Title
WIDALS & FRKALS: Facile Spacio-Temporal Prediction for Massive Data

Author
Zes, Dave

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

WIDALS & FRKALS:

Facile Spacio-Temporal Prediction for Massive Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Dave Anthony Zes

2012

Abstract of the Dissertation

WIDALS & FRKALS:

Facile Spacio-Temporal Prediction for Massive Data

by

Dave Anthony Zes

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2012

Professor Ying Nian Wu, Chair

In the following we unite Adaptive Least Squares (ALS) and Inverse Distance Weighting as a computationally frugal means of modeling very large space-time data. This technique, dubbed weighting by inverse distance with adaptive least squares (WIDALS) boasts several merits, including a small and readily interpretable hyperparameter space, and relative ease of implementation. We include RMSE comparisons between WIDALS and various solutions including the Kalman solution on small simulated data sets. We culminate our work with a large scale imputation/model fitting ritual (dubbed "phyning") using WIDALS on 6 contemporaneous climatological data set, provided by the National Climatological Data Center, possessing $6 \cdot 11669 = 70014$ covariates/responses over 1016 time points accompanied by about 24 million exogenous covariates for a total of about 96,000,000 scalar data.

The dissertation of Dave Anthony Zes is approved.

Zhuowen Tu

Rick Paik Schoenberg

Jan De Leeuw

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2012

Table of Contents

List of Figures

List of Tables

Acknowledgments

To Benevolence.

Vita

2005	B.S. (Mathematics), California State University, Dominguez Hills.
2006	M.S. (Statistics), UCLA.
2006–2011	Teaching Assistant, Statistics Department, UCLA.
2012–present	Director of Analytics, Korn/Ferry International.

Publications

SSsimple Reference Manual and Vignettes. v0.6.0. The Comprehensive R Archive Network, 2011, 2012. Guide and tutorial for R package **SSsimple**.

Facile Spacio-Temporal Modeling: Forecasting with Adaptive Least Squares and the Kalman Filter. Journal of Environmental Statistics (Accepted for publication, *Journal of Environmental Statistics*).

Crittenden AN, NL Conklin-Brittain, MJ Schoeninger, DA Zes, and FW Marlowe. Juvenile Foraging among the Hadza: Implications for Human Life History. In Revision. *Evolution and Human Behavior.*

Crittenden AN, FW Marlowe, and DA Zes. In review. Kin Selection and Reciprocity in the Food Sharing Networks of Hadza Hunter-Gatherer Children. *Proceedings of the Royal Academy of Sciences, Biology.*

WIDALS Reference Manual and Vignettes. v0.5.1. The Comprehensive R Archive Network, 2012. Guide and tutorial for R package **widals**.

Facile Spacio-Temporal Modeling for Massive Data: WIDALS, Weighting by Inverse Distance with Adaptive Least Squares. Journal of Environmental Statistics (Submitted for publication, *Journal of Environmental Statistics*).

CHAPTER 1

Introduction

The sub-discipline of statistics that focuses on observations over the space-time domain has received considerable interest recently. The broad range of application, along with the perceived import of these applications being major motivators. Spacio-temporal prediction can provide valuable answers in many different contexts, such as ecology (pollution, new growth, habitat), industry (resource availability, regulation attainment), climate (in every imaginable way), even in marketing (demographics). Space-time modeling, or *systems*, has undergone a parallel development in engineering, where much emphasis is given to guidance and signal tracking (e.g., [AM79, Hay02, Lju98, Say03]). While differences in application between the two disciplines necessitates unique attention to context, fruitful discussions of methodology in either field necessitates appreciating analogous methods known to the other. It is also true that particular analytical difficulties that arise in one will also at some point arise in the other, since both ultimately call upon the same theoretical underpinnings. The growing availability of massive data spawns, for the statistician, one such difficulty. When performing inference, one has an almost pavlovian reaction to handling more and more data, having learned early on that the addition of more predictors can drop the standard error of the sought estimate — in the best case — by $\sqrt{n}$. The well known analytical inhibition is that the computational expense of an exact linear (or a translated non-linear) solution is $O(n^3)$ (this property of solving the relevant normal system of equations is well known — see, e.g., [KJ99], pg 673, or [JCH07]).

Moreover, exactly solving a solution for large n potentially introduces numerical instability. In engineering, attention has come in the form of variants of certain solutions. The square-root Kalman filter, for example, mitigates numeric instability [Say03, Hay02], and its "unscented" variant reduces computational complexity of retrospective hyper-paramenterization, but the system solution is still $O(n^3)$ [MW01]. In the statistical setting, for example when examining terrestrial processes, there commonly exists the constraint that observations over space covary as some well behaved function (a *covariogram*) of distance. Sahu *et al* successfully employ multilevel systems to describe and predict rainfall and particulate matter (PM) on relatively small, spacially local data [LSM07, SM05b, SM05a]. Cressie and Wilke coalesce a familiar state-space formulation and space-time modeling in [CW02]. The authors make note that estimating the proffered system parameters (i.e., *identifying* their system) given data can pose challenges, especially in high dimension. This may be a slight overstatement, since a researcher data-in-hand can often intuit a sense of the system covariance structure and its autoregressive properties. *Solving* the system, though, for fixed number of time points, τ, is still $O(n^3)$. Indeed, the constraint that the observation-space covariance has structure imparted to it through a covariogram, in-and-of-itself, does not mitigate the computational burden — although using iterative matrix inversion techniques can provide some relief, especially when sites are spacially located in a regular pattern [BBN02].

An early spacial modeling technique, known generically as "inverse distance weighting" (IDW), interpolates (meaning to predict to an unmonitored site) using a linear combination of recorded observations and weights created directly as some function of distance (see [She68] and references therein). It is not a particular irony that this easy (typically) $O(n)$ method long predates our current era of big data. Fifty years ago, prior to personal computing, a hundred data were "big." In its earlier manifestations, IDW was taken as a standalone method by which one could

fit deterministic surface shapes. In concept, this sort of approach can be thought of as proginating distance-based deterministic spacial transformations, such as radial bases — though as part of larger stochastic systems. As a standalone means of modeling spacial phenomenon, it is sufficiently archaic so as to be ignored by many modern authoritative texts, e.g., [Cre93, GDG10].

Abiding desire to make use of more flexible multilevel models (systems) in the statistical setting, much emphasis in large data analysis, therefore, has been on approximating an ideal solution. One common approach interpolates using a local neighborhood of observations. This $O(n)$ approach has some rigorous justification. Michael Stein has demonstrated that, given certain assumptions about the covariance function and the location of sites, under asymptotic infill the normal system weights are dominated by nearby locations — the ratio of the MSE over the ideal MSE converges to 1 [Ste02]. The effect of distant observations are, in essence, "screened" by local observations — a conjecture that had been erstwhile empirically supported in geostatistics [Jea99]. The most obvious implementation involves using only a neighborhood of some new location, $\omega^\star$, as support. Regional discontinuity of spacial predictions typically produced by this method is a poor analogy to "the mosaic of regional patterns" produced by tessellation-driven spacial *point process* solutions developed in, e.g., [Sch11], and also [Bar08]. [1] Moreover, if we seek to interpolate to $n^\star$ unmonitored locations, this will mean the sequential inversion of all the small concomitant local observation matrices — perhaps n or so inversions if $n^\star > n$. An artful alternative involves processing the support covariance with a (preferably positive definite) function over distance. The correct choice of functions, in essence, compresses, or *tapers* the distance-scale of the covariance. The result is a sparse matrix [FGN06, FS09]. Sparsification reduces the computational expense of inversion, and numerous computational approaches exist for handling sparse matrix manipulations, [CR05]. Moreover, unlike in the

[1] In fact, Voronoi tessellations appear (asymptotically) in many spacial hierarchical solutions, including the Kalman Filter, and, yes, WIDALS too.

prior approach of segregating data into local neighborhoods, tapering requires only a single inversion for arbitrarily large $n^\star$. The central drawback is that as the scale of the covariogram becomes relatively large compared to the range of the distance support, the tapered covariance prediction deteriorates rapidly. Perhaps the most interesting property of tapering is that it is amenable to — so to say — model optimism. If the researcher can explain away long range stochastic covariation by carefully selecting covariates to contribute to the (deterministic) mean function, then tapering rules the day. In practice, of course, explaining away long range spacial covariation is often not possible, if for no other reason than the lack of such covariates. Another approach known an Fixed Rank Filtering [CSK10], in fact a temporal extension of Fixed Rank Kriging [JC04], estimates the system covariance in lower dimension. The authors accomplish this by recasting mutual distances between sites into a smaller subspace of relative distances from strategically placed "resolutional" points; smoothing the empiric spacial covariance; invoking the computationally less intensive Sherman-Morrison-Woodbury equality for inversion. Ultimately, the computational burden becomes a function of the amount of smoothing that is imposed on the empiric covariance. One potential shortcoming occurs when the monitored sites appear in clusters, as locating smoothing loci for the empiric covariance can pose challenges. Another approach makes use of multiresolutional modeling (MRM). The spacial adaptation, MRSM, has been applied with great success to global Total Column Ozone (TCO) [HC96, HCG02, JCH07, JC02]. MRSM can be thought of as either a spacial analogue, or adaptation, of wavelet analysis. Each point (or small, compact region) in the spacial domain is inherited by a sequence of hierarchical regions. Any point, then, can be assigned a response prediction by using a linear combination of properties (e.g., the mean) of each inheriting region. If the resolution is fixed, i.e., the number of hierarchical levels are fixed, then a static MRSM is $O(n)$, [JCH07]. Space-time generalizations have arisen in numerous fields for a variety of

applications, from cardiology [BBR09], to video processing [Gar99]. One drawback of resolutional modeling is that drawback known to wavelet decomposition: The choice of the transformation forces the interpretation of the domain. Wavelets that can describe *continuous* (non-discrete) non-constant responses within the highest resolutional windows can be computationally expensive [ABS00]. Obviating this, we interpolate to, or predict from, regions, not points. Especially in the context of global modeling, this is a fairly trivial concern. For example, near the equator one can zoom into a region about a kilometer square with only 14 hierarchical levels ($20000 \cdot 2^{-14} = 1.2$). If we regard the number of levels, m, as an arbitrary constant, then space-time wavelet modeling is $O(m^3 n)$.

Weighting by Inverse Distance with Adaptive Least Squares (WIDALS), the central theme of this present work, must be regarded as the ultimate in computational frugality. WIDALS is an approximation; it mimics the BLUP of a general, dynamic linear system. WIDALS will not strike awe into anyone — it is barbarically simple. Despite, WIDALS possesses several remarkable virtues.

❶ Very low computational expense (e.g., forecasting or interpolating to fixed $n^\star$ sites at τ time steps is $O(n\tau)$).

❷ Very low computer memory expense. There is no need to ever store a full distance matrix, or its concomitant covariance, or the covariance inverse in memory during execution. §4.1.3.2

❸ Implicit handling of temporal correlation of model noise. (4.7)-(4.9)

❹ Small hyperparameter space (3-5 readily interpretable positive scalar values). Table 5.1

❺ Hyperparameter robustness. Miss-location of hyperparameters in fitting typically results in relatively mild deterioration of testing RMSE.

❻ Because of the small hyper-parameter space, metaheuristic fitting techniques can be employed.

