2/\rho)e^{-\phi_j}))}$$

By some tedious but straightforward computation we can obtain the second derivatives:

$$\frac{\partial^2 f^*(\xi)}{\partial \xi_j^2} = \frac{\partial x_j^*}{\partial \xi_j} - \rho\frac{\partial^2 x_j^*}{\partial \xi_j^2}\left[\phi_j + x_j^* - \tilde{z}_j + u_j\right] + \frac{\partial x_j^*}{\partial \xi_j}\left[1 - y_j^2\frac{\partial x_j^*}{\partial \xi_j}e^{-x_j^*} - \rho\frac{\partial x_j^*}{\partial \xi_j}\right]$$

$$\frac{\partial^2 x_j^*}{\partial \xi_j^2} = \frac{\mathscr{W}((y_j^2/\rho)e^{-\phi_j})}{\rho^2(1 + \mathscr{W}((y_j^2/\rho)e^{-\phi_j}))^3}$$

C Data Exploration

In this appendix we examine some of the properties of the time-series of the temperature in ERA-20C dataset. The goal here is to demonstrate some of the difficulties in modeling the trend in the temperature volatility and motivate our methodology.

Figure Figure 8 shows the time-series of the temperature of three cities: Indianapolis (USA), San Diego (USA) and Manaus (Brazil). The time-series of Indianapolis and San Diego show clear cyclic behavior. However, while it seems (qualitatively) that these cycles can be modeled by a sinusoidal function for Indianapolis, the same is not true for San Diego. Also, the amplitude of the cycles changes from some years to others. The time-series of Manaus does not show any regular cyclic behavior. This demonstrates the first difficulty in analyzing the variance of this data: to analyze the variance, we first need to remove the cyclic terms from all time-series. However, there is a lot of variations in the cyclic behavior of the time-series of different locations. In addition, some of these cycles cannot be easily modeled by a parametric function [5]. To overcome these issues, we use a non-parametric approach to remove the cyclic terms from the time-series and de-trend them. This approach, called ℓ_1-*trend filtering* is explained in Section 2 of the text. We detrended each time-series separately using this method. For each time-series, we found the optimal value of the penalty parameter using *k-fold cross-validation* with $k = 5$. We used the R package **genlasso** to perform these computations (Arnold and Tibshirani 2016).

The blue curve in the left panel of Figure Figure 9 shows the time-series of the temperature of Indianapolis after detrending using this method. This figure, reveals another difficulty in estimating the trend of volatility in this data: the variance of this signal, shows cyclic behavior. Also, the cycles are not regular and their amplitude and frequency change. Even if one can describe the behavior of the variance of the time-series at all locations using a single parametric model (for example a variant of the GARCH models (Bollerslev 1986)), it is not clear how the trend in the variance should be investigated in this framework. These observations motivate the need to develop a non-parametric framework for the problem at hand.

[5] One might try to model the cycles by the summation of sinusoidal terms with different frequencies. However, for some time-series this may need many terms to be included in the summation to achieve a reasonable level of accuracy. In addition, this model cannot capture the non-stationarity in the cycles.

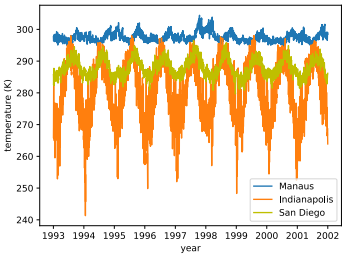

Figure 8: Time-series of the temperature (in Kelvin) of three cities.

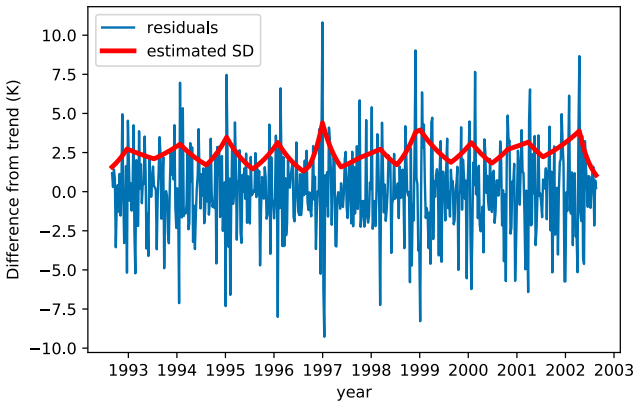

Figure 9: Left: The residuals of the time-series of Indianapolis (averaged weekly) and the estimated SD obtained from the method of Section 2.1 (red). Middle: the estimated SDs (red) and their annual average (black) without imposing the long horizon penalty. Right: the same as middle panel but here the long horizon penalty is imposed. See the text for more details.

The red curve in the left panel of Figure Figure 9 shows the estimated SD (which is $\exp(h_t/2)$) of the residuals of the time-series of Indianapolis obtained from out proposed model. To reduce the number of time-steps we work on the weekly averaged of the data. The curve of the estimated SD captures the periodic variations in the SD of the signal. Just by looking at this curve, it is hard to say if the SD is decreasing or increasing. Therefore, we compute the average of the estimated SD for each year. The estimated SD together with this annual average is shown in the middle panel of Figure 9. As it can be seen, the annual trend is not smooth. This is because in the optimization problem (2), the smoothness of the annual trend is not encouraged. To remedy this, we add the following long horizon penalty to (2):

$$\sum_{i=1}^{N_{year}-2} \left| \sum_{t=1}^{52} h_{t_1} - 2h_{t_2} + h_{t_3} \right| \tag{7}$$

where $t_1 = 52(i-1) + t$, $t_2 = 52i + t$ and $t_3 = 52(i+1) + t$. Also, N_{year} is the number of years over which we are performing our analysis (here $N_{year} = 10$). Since we are working on the weekly averaged data, each year corresponds to 52 observations. In the matrix form, the penalty (7) adds N_{year} rows to the matrix D. The estimated SDs using this penalty matrix is shown in the right panel of Figure 9. The annual average of the estimated SDs shows a

linear trend with a positive slope.