❼ The Metaheuristic Stochastic-Search (MSS) allows for minimization of an arbitrary cost function, and is perfectly suited for parallel processing.

❽ Because the WIDALS solution explicitly segregates the deterministic and stochastic parts, it is readily extensible, e.g., the researcher can easily adapt the algorithm to account for time-varying, or site-varying spacial covariation.

❾ Relative ease of implementation.

As is typically the case with approximations, the central detraction is the escalation in the cost function. If one considers the MSE of a two-level state-space system, for example, WIDALS is suboptimal in three ways. Two occur at the adaptive least squares (ALS) stage (mean function estimation). The third degradation occurs during stochastic interpolation. A more particular hypothesis of this present work is that, in a retrospective training-testing setting (RTT), owing to the relative simplicity of its hyper-parameterization, WIDALS can produce comparable MSE even against an optimal solution. Moreover, WIDALS may outperform an "optimal" solution in the case of minor system misspecification, e.g., the presence of slight temporal correlation in the measurement space.

A reader disenchanted with WIDALS's lack of elegance, and its likely *a prior* sub-optimality, should at least view WIDALS's speed and easy implementation as an invitation to use it as a benchmark by which the incredulous researcher can test their own candidate solutions.

1.1 Notation

Important objects and their respective symbols are given in Table 1.1. The last four entries are superscripts used at times in this submission to remind the reader of the dimensionality of certain operations.

Symbol	Dim.	Description
$\mathbf{Z}$	$(\tau \times n)$	Data matrix. Time indexed by row, location indexed by column.
$\mathbf{z}_t$	$(1 \times n)$	Observation vector at time t.
$\mathbf{H}_t$	$(n \times p)$	A deterministic mapping/transformation into the observation space containing $\mathbf{z}_t$; the regressor matrix at time t; a bases transform.
$\boldsymbol{\beta}_t$	$(p \times 1)$	The partial slopes for $\mathbf{H}_t$.
$\mathbf{b}_t$	$(p \times 1)$	An estimate of $\boldsymbol{\beta}_t$.
$\mathbf{R}_t$	$(n \times n)$	The true variance of the measurement stochasticity of $\boldsymbol{z}_t$.
$\mathbf{C}_t$	$(n \times n)$	An estimate of $\mathbf{R}_t$.
$\mathbf{L}_{\mathrm{HH}\,t}$	$(p \times p)$	An estimate of $\mathsf{E}\left[\,\mathbf{H}_t^T\,\mathbf{H}_t\,\right]$.
$\mathbf{L}_{\mathrm{Hz}\,t}$	$(p \times 1)$	An estimate of the $\mathsf{E}\left[\,\mathbf{H}_t^T\,\boldsymbol{z}_t^T\right]$.
$\diamond$	—	Used as a superscript to indicate the object applies to the predictor space (i.e., the monitored sites).
$\star$	—	Used as a superscript to indicate the object applies to the response space (i.e., the unmonitored sites).
$\diamond\times\diamond$	—	Used as a superscript to indicate the object is either a product of predictor objects, or a function mapping from predictor space to predictor space.
$\diamond\times\star$	—	Used as a superscript to indicate the object is either a product of a predictor space object and a response space object, or a function mapping from predictor space to response space.

Table 1.1: Essential symbols

CHAPTER 2

Adaptive Least Squares

Adaptive Least Squares (ALS) is the cornerstone of Weighting by Inverse Distance with ALS (WIDALS) and Fixed Rank Kriging with ALS (FRKALS). ALS has several interpretations and hence implementations; however, the common thread in all is that ALS estimates potentially time-varying partial slopes by recursively adjusting system covariances. In this Chapter we will first detail applying ALS under the assumption of a purely spacially and temporally auto-regressive system. We will then detail applying ALS to solve a system assuming a potentially time-varying mean function of the form $\boldsymbol{\beta}_t\mathbf{H}$. We will lastly discuss a hybrid variant. WIDALS and FRKALS presented later rely on the latter two types of ALS implementations.

2.1 ALS Under a Purely Auto-Regressive Response

Consider an auto-regressive system independent of Ω, and only dependent on time labels signaling $\Delta t = 1$.

$$\boldsymbol{z}_t = B\big(\mathbf{Z}_{1:t-1}\big) + \boldsymbol{\varepsilon}_t \tag{2.1}$$

If $B(\cdot)$ is linear, then the optimal forecasting estimator comes by way of

$$B = \mathsf{E}\big[\boldsymbol{\zeta}_{1:t-1}^T\, \boldsymbol{\zeta}_{1:t-1}\big]^{-1}\, \mathsf{E}\big[\boldsymbol{\zeta}_{1:t-1}^T\, \boldsymbol{z}_t\big] \tag{2.2}$$

$$\widehat{\boldsymbol{z}}_t = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{t-1})\, B \tag{2.3}$$

with

$$\boldsymbol{\zeta}_{1:t-1} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_{t-1}) \tag{2.4}$$

the flattened, or vectorized version of the random variable, $\boldsymbol{Z}_{1:t-1}$. The challenge, of course, is estimating the covariance across lag-0 (the regressor-regressor covariance, the first expectation in (2.2)) and the covariance between lag-0 and lag-1 (the regressor-response covariance, the second expectation).

We now consider a special AR1 case of (2.1):

$$\mathbf{B}_t = \mathbf{I}\mathbf{B}_{t-1} + \boldsymbol{\nu}_t \tag{2.5}$$

$$\mathbf{z}_t = \mathbf{z}_{t-1}\,\mathbf{B}_t + \boldsymbol{\varepsilon}_t \tag{2.6}$$

with $\boldsymbol{\nu}_t \sim \mathcal{N}\big[\mathbf{0}, \sigma_\nu\,\mathbf{I}\big]$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}\big[\mathbf{0}, \sigma_\varepsilon\,\mathbf{I}\big]$. This AR1 variant can be cast as a state space system, and can be solved utilizing Adaptive Least Squares (ALS):

$$\widehat{\mathbf{B}} = \Big[\mathbf{L}_{t-1}^{(0)}\Big]^{-1}\,\mathbf{L}_{t-1}^{(1)} \tag{2.7}$$

$$\widehat{\boldsymbol{z}}_t = \mathbf{z}_{t-1}\,\widehat{\mathbf{B}} \tag{2.8}$$

$$\mathbf{L}_t^{(0)} = \mathbf{L}_{t-1}^{(0)} + \mathrm{g}_t\,\big(\,\mathbf{z}_{t-1}^T\mathbf{z}_{t-1} - \mathbf{L}_{t-1}^{(0)} + \mathbf{I}\lambda\,\big) \tag{2.9}$$

$$\mathbf{L}_t^{(1)} = \mathbf{L}_{t-1}^{(1)} + \mathrm{g}_t\,\big(\,\mathbf{z}_{t-1}^T\mathbf{z}_t - \mathbf{L}_{t-1}^{(1)}\,\big) \tag{2.10}$$

$$\mathrm{g}_{t+1} = (\mathrm{g}_t + \rho)/(\mathrm{g}_t + \rho + 1) \tag{2.11}$$

with $\rho = \sigma_\nu/(\sigma_\nu + \sigma_\varepsilon)$. ALS (2.8)-(2.11) is a *stochastic descent* operation. "Stochastic" since the covariances, $\mathbf{L}^{(0)}$ and $\mathbf{L}^{(1)}$, are empirically estimated from noisy observations, "descent" since as we collect more and more data as t increases, the objective function decreases. For an attractive introduction into ALS and some filter variants, see the yet unpublished work, [McC05]. The superscript in $\mathbf{L}_t^{(0)}$ indicates that we are interested in an estimate of system covariance of observations separated by 0 time epochs. The covariance estimates are recursively updated with uniform adjustments to each element determined by the scalar gain, here notated as convention dictates with g (rather than K). It so happens that our gain update

(2.11) is equivalent to that given previously in (6.5) when the state is scalar, where $\rho = \sigma_\nu/(\sigma_\nu + \sigma_\varepsilon)$. This Kalman-like gain distinguishes this ALS method from the otherwise similar usual formulations of Recursive Least Squares (RLS), which instead updates covariance estimates using a constant *forgetting factor*. Suppose that the signal-to-noise ratio is zero, so that the gain becomes a harmonic sequence, $\left(1, \frac{1}{2}, ..., \frac{1}{t}\right)$, and the covariance estimates at time zero are set to zero. At any time t, the solution is the usual least squares solution with uniform weights. That is, $\mathbf{L}_t^{(0)} = t^{-1}\, \mathbf{Z}_{1:t}^T\, \mathbf{Z}_{1:t}$. The most well known version of RLS possesses the potentially desirable property that it does not require inversion of the regressor-regressor variance matrix, $\mathbf{L}_t^{(0)}$. Rather, it makes use of a Sherman-Woodbury-Morrison (SWM) update. For large datasets, say where the number of sites exceeds a few hundred, RLS offers an attractive alternative to (2.8)-(2.11). It lacks, though, the usual introduction of a regularizer. The importance of regularization has inspired clever investigations into uniting it with the computationally advantageous SWM RLS, [Gun96], [Gun94], [LYS99].

We can extend the process by including more lagged observations. Since the predictor and predictand force the choice of covariances to be estimated, we can write the more general form as

$$\widehat{\boldsymbol{z}}_t = \mathbf{z}^\diamond \left[\mathbf{L}_{t-1}^{\diamond\diamond}\right]^{-1} \mathbf{L}_{t-1}^{\diamond\star} \tag{2.12}$$

$$\mathbf{L}_t^{\diamond\diamond} = \mathbf{L}_{t-1}^{\diamond\diamond} + g_t \left(\mathbf{z}^{\diamond T}\mathbf{z}^\diamond - \mathbf{L}_{t-1}^{\diamond\diamond} + \mathbf{I}\lambda \right) \tag{2.13}$$

$$\mathbf{L}_t^{\diamond\star} = \mathbf{L}_{t-1}^{\diamond\star} + g_t \left(\mathbf{z}^{\diamond T}\mathbf{z}_t^\star - \mathbf{L}_{t-1}^{\diamond\star} \right) \tag{2.14}$$

$$g_{t+1} = (g_t + \rho)/(g_t + \rho + 1) \tag{2.15}$$

We always predict $\boldsymbol{z}_t$. The predictor, for example, may be $\mathbf{z}^\diamond = (\mathbf{z}_{t-3}, \mathbf{z}_{t-2}, \mathbf{z}_{t-1})$, or if a smoother is sought for imputation, $\mathbf{z}^\diamond = (\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1})$. In this general form the only thing open to decision is what observations are used to estimate the covariances. This decision only occurs when smoothing; in forecasting, our hands are tied. If forecasting to t, we are only permitted use of observations prior to t.

The appeal of ALS is its great parsimony. We can run ALS with as few as a single scalar hyperparameter. But how good a proxy is the time-varying parameter AR1 system (2.5)-(2.6) to the state space system (6.1)-(6.2) — or, to the point, under what conditions will ALS compete with an optimal Kalman Filter?

2.1.1 Relating Auto-Regressive ALS and the Kalman Filter

The rudimentary connection between ALS and the KF as applied to the presently considered state space system, (6.1) and (6.2), is most easily revealed through two equalities.

$$\mathbf{\Lambda}^{(0)} := \mathsf{E}\big[\boldsymbol{z}_t^T \boldsymbol{z}_t\big] = \mathbf{H}\,\mathbf{Q}\left(\mathbf{I} - \mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{H}^T + \mathbf{R} \tag{2.16}$$

$$\mathbf{\Lambda}^{(k)} := \mathsf{E}\big[\boldsymbol{z}_t^T \boldsymbol{z}_{t-k}\big] = \mathbf{H}\,\mathbf{Q}\,\mathbf{F}^k\left(\mathbf{I} - \mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{H}^T \tag{2.17}$$

with $k \in (1, 2, ..., t-1)$. Keeping our sights fixed on 1-step ahead forecasting to time t, we can rewrite (2.3)-(2.4) to obtain

$$\mathbf{\Lambda}^{\diamond\times\diamond} = \begin{pmatrix} \mathbf{\Lambda}^{(0)} & \mathbf{\Lambda}^{(1)} & \dots & \mathbf{\Lambda}^{(t-2)} \\ \mathbf{\Lambda}^{(1)} & \mathbf{\Lambda}^{(0)} & \dots & \mathbf{\Lambda}^{(t-3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Lambda}^{(t-2)} & \mathbf{\Lambda}^{(t-3)} & \dots & \mathbf{\Lambda}^{(0)} \end{pmatrix}, \qquad \mathbf{\Lambda}^{\diamond\times\star} = \begin{pmatrix} \mathbf{\Lambda}^{(t-1)} \\ \mathbf{\Lambda}^{(t-2)} \\ \vdots \\ \mathbf{\Lambda}^{(1)} \end{pmatrix} \tag{2.18}$$

and

$$\widetilde{\boldsymbol{z}}_t = (\mathbf{z}_1,\, \mathbf{z}_2, ...,\, \mathbf{z}_{t-1})\left[\mathbf{\Lambda}^{\diamond\times\diamond}\right]^{-1}\mathbf{\Lambda}^{\diamond\times\star} \tag{2.19}$$

and arrive at the following assertion.

Statement 1: The solution, $\widetilde{\boldsymbol{z}}_t$, given in (2.19), *is* the Kalman solution.

The argument is trivial.

Pf: Since the joint distribution of any arbitrary subsets of observations is Gaussian, by Guass-Markov we have that $\widetilde{\boldsymbol{z}}_t = \mathsf{E}\big[\boldsymbol{z}_t \big|\, \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ..., \mathbf{z}_1, \mathbf{H}, \mathbf{Q}, \mathbf{F}, \mathbf{R}\big]$. We also have the well-published equality for the Kalman Solution

$\widehat{\boldsymbol{z}}_t = \mathsf{E}\big[\boldsymbol{z}_t | \, \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ..., \mathbf{z}_1, \mathbf{H}, \mathbf{Q}, \mathbf{F}, \mathbf{R}, \mathbf{P_0} = \mathbf{Q}\big(\mathbf{I} - \mathbf{F}^T\mathbf{F}\big)^{-1}, \mathbf{b}_0 = \mathbf{0}\big]$. Since the expectation of an RV is unique, it must be the case that $\widetilde{\boldsymbol{z}}_t = \widehat{\boldsymbol{z}}_t$.

The upshot of all this is that the distinction between the Kalman solution and ALS is twofold. First, in ALS, we use $\mathbf{L}^{\diamond\diamond}$ and $\mathbf{L}^{\diamond\star}$ as proxies for $\mathbf{\Lambda}^{\diamond\diamond}$ and $\mathbf{\Lambda}^{\diamond\star}$ respectively, but, importantly, only through a usually small number of pre-specified lags (e.g., 1 or 2). Second, because we only estimate these covariances across one or two lags, we necessarily only use one or two time-lagged observations as our predictors.

We can gain some insight into the difference between the quality of the ALS and Kalman predictions using a single lag simplification of (2.19) as an intermediate.

$$\widetilde{\boldsymbol{z}}_t = \mathbf{z}_{t-1}\left[\mathbf{\Lambda}^{(0)}\right]^{-1}\mathbf{\Lambda}^{(1)} \tag{2.20}$$

where the covariances are defined as in (2.16) and (2.17). Notice that this solution is entirely deterministic; both covariances are fixed, and the predictors, once recorded, are treated as fixed.

If we condition on knowing $\mathbf{\Lambda}^{(0)}$ and $\mathbf{\Lambda}^{(1)}$, we can use the fundamental least squares relation $\sigma_e^2 = \mathsf{E}[(z-\mu_z)^2] - \mathsf{E}[(\widehat{z}-\mu_z)^2]$, since $\widehat{z} \perp z - \widehat{z}$, to derive the one-step-ahead forecast error. Without loss of generality, say that $\mu_z = 0$. Define

$$\mathbf{V} := \left[\mathbf{\Lambda}^{(0)}\right]^{-1}\mathbf{\Lambda}^{(1)} \tag{2.21}$$

our matrix of slopes that map $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$, for example, $\widehat{z}_{i,t} = \mathbf{z}_{t-1}\,\mathbf{v}_{\cdot,i}$, where $\mathbf{v}_{\cdot,i}$ is the ith column of $\mathbf{V}$. Notice that since $\mathsf{E}\left[\boldsymbol{z}_t^T\boldsymbol{z}_t\right] = \mathsf{E}\left[\boldsymbol{z}_{t-1}^T\boldsymbol{z}_{t-1}\right]$, $\mathbf{V}$ also serves as the matrix of partial correlations between $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t-1}$. Let $\widehat{e}_i$ be the observed one-step-ahead forecast error for site i. Our fitted variances for site i is

$$\mathbf{\Xi}_i = \mathbf{v}_{\cdot,i}\ \mathbf{v}_{\cdot,i}^T\ \circ \mathbf{\Lambda}^{(0)} \tag{2.22}$$

where $\circ$ represents the element-wise product. If we have $\mathbf{1}$ denote a column vector of 1's, we arrive at

$$\mathsf{E}\big[\widehat{e}_i^2\big] = \mathbf{\Lambda}^{(0)}_{[i,i]} - \mathbf{1}^T\,\mathbf{\Xi}_i\,\mathbf{1} \tag{2.23}$$

The result (2.23) offers a means by which we can assess how using only 1-lagged observations degrades the estimate. Firstly, one may compare this time-censored error with the error resulting from the full covariance structure, i.e., substituting $\mathbf{\Lambda}^{\diamond\diamond}$ and $\mathbf{\Lambda}^{\diamond\star}$ into (2.21)-(2.23). Or, equivalently comparing the error in (2.23) with that created from the steady-state value of $\mathbf{P}^{(-)}_{[i,i]}$ in the corresponding Kalman Filter.

In Chapter 5 we will detail an empiric counterpart to the above that allows for the creation of *standard errors* for $\widehat{\mathbf{B}}$, which will serve us later in §7.3.

2.2 ALS to Locate an Explicit System Mean Function

Suppose our observations arise through

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\nu}_t \tag{2.24}$$

$$\boldsymbol{z}_t^T = \mathbf{H}_t\,\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \tag{2.25}$$

with $\boldsymbol{\nu}_t \sim \mathcal{N}[\,\mathbf{0}, \mathbf{Q}_t\,]$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}[\,\mathbf{0}, \mathbf{R}_t\,]$. Our system, (2.24)-(2.25), may readily be cast as a time-varying parameter regression problem, where $\boldsymbol{\beta}_t$ is our $(p \times 1)$ parameter (i.e., partial slopes), and $\mathbf{H}_t$, $(n \times p)$, is our (known) regressor matrices (i.e., design matrices). We may now predict $\boldsymbol{z}$ through an estimate of $\boldsymbol{\beta}$. Suppose the general case where the observation errors are cross-correlated across time, e.g., $\mathbf{R}_t^{(1)} = \mathsf{Cov}[\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t-1}]$, and $\mathbf{R}_t^{(t-1)} = \mathsf{Cov}[\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_1]$, etc.

2.2.1 1-Step Ahead Forecast

We first have

$$\mathbf{\Sigma}^{\diamond\times\diamond} := \begin{pmatrix} \mathbf{R}_t^{(0)} & \mathbf{R}_1^{(t)} & \dots & \mathbf{R}_t^{(t-2)} \\ \mathbf{R}_t^{(1)} & \mathbf{R}_t^{(0)} & \dots & \mathbf{R}_t^{(t-3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_t^{(t-2)} & \mathbf{R}_t^{(t-3)} & \dots & \mathbf{R}_t^{(0)} \end{pmatrix}, \qquad \mathbf{\Sigma}^{\diamond\times\star} := \begin{pmatrix} \mathbf{R}_t^{(t-1)} \\ \mathbf{R}_t^{(t-2)} \\ \vdots \\ \mathbf{R}_t^{(1)} \end{pmatrix} \tag{2.26}$$

and

$$\boldsymbol{\zeta}_{1:t-1} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{t-1}) \tag{2.27}$$

If we create a vertically stacked regressor matrix

$$\mathbf{\Gamma}_{t-1} := \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_{t-1} \end{pmatrix} \tag{2.28}$$

The LS BLUP comes by way of

$$\mathbf{b} = \left(\mathbf{\Gamma}_{t-1}^T \mathbf{\Sigma}^{\diamond\times\diamond} \mathbf{\Gamma}_{t-1}^T\right)^{-1} \mathbf{\Gamma}_{t-1}^T \mathbf{\Sigma}^{\diamond\times\diamond} \boldsymbol{\zeta}_{1:t-1} \tag{2.29}$$

$$\widehat{\boldsymbol{\zeta}}_{1:t-1}^T = \mathbf{\Gamma}_{t-1}^T \mathbf{b} \tag{2.30}$$

$$\widehat{\boldsymbol{z}}_t = \mathbf{b}^T \mathbf{H}_t^T \tag{2.31}$$

$$\widetilde{\boldsymbol{z}}_t = \widehat{\boldsymbol{z}}_t + \left(\boldsymbol{\zeta}_{1:t-1} - \widehat{\boldsymbol{\zeta}}_{1:t-1}\right) \left[\mathbf{\Sigma}^{\diamond\times\diamond}\right]^{-1} \mathbf{\Sigma}^{\diamond\times\star} \tag{2.32}$$

The solution (2.29)-(2.32) is simply the generalized least squares solution conformed to our present system. This $O(n^3\tau^3)$ solution is, at the very best, computationally impractical.

Omitting details for the moment, we now propose a computationally easy approximation. For some easily computed Υ to be introduced explicitly in Chapter

4, and also implied in a dimensionally reduced space in Chapter 3, we solve *via*

$$\mathbf{b}_{t-1} = \Big[\mathbf{L}_{\mathrm{HH}\,t-1}\Big]^{-1} \mathbf{L}_{\mathrm{Hz}\,t-1} \tag{2.33}$$

$$\widehat{\boldsymbol{z}}_t^{\,T} = \mathbf{H}_t\, \mathbf{b}_{t-1} + \Upsilon_t \tag{2.34}$$

$$\mathbf{L}_{\mathrm{HH}\,t} = \mathbf{L}_{\mathrm{HH}\,t-1} + \mathrm{g}_t \left(\mathbf{H}_t^T \mathbf{H}_t - \mathbf{L}_{\mathrm{HH}\,t-1} + \mathbf{I}\lambda \right) \tag{2.35}$$

$$\mathbf{L}_{\mathrm{Hz}\,t} = \mathbf{L}_{\mathrm{Hz}\,t-1} + \mathrm{g}_t \left(\mathbf{H}_t^T \mathbf{z}_t - \mathbf{L}_{\mathrm{Hz}\,t-1} \right) \tag{2.36}$$

$$\mathrm{g}_{t+1} = (\mathrm{g}_t + \rho)/(\mathrm{g}_t + \rho + 1) \tag{2.37}$$

being reminded that $\mathbf{L}_{\mathrm{HH}}$ is $(p \times p)$, and $\mathbf{L}_{\mathrm{Hz}}$, $(p \times 1)$.

2.3 ALS Hybrid

In §2.1 we view our predictors (i.e., those covariates used to create our solution partial slopes, $\widehat{\mathbf{B}}$, from(2.7)), as being provided by the response itself — it is *autoregressive* in the most literal sense. In the solution sequence (2.33)-(2.37), we make use of a generic predictor placeholder called $\mathbf{H}_t$. This latter solution sequence shall be that used for mean function estimation/prediction in WIDLAS. The term "Hybrid" in this Section title, serves as a reminder that this generic placeholder, $\mathbf{H}_t$, can be used to house *anything* that the researcher suspects may serve to predict each $\boldsymbol{z}_t$. For example, we might have $\mathbf{H}_t = \big[\mathbf{H}^{(\boldsymbol{\Omega},\mathbf{t})}, \mathbf{z}_{t-1}^T\big]$, where $\mathbf{H}^{(\boldsymbol{\Omega},\mathbf{t})}$ is some spacial, temporal, or *non-seperable* spacio-temporal transformation of our choosing [Efr99, Was06, ABS00]. Moreover, in "phyning," to be introduced in §7.1, where we simultaneous impute and fit across multiple contemporaneous data sets, $\mathbf{H}_t$, will contain exogenous responses.

CHAPTER 3

Fixed Rank Kriging with Adaptive Least Squares (FRKALS)

FRKALS utilizes ALS to temporally extend Fixed Rank Kriging (FRK), advanced by Noel Cressie along with Gardar Johannesson [CJ08]. The inclusion of FRKALS in this present work serves two ends. One is simply to document its existence. The other is to provide an additional, computationally reduced method for the sake of comparison, upcoming in Chapter 6. Our treatment here will be terse, but sufficiently thorough. Science is a cauldron of irony. While the two computational sequences offered below look onerous, they are nothing more than the trivial result of inserting the dimensionally reduced objects described in [CJ08] into ALS.

3.1 Fixed Rank Kriging (FRK)

Much like this present work, FRK is proffered as a response to the growing need to analyze or predict scalar-valued observations in space where their number, n, is very large, say $n > 1000$. FRK bypasses the usual requirement of the inversion of the $n \times n$ predictor covariance matrix, $\mathbf{\Sigma}$. Cressie and Johannesson obviate this burden by developing three key insights. First, they have $\mathbf{\Sigma} \approx \mathbf{SKS}^T$, with $\mathbf{S}$ an $n \times r$ (with $r << n$) matrix containing distance information between the n sites and r fixed *centers*, and $\mathbf{K}$ an $r \times r$ matrix estimated empirically. Second, they bin up the observations, $\mathbf{z}$ into $\bar{\mathbf{z}}$, with length m, with $r < m < n$. And lastly

they invoke the Sherman-Morrison-Woodbury equality.

For the remainder of this Chapter we assume the reader possesses facility with FRK. We furthermore adapt the notation of Cressie and Johannesson: $\mathbf{T} = \mathbf{H}$, and $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Note also, that $\mathbf{W}$ is in no way related to our WIDALS weight $\mathbf{W}$ to be detailed in the next Chapter — we abide the authors' original notation as a matter of deference.

3.2 Uniting ALS and FRK

As is generally true with solutions to linear systems, estimators and predictors are functions of system covariances. So too with FRK. As discussed in Chapter 2, ALS provides a means of estimating a potentially time-varying linear function between a predictor and response through recursive adjusting of the predictor-predictor and predictor-response covariance matrices, [Hay02], [Say03], [McC05]. FRK allows for the natural insertion of ALS in two places. First, for the sake of estimating/predicting the mean function parameters, $\widehat{\boldsymbol{\alpha}}_t$, and second, estimating/predicting the binned de-trended response variance matrix, $\bar{\boldsymbol{\Sigma}}_{M\,t}$. Our proposed ALS extension is made yet more desirable because many FRK computations can be executed prior to the process loop, e.g., the creation of $\mathbf{S}$, $\mathbf{S}_0$, $\mathbf{W}$, as well as the QR decomposition of $\mathbf{S}$.

3.2.1 0-Step-Ahead Prediction

Here we suppose we have been charged with producing $\widehat{\mathbf{y}}_{0\,t} \mid \mathbf{Z}_{1:t}$. The following 11 assignments are recursed over time.

$$\widehat{\mathbf{L}}_{\mathrm{TT}\,t} = \widehat{\mathbf{L}}_{\mathrm{TT}\,t-1} + g_{1\,t} \cdot \left(\mathbf{T}_t^T \, \widehat{\mathbf{\Sigma}}_{t-1}^{-1} \, \mathbf{T}_t - \widehat{\mathbf{L}}_{\mathrm{TT}\,t-1} \right) \tag{3.1}$$

$$\widehat{\mathbf{L}}_{\mathrm{Tz}\,t} = \widehat{\mathbf{L}}_{\mathrm{Tz}\,t-1} + g_{1\,t} \cdot \left(\mathbf{T}_t^T \, \widehat{\mathbf{\Sigma}}_{t-1}^{-1} \, \mathbf{z}_t - \widehat{\mathbf{L}}_{\mathrm{Tz}\,t-1} \right) \tag{3.2}$$

$$\widehat{\boldsymbol{\alpha}}_t = \left(\widehat{\mathbf{L}}_{\mathrm{TT}\,t} + \lambda \mathbf{I} \right)^{-1} \widehat{\mathbf{L}}_{\mathrm{Tz}\,t} \tag{3.3}$$

$$\widetilde{\mathbf{z}}_t = \mathbf{z}_t - \mathbf{T}_t \, \widehat{\boldsymbol{\alpha}}_t \tag{3.4}$$

$$\bar{\mathbf{z}}_t = \mathbf{W}^T \, \widetilde{\mathbf{z}}_t \; * \boldsymbol{\gamma} \tag{3.5}$$

$$\bar{\mathbf{\Sigma}}_{M\,t} = \bar{\mathbf{\Sigma}}_{M\,t-1} + g_{2\,t} \cdot \left(\mathsf{Var}[\bar{\mathbf{z}}_t] - \bar{\mathbf{\Sigma}}_{M\,t-1} \right) \tag{3.6}$$

$$\widehat{\mathbf{K}}_t = \mathbf{R}^{-1} \mathbf{Q}^T \left(\bar{\mathbf{\Sigma}}_{M\,t} - \sigma^2 \bar{\mathbf{V}} \right) \mathbf{Q} \left(\mathbf{R}^{-1} \right)^T \tag{3.7}$$

$$\widehat{\mathbf{\Sigma}}_t^{-1} = (\sigma^2 \mathbf{V})^{-1} - (\sigma^2 \mathbf{V})^{-1} \mathbf{S} \, \left(\widehat{\mathbf{K}}_t^{-1} + \mathbf{S}^T \, (\sigma^2 \mathbf{V})^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^T \, (\sigma^2 \mathbf{V})^{-1} \tag{3.8}$$

$$g_{1\,t+1} = (g_{1\,t} + \rho_1)/(g_{1\,t} + \rho_1 + 1) \tag{3.9}$$

$$g_{2\,t+1} = (g_{2\,t} + \rho_2)/(g_{2\,t} + \rho_2 + 1) \tag{3.10}$$

$$\widehat{\mathbf{y}}_{0\,t} = \mathbf{T}_0 \, \widehat{\boldsymbol{\alpha}}_t + \mathbf{S}_0 \, \widehat{\mathbf{K}}_t \, \mathbf{S}^T \, \widehat{\mathbf{\Sigma}}_t^{-1} \, \widetilde{\mathbf{z}}_t \tag{3.11}$$

Recall that $\mathbf{T}_t$, $(n \times p)$, represents the known regressors.

Our process requires 3 scalar hyperparameters, $(\lambda, \rho_1, \rho_2)$. Equations (3.1) - (3.3) are ALS updates of the (deterministic) predictor-predictor variance matrix, i.e., $\mathbf{T}^T\mathbf{T}$, and the predictor-response covariance matrix, i.e., $\mathbf{T}^T\mathbf{z}$, respectively; (3.4) de-trends the response; (3.5) creates the binned response, with $\boldsymbol{\gamma}$ the reciprocal of the column sums of $\mathbf{W}$; (3.6) creates an ALS update of the binned response variance matrix. Finally, (3.9) and (3.10) update the update gains.

In absence of any prior conviction concerning the initial state of the three covariance matrices, a sensible way to commence the algorithm is to have $g_{1\,1} = g_{2\,1} = 1$, and $\widehat{\mathbf{L}}_{\mathrm{TT}\,0} = \mathbf{0}$, $\widehat{\mathbf{L}}_{\mathrm{Tz}\,0} = \mathbf{0}$, $\bar{\mathbf{\Sigma}}_{M\,0} = \mathbf{0}$, $\widehat{\mathbf{\Sigma}}_0^{-1} = \mathbf{I}$.

3.2.2 1-Step-Ahead Prediction

Here we create $\widehat{\mathbf{y}}_{0\,t} \mid \mathbf{Z}_{1:(t-1)}$.

$$\widehat{\mathbf{L}}_{\mathrm{TT}\,t} = \widehat{\mathbf{L}}_{\mathrm{TT}\,t-1} + g_{1\,t} \cdot \left(\mathbf{T}_t^T \, \widehat{\mathbf{\Sigma}}_{t-1}^{-1} \, \mathbf{T}_t - \widehat{\mathbf{L}}_{\mathrm{TT}\,t-1} \right) \tag{3.12}$$

$$\widehat{\mathbf{L}}_{\mathrm{Tz}\,t} = \widehat{\mathbf{L}}_{\mathrm{Tz}\,t-1} + g_{1\,t} \cdot \left(\mathbf{T}_t^T \, \widehat{\mathbf{\Sigma}}_{t-1}^{-1} \, \mathbf{z}_t - \widehat{\mathbf{L}}_{\mathrm{Tz}\,t-1} \right) \tag{3.13}$$

$$\widehat{\boldsymbol{\alpha}}_t = \left(\widehat{\mathbf{L}}_{\mathrm{TT}\,t} + \lambda \mathbf{I} \right)^{-1} \widehat{\mathbf{L}}_{\mathrm{Tz}\,t} \tag{3.14}$$

$$\widetilde{\mathbf{z}}_t = \mathbf{z}_t - \mathbf{T}_t \, \widehat{\boldsymbol{\alpha}}_t \tag{3.15}$$

$$\bar{\mathbf{z}}_t = \mathbf{W}^T \, \widetilde{\mathbf{z}}_t \; * \boldsymbol{\gamma} \tag{3.16}$$

$$\bar{\mathbf{\Sigma}}_{M\,t}^{(0)} = \bar{\mathbf{\Sigma}}_{M\,t-1}^{(0)} + g_{2\,t} \cdot \left(\mathsf{Var}[\bar{\mathbf{z}}_t] - \bar{\mathbf{\Sigma}}_{M\,t-1}^{(0)} \right) \tag{3.17}$$

$$\bar{\mathbf{\Sigma}}_{M\,t}^{(1)} = \bar{\mathbf{\Sigma}}_{M\,t-1}^{(1)} + g_{2\,t} \cdot \left(\mathsf{Cov}[\bar{\mathbf{z}}_{t-1}, \bar{\mathbf{z}}_t] - \bar{\mathbf{\Sigma}}_{M\,t-1}^{(1)} \right) \tag{3.18}$$

$$\widehat{\mathbf{K}}_t^{(0)} = \mathbf{R}^{-1} \mathbf{Q}^T \left(\bar{\mathbf{\Sigma}}_{M\,t}^{(0)} - \sigma^2 \bar{\mathbf{V}} \right) \mathbf{Q} \left(\mathbf{R}^{-1} \right)^T \tag{3.19}$$

$$\widehat{\mathbf{K}}_t^{(1)} = \mathbf{R}^{-1} \mathbf{Q}^T \, \bar{\mathbf{\Sigma}}_{M\,t}^{(1)} \, \mathbf{Q} \left(\mathbf{R}^{-1} \right)^T \tag{3.20}$$

$$\widehat{\mathbf{\Sigma}}_t^{-1} = \left(\sigma^2 \mathbf{V} \right)^{-1} \tag{3.21}$$

$$- \left(\sigma^2 \mathbf{V} \right)^{-1} \mathbf{S} \, \left(\left(\widehat{\mathbf{K}}_t^{(0)} \right)^{-1} + \mathbf{S}^T \left(\sigma^2 \mathbf{V} \right)^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^T \, \left(\sigma^2 \mathbf{V} \right)^{-1} \tag{3.22}$$

$$g_{1\,t+1} = (g_{1\,t} + \rho_1)/(g_{1\,t} + \rho_1 + 1) \tag{3.23}$$

$$g_{2\,t+1} = (g_{2\,t} + \rho_2)/(g_{2\,t} + \rho_2 + 1) \tag{3.24}$$

$$\widehat{\mathbf{y}}_{0\,t} = \mathbf{T}_{0\,t} \, \widehat{\boldsymbol{\alpha}}_{t-1} + \mathbf{S}_0 \, \widehat{\mathbf{K}}_{t-1}^{(1)} \, \mathbf{S}^T \, \widehat{\mathbf{\Sigma}}_{t-2}^{-1} \, \widetilde{\mathbf{z}}_{t-1} \tag{3.25}$$

This 1-step-ahead estimation/prediction process is distinguished from the 0-step-ahead process by the need to estimate response covariation across 1 time lag. We accomplish this through (3.18) and (3.20).

We may coax the notation here into that common elsewhere in this paper by having $\mathbf{C}^{\diamond\times\star} = \mathbf{C}_{t-1}^{(1)} = \mathbf{S}_0 \, \widehat{\mathbf{K}}_{t-1}^{(1)} \, \mathbf{S}^T$, $\mathbf{C}^{\diamond\times\diamond} = \mathbf{C}_{t-2}^{(0)} = \widehat{\mathbf{\Sigma}}_{t-2}$, $\mathbf{H}_t = \mathbf{T}_{0\,t}$, $\mathbf{b}_{t-1} = \widehat{\boldsymbol{\alpha}}_{t-1}$.

So now (3.25) can be written as

$$\widehat{\mathbf{y}}_{\,t} = \mathbf{H}_t \, \mathbf{b}_{\,t-1} + \mathbf{C}^{\diamond\times\star} \left[\mathbf{C}^{\diamond\times\diamond} \right]^{-1} \left(\mathbf{z}_{\,t-1} - \mathbf{H}_{t-1} \, \mathbf{b}_{\,t-1} \right) \tag{3.26}$$

which is spiritually identical to the GLS solution phrasing given in, e.g., (2.32). The second term in the RHS of (3.26) becomes the mysterious Υ alluded to in Chapter 2.

When the system arises according to (2.24) and (2.25) where the measurement shocks are uncorrelated across time, i.e., $\mathsf{E}[\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t-1}] = \mathbf{0}$, then $\mathsf{E}[\mathbf{C}^{\diamond\times\star}] = \mathbf{0}$, so the Υ adjustment is superfluous.

CHAPTER 4

Weighting by Inverse Distance with Adaptive Least Squares (WIDALS)

WIDALS solves a system by making use of the following recursive sequence:

$$\Upsilon_t = \widehat{\boldsymbol{\varepsilon}}^{\diamond}\, \mathbf{W} \tag{4.1}$$

$$\widehat{\boldsymbol{z}}_t^T = \mathbf{H}_t\, \mathbf{b}_{t-1} \tag{4.2}$$

$$\widetilde{\boldsymbol{z}}_t^T = \widehat{\boldsymbol{z}}_t^T + \Upsilon_t \tag{4.3}$$

where $\widehat{\boldsymbol{\varepsilon}}^{\diamond}$ is a truncated version of $\boldsymbol{\zeta}_{1:t-1} - \widehat{\boldsymbol{\zeta}}_{1:t-1}$ from (2.32), and where $\mathbf{b}_{t-1}$ is the ALS partial slopes estimate/prediction from (2.33)-(2.37). From GLS, the weighting function, $\mathbf{W}$, would be the familiar solution to the LS normal system of equations,

$$\mathbf{W} = \mathsf{E}\left[\widehat{\boldsymbol{\varepsilon}}^{\diamond T}\widehat{\boldsymbol{\varepsilon}}^{\diamond}\right]^{-1} \mathsf{E}\left[\widehat{\boldsymbol{\varepsilon}}^{\diamond T}\widehat{\boldsymbol{\varepsilon}}_t^{\star}\right] \tag{4.4}$$

Several authors have proposed computationally simplified means of producing this $\mathbf{W}$. Furrer, Genton, and Nychka [FGN06] make use of *tapering* — thresholding covariogram estimates of these two covariance objects by zeroing sufficiently small matrix entries, hence allowing for sparse matrix inversion. Thresholding in this way is a very natural approach, and occurs incidentally in the application of state-space solutions such as the Kalman Filter to natural phenomenon (e.g., climate, ecology, pollution, etc.), since these solutions commonly regard $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_{t-1}$ as independent, an assumption most likely untrue. So, a researcher applying such

a solution is at least tacitly thresholding temporal covariation. In fact, temporal tapering occurs in ALS as well. In the GLS formulation, $\mathbf{W}$ may be thought of as an expression of *conditional correlations*. While the broader theory describing the behavior of these conditional correlations is complex [FGN06], intuition and empirical research, e.g., [Ste02], can provide profound insight into the rather *local* nature of random fields even in systems possessing long range covariation. It is this very attribute that compels treatments such as tapering, or local neighborhoods, or even inverse distance weighting.

Usage of tapering to produce $\widetilde{\boldsymbol{z}}_t$ is an attractive frontier for future research; WIDALS, as we currently present it, is far less elegant. We have $\mathbf{W}$ be some weighting function as an inverse of distance, e.g., $\exp[-\alpha\,\mathbf{D}^{\diamond\times\star}]$.

Before proceeding, we will take a moment to advance some notation through example. Let us suppose we wish to forecast (to time t) to all locations, $\mathbf{s}$, one step ahead utilizing lag-1 and lag-2 information in the stochastic adjustment, Υ. We supply $\widehat{\boldsymbol{\varepsilon}}^{\diamond} = (\mathbf{z}_{t-2},\, \mathbf{z}_{t-1}) - (\widehat{\boldsymbol{z}}_{t-2},\, \widehat{\boldsymbol{z}}_{t-1})$. We construct distance over $\Omega \times T$ in the following way. Have $d^{\Omega}_{i,j}$ represent the i,j entry of our $(n \times n)$ spacial distance matrix, $\mathbf{D}^{\Omega}$.

$$d^{\Omega}_{i,j} = \sqrt{\Delta^2_{x_{i,j}} + \Delta^2_{y_{i,j}}} \tag{4.5}$$

Distance in time is constructed by

$$d^{(\Delta_t)}_{i,j} = \gamma \cdot \Delta_t \tag{4.6}$$

E.g., $d^{(0)}_{i,j} = 0$, and $d^{(1)}_{i,j} = \gamma$, and $d^{(2)}_{i,j} = 2 \cdot \gamma$, and so on. So that, in our lag $(-2, -1)$ example,

$$\mathbf{D}^{\diamond\times\star} = \sqrt{\begin{pmatrix} \mathbf{D}^{(2)} \\ \mathbf{D}^{(1)} \end{pmatrix}^2 + \begin{pmatrix} \mathbf{D}^{\Omega} \\ \mathbf{D}^{\Omega} \end{pmatrix}^2} \tag{4.7}$$

noting that $\mathbf{D}^{\diamond\times\star}$ is the desired $(2n \times n)$. Also note that each time distance matrix is constant.

For a second example, suppose we wish to retrospectively predict to a collection of $n^\star$ unmonitored sites, $\mathbf{s}^\star$ using lags $(-1, 0, 1)$. Then $\widehat{\boldsymbol{\varepsilon}}^\diamond = (\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}) - (\widehat{\boldsymbol{z}}_{t-1}, \widehat{\boldsymbol{z}}_t, \widehat{\boldsymbol{z}}_{t+1})$. As in the prior forecast example, the i counter from (4.5) will index over the n known locations, but j will index over the $n^\star$ new sites, $\mathbf{s}^\star$

In this present work, we have $\mathbf{W}$ be a function of exponential (negative) distance. Specifically,

$$\mathbf{A} = \exp[-\alpha\, \mathbf{D}^{\diamond\times\star}] \tag{4.8}$$

$$\mathbf{W} = \phi\ \mathbf{A}\, \mathbf{U}^{-1} \tag{4.9}$$

where $\mathbf{U}$ is a diagonal $n \times n$ matrix of column sums of $\mathbf{A}$; this matrix product (4.9) creates columns that sum to one. This is scaled by a scalar *flattening* hyperparamter, ϕ. Ultimately, then, the columns of $\mathbf{W}$ each sum to ϕ. Notice that this column-wise standardization of $\mathbf{W}$ acts as a WIDALS analogy to Ordinary Kriging.

4.0.3 The WIDALS Hyperparameter Space

All told, the WIDALS solution requires 5 scalar hyperparameters,
$\Pi^{\text{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\}$, with all values naturally being non-negative. The first two, ρ and λ are the ALS SNR and regularizer respectively. Distance decay of the mutual influence between sites is determined by α. Increasing alpha has the direct computational effect of removing the influence of distant site response values on the final estimated response value of another site. Finally, γ is our temporal distance scale. Suppose, for example, two sites, s_1 and s_2, are separated in space by 10 units. Now, thinking in terms of distance in space-time, the distance between $s_{1,t}$ and $s_{2,t}$ is still $\sqrt{10^2 + (0 \cdot \gamma)^2} = 10$ units. But the distance between, say, $s_{1,t-1}$ and $s_{2,t+1}$ (i.e., these same two sites "separated" in time by two epochs) is $\sqrt{10^2 + (2 \cdot \gamma)^2}$ units.

4.1 Prediction Quality

In a strictly parametric context, imagining the generative system (2.24)-(2.25), where the errors are uncorrelated across time, suboptimality enters our WIDALS process in three places. The first two appear in the deterministic part (i.e., ALS). As discussed in §2.2, ALS, as described, excludes some response covariation in the production of $\mathbf{b}$. Moreover, to avoid inverting the error covariance matrix, which is at least $(n \times n)$, we have excluded it from the process. That is, in (2.35) we present $\mathbf{L}_{\mathrm{HH}\,t}$ as an estimator for $\mathbf{H}_t^T\mathbf{H}_t$ rather than $\mathbf{H}_t^T\Sigma_t^{-1}\mathbf{H}_t$; likewise, Σ_t^{-1} is absent in the creation of $\mathbf{L}_{\mathrm{Hz}\,t}$. Now, we have introduced suboptimality through the proxy function Υ. A model-specific grand theory concerning the unified contribution of these three error sources in the overall WIDALS prediction error is made particularly challenging since they will be cross-correlated. The reader is directed to Figure 6.6 (far left boxplots). Here we find strong evidence of WIDALS's *self correction*. No particular erudition is required of the reader here; unless these three WIDALS error sources are mutually correlated, we'd need a tarot reading to explain the otherwise inexplicably low RMSE.

4.1.1 Miss IDed W, Theory and Examples

For the following suppose we are predicting a single response, $\widetilde{z}$. Suppose $\mathbf{\Sigma} = C(\mathbf{D}^{\diamond\diamond})$ is the true measurement-space error variance, and $\mathbf{C} = \widehat{C}(\mathbf{D}^{\diamond\diamond})$ is some estimate of $\mathbf{\Sigma}$. Have $\boldsymbol{\sigma} = C(\mathbf{d}^{\diamond\star})$ and $\mathbf{c} = \widehat{C}(\mathbf{d}^{\diamond\star})$. The prediction error attributable to the stochastic adjustment comes by way of

$$\mathsf{MSE}[\,\widetilde{\boldsymbol{z}}] = C(0) - 2\,\mathbf{c}^{\,T}\mathbf{C}^{-1}\,\boldsymbol{\sigma} + \mathbf{c}^{\,T}\mathbf{C}^{-1}\mathbf{\Sigma}\,\mathbf{C}^{-1}\,\mathbf{c} \tag{4.10}$$

(See [SH89])

Notice that when $\mathbf{C} = \mathbf{\Sigma}$,

$$\mathsf{MSE}[\widetilde{\boldsymbol{z}}] = C(0) - \mathbf{c}^{\,T}\mathbf{C}^{-1}\,\mathbf{c} \tag{4.11}$$

$$= C(0) - \boldsymbol{\sigma}^{\,T}\,\mathbf{\Sigma}^{-1}\boldsymbol{\sigma} \tag{4.12}$$

whose square root is recognized commonly as the *Kriging* prediction error [SH89], [FGN06].

These results provide for straight forward assessment of a candidate $\mathbf{w}^{\diamond\times\star}$ against some reckoned system covariance, since we may rewrite (4.10) as

$$\mathsf{MSE}[\widetilde{\boldsymbol{z}}] = C(0) - 2\,\mathbf{w}^{\diamond\times\star\,T}\,\boldsymbol{\sigma} + \mathbf{w}^{\diamond\times\star\,T}\,\mathbf{\Sigma}\,\mathbf{w}^{\diamond\times\star} \tag{4.13}$$

4.1.1.1 The WIDALS W

To get some sense of the quality of the WIDALS weighting matrix given in (4.9), we utilize an *infill* simulation. We apply (the square root of) (4.13) to both the WIDALS weights and truth (i.e., $\mathbf{\Sigma}^{-1}\boldsymbol{\sigma}$) to produce a cost ratio.

During our investigations we explored numerous infilling sequences, never did one speak particularly poorly of our WIDALS $\mathbf{W}$. Here we detail two scenarios from (about) opposite ends of the spectrum of performance. The most complementary is the "cross" infill. The "grid" infill is less complementary. For each we consider 9 different parameterizations of $C(\cdot)$ as a Matérn covariogram.

The Grid Infill

Here, we start with a square region in $[-1,1]\times[-1,1] \subset \mathbb{R}^2$, and position a site, $s^\star$, at the center (the origin) to which we wish to interpolate. We then manufacture a cohort of "monitored" sites by iteratively adding groups of 4, one on each half axis whose distance from the origin is in powers of 1/2. For example, the west points (on the negative x-axis), are located at $x = (-1/2^0, -1/2^1, -1/2^2, ..., -1/2^{35})$, $y = (0, 0, 0, ..., 0)$.

The respective RMSE ratios are shown in Figures 4.1-4.3. Like in the upcoming

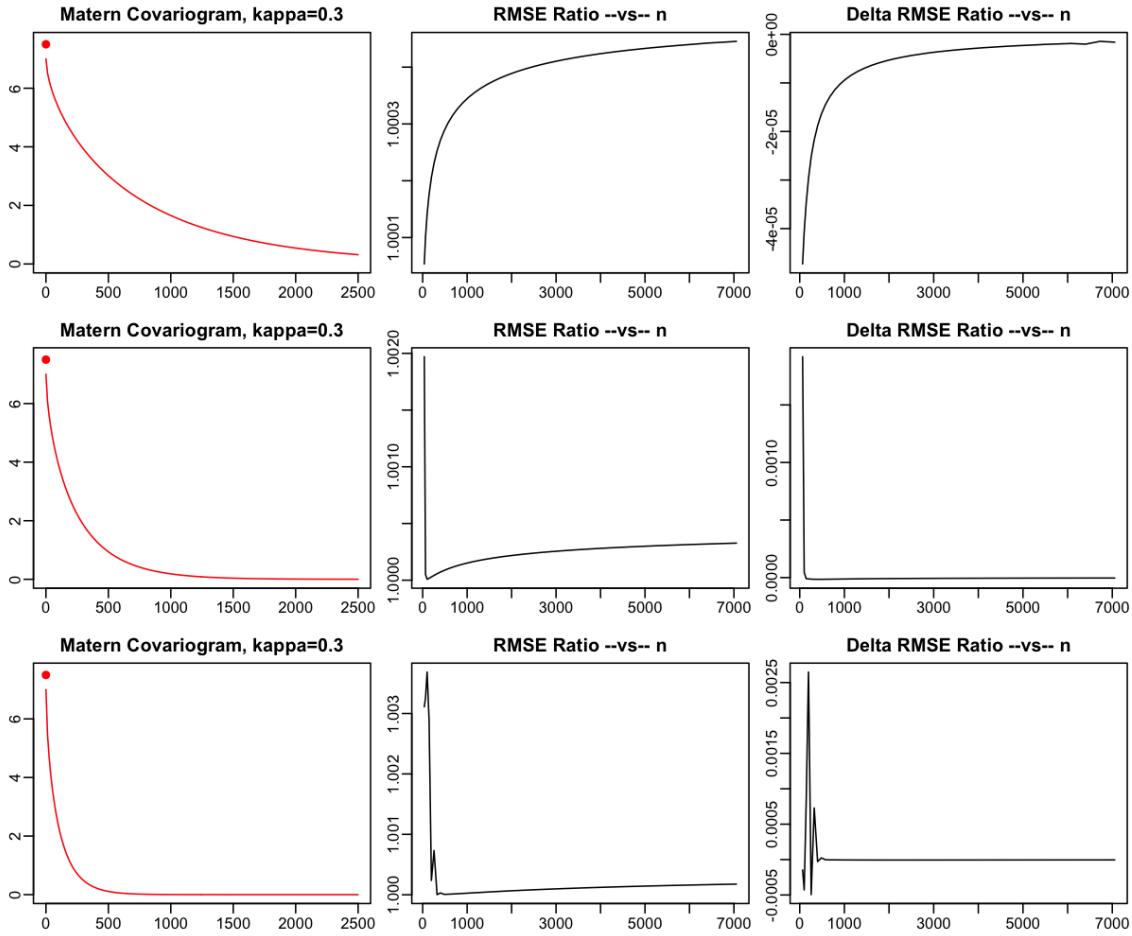

Figure 4.1: Grid Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 0.3$, "nugget" effect set to 0.5.

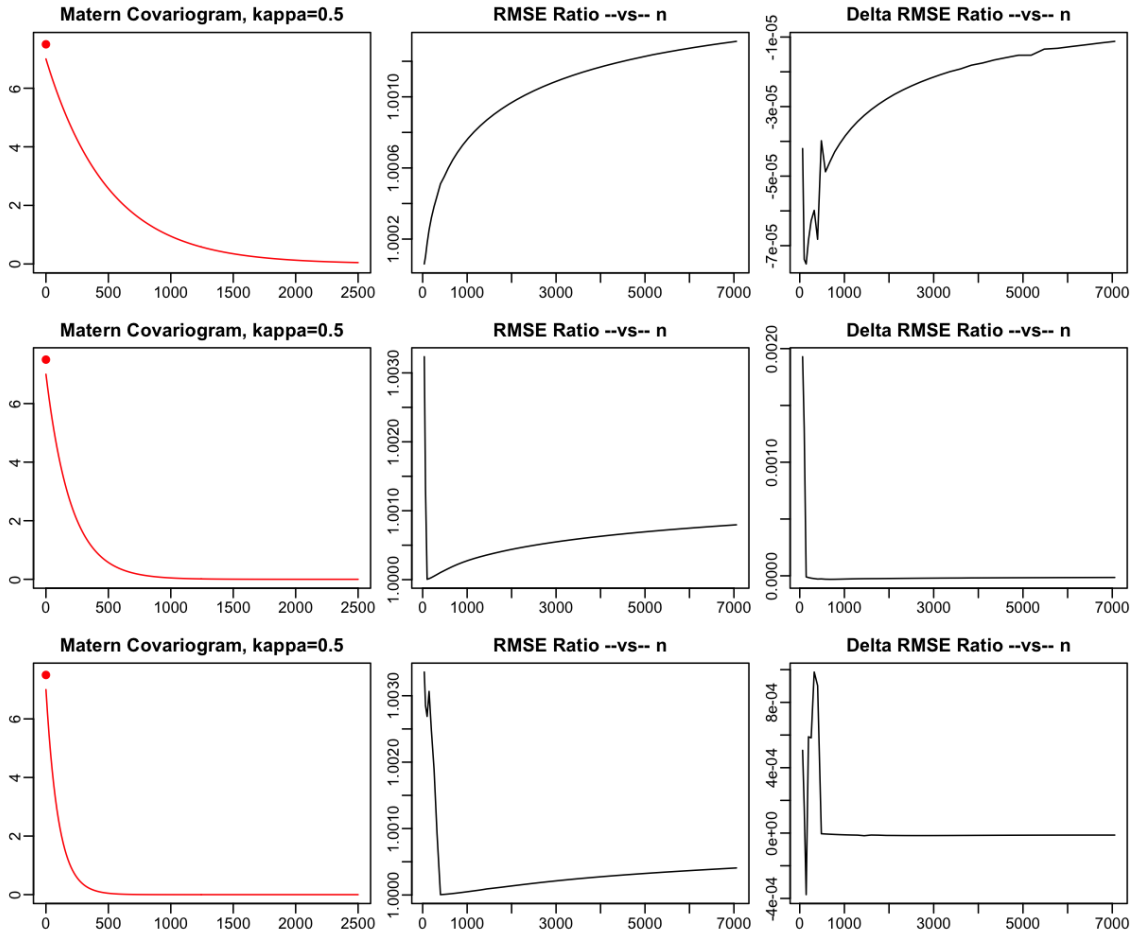

Figure 4.2: Grid Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 0.5$, "nugget" effect set to 0.5.

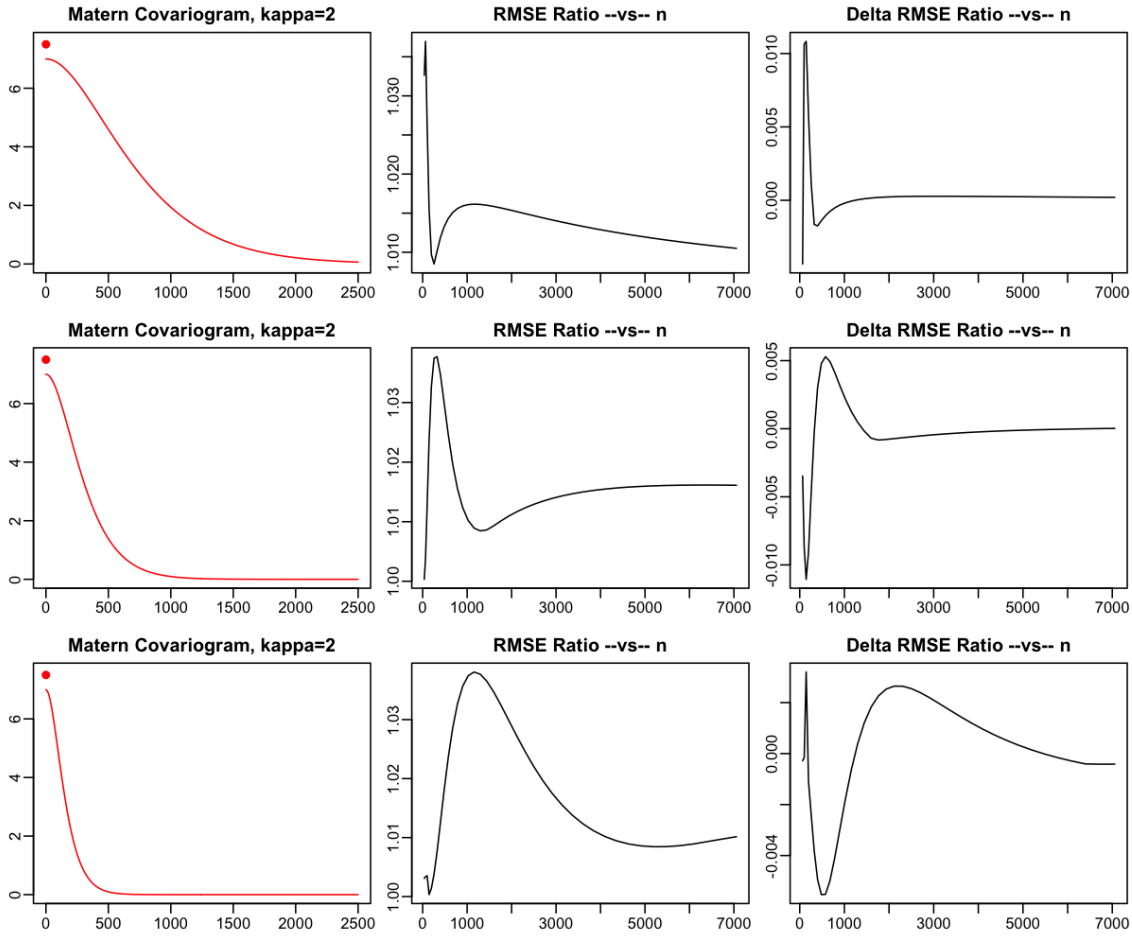

Figure 4.3: Grid Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 2$, "nugget" effect set to 0.5.

Cross infill, the poorest performance occurs when the shape parameter is large.

The rightmost plots in each panel show the change in the RMSE ratio as n increases. It is clear even upon brief inspection that our WIDALS $\mathbf{W}$ acts as an eerie doppelgänger for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$. The poorest RMSE ratio occurs when the Matérn shape paramenter is comparatively high ($\kappa = 2$), and the region is sparsely populated with supporting sites (about 20). In all 9 cases, the ratio runs rapidly to (near) unity as the region becomes more populated.

The Cross Infill

Again we start with a square region in $[-1, 1]\times[-1, 1] \subset \mathbb{R}^2$, and position a site, $s^\star$, at the origin. We then manufacture a cohort of evenly gridded "monitored" sites in our square region whose number abides the sequence $\{n_j : 2j \cdot 2j\}$ for $(3 \leq j \leq 42) \in \mathbb{N}$, i.e., $\{36, 64, 100, 144, ..., 7056\}$.

In Figures 4.4-4.6, the leftmost plots in each panel show the covariogram over distance (note that the corner-to-corner distance of our domain is 2.83). The ordinate scale of these plots is unimportant since we are ultimately only concerned with the cost ratio. The middle column plots in each panel show the RMSE of $\mathbf{W}$ divided by the ideal RMSE from (4.12) against number of sites, n, where we utilize Stochastic-Search to locate α and ϕ to produce minimum RMSE for $\mathbf{W}$ according to (4.13).

It should not be a grand revelation that the addition of a small nugget effect will, at least to a point, tend to pull the cost ratio to unity. But the rapid and rather precise convergence to unity in all 9 parameterizations of the Matern covariogram might be something of a surprise. The story, for our purposes, lives in the graphs, but said in brief, $\mathbf{W}$ provides a remarkable mimic of truth as our region becomes more densely clustered.

If this were the end of the story, one would never need invert a matrix for the sake of spacial modeling ever again. WIDALS, alas, is not all daisies and

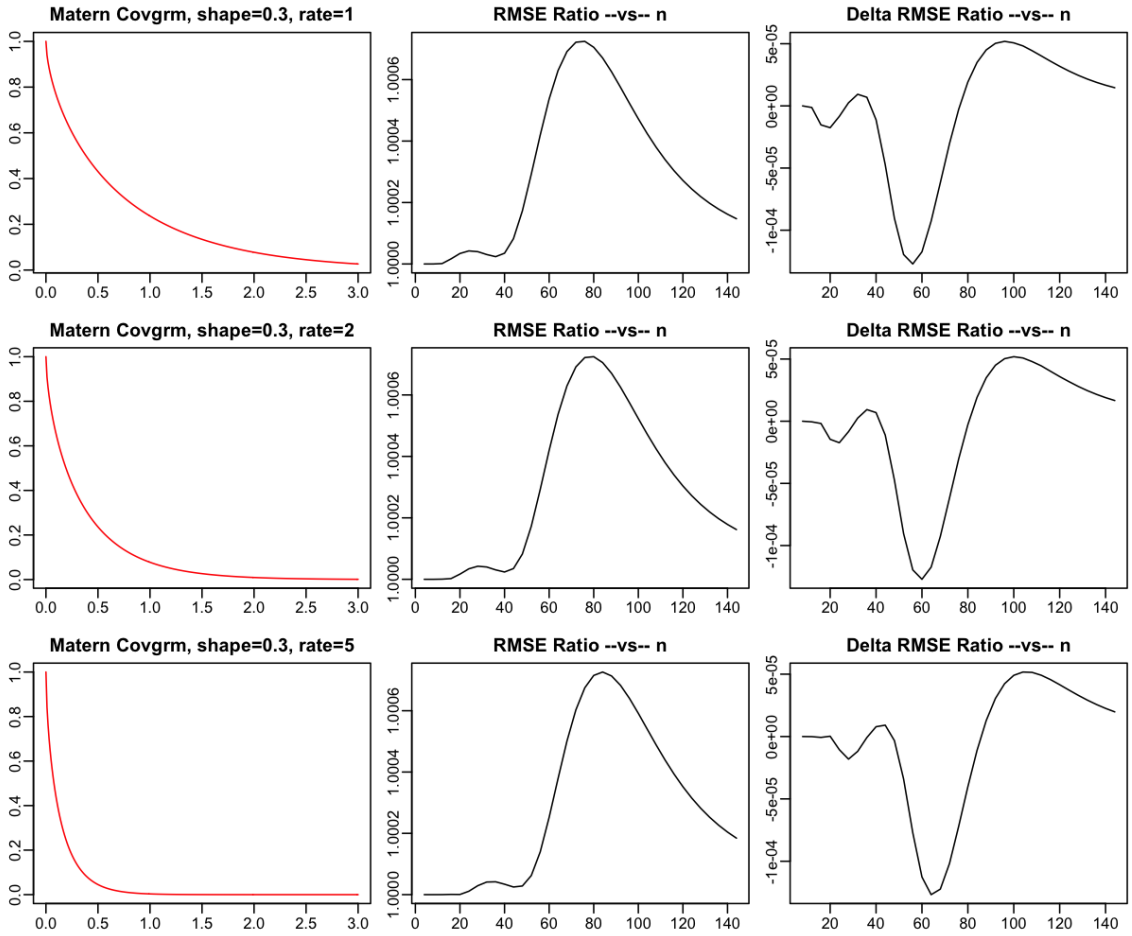

Figure 4.4: Cross Infill error of miss-specification implied by W for 3 covariogram rates, with shape parameter $\kappa = 0.3$, "nugget" effect set to 0.01.

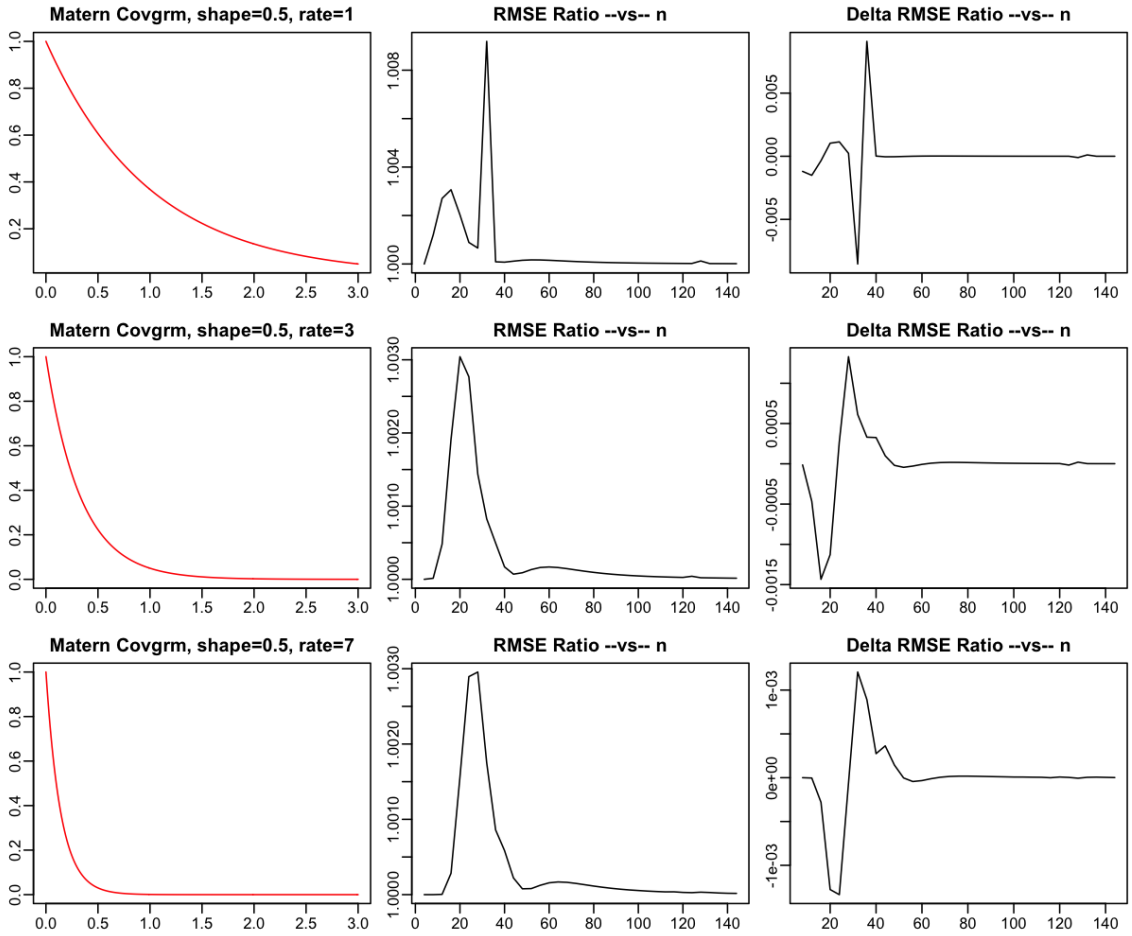

Figure 4.5: Cross Infill error of miss-specification implied by W for 3 covariogram rates, with shape parameter $\kappa = 0.5$, "nugget" effect set to 0.01.

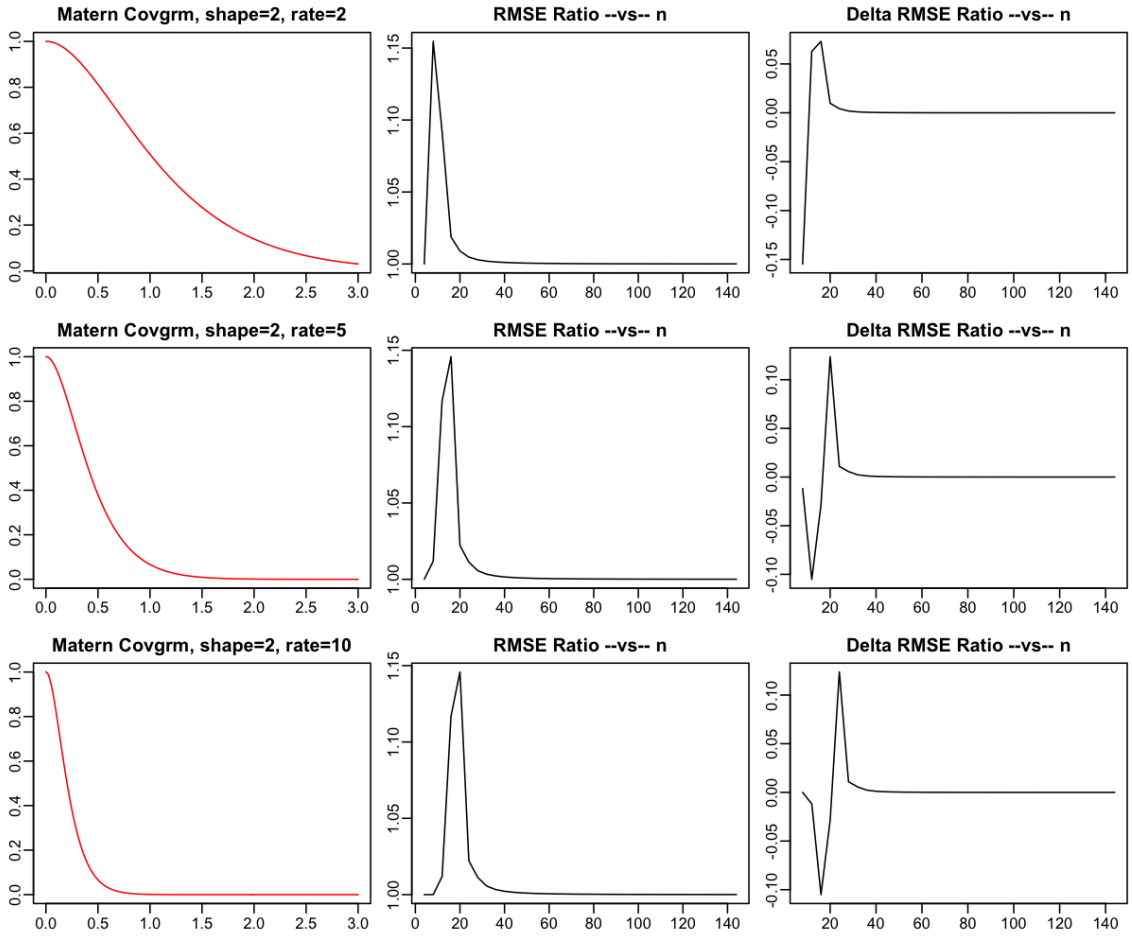

Figure 4.6: Cross Infill error of miss-specification implied by W for 3 covariogram rates, with shape parameter $\kappa = 2$, "nugget" effect set to 0.01.

rainbows. WIDALS has a dark side.

4.1.2 The Dark Side

In the previous infill scenario, we interpolated to a single site. In practice, a researcher will likely seek interpolating to many sites. In the current formulation we use a single rate parameter, α, in the construction of our WIDALS weights, $\mathbf{W}$ — we are stuck with a single weighting scale for every site to which we wish to interpolate. It so happens that this potential shortcoming can be greatly mitigated by an almost incidental property in the computation of (4.1). This will be discussed shortly in §4.1.3.2. Immediately following, in §6, we shall take WIDALS to the test by pitting it against canonical solutions in a more real-life-like setting.

4.1.3 More WIDALS Machinery

In addition to avoiding large matrix inversion, WIDALS makes available two other major computational dispensations.

4.1.3.1 Thresholding

Consider a large invertible matrix, $\mathbf{C}$. The computational burden of inverting $\mathbf{C}$ can be mitigated if it is sparse. This fact is particularly salient in our current context, since if the spacial covariogram decays rapidly with distance relative to the distances in the spacial domain, Ω, then many elements of $\mathbf{C}$ will be nearly zero. Elements that are sufficiently small — smaller than some thresholding value — can be set to zero.

In most all applications, the normal system of equations will produce weights (partial slopes) that decay much more rapidly with distance than the covariogram that created them. Examples are easy to come by; one is provided, using 196 evenly gridded sites, in Figure 4.7.

This expectation of a rapid decay in the weight function is titillating for WIDALS, since this suggests that the last remaining computational hurdle, the large matrix product from (4.1), may be greatly simplified.

4.1.3.2 Monitored Site-wise Execution

Now, with a careful look at (4.1)-(4.3), and (4.8)-(4.9), we propose the following computational sequence.

For $j \in 1:n$.

Calculate the jth column of $\mathbf{D}^{\diamond\times\star}$ from (4.7), call this vector $\mathbf{d}_j^{\diamond\times\star}$.

Use $\mathbf{d}_j^{\diamond\times\star}$ to calculate the $j$the column of $\mathbf{A}$ using (4.8), call this vector $\mathbf{a}_j^{\diamond\times\star}$.

Divide the elements of $\mathbf{a}_j^{\diamond\times\star}$ by its sum, multiply each element by the flattening parameter, ϕ. Call the resulting vector $\mathbf{w}_j$.

Locate only those elements of $\mathbf{w}_j$ that are greater than some threshold — say, 10^{-16}, for example.

Finally, employing only those indices indicated in the prior step, perform the matrix-vector multiplication, $\widehat{\boldsymbol{\varepsilon}}^{\diamond}\,\mathbf{w}_j$ to yield the jth column of $\boldsymbol{\Upsilon}_t$

End sequence.

In this way, additionally, **we bypass the need for memory storage of any full distance matrix, its concomitant covariance matrix or its inverse**.

Finally, we had mentioned in the previous Section that we employ the same WIDALS weight rate hyperparameter, α, for every site to which we predict. It so happens that the above computational sequence permits the usage of different α's for each new site. At the time of this writing this capability has only been yet minimally explored. Indeed, changing α does reduce prediction RMSE, but may wildly complicate model fitting. For the sake of this present submission, this variant of WIDALS will be regarded as a potential area of future investigation.

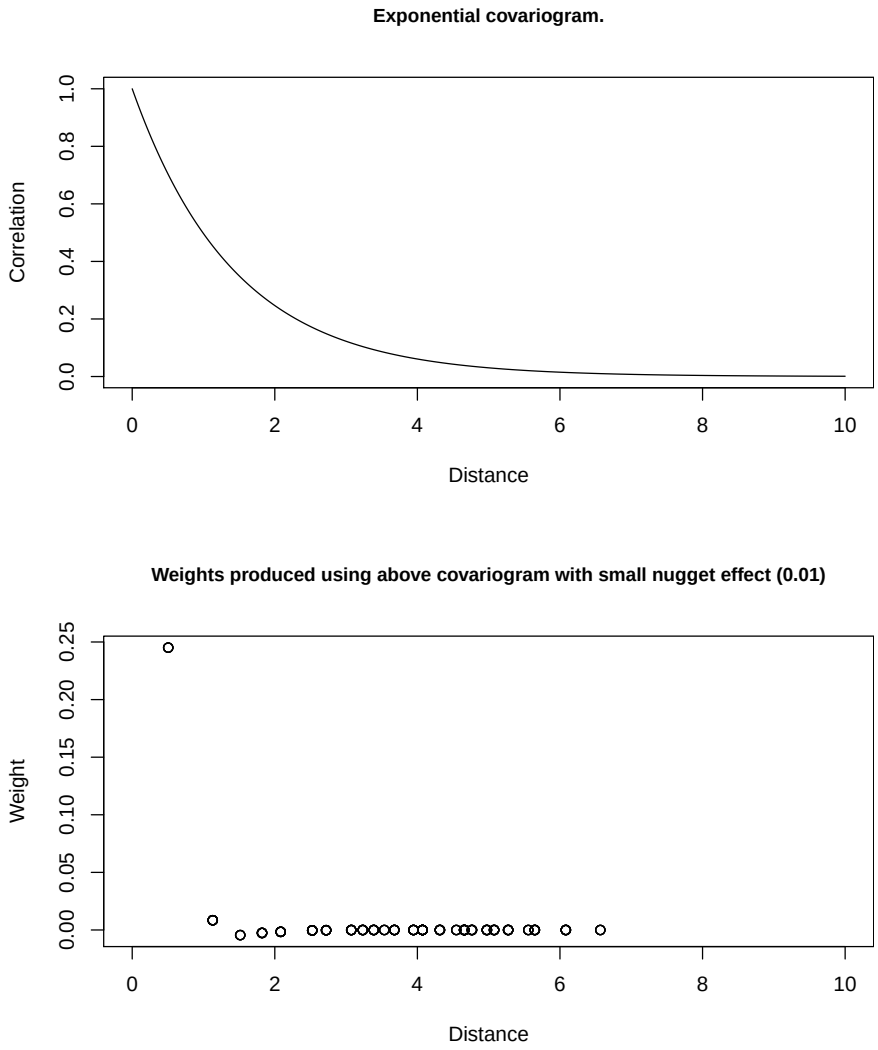

Figure 4.7: Covariogram and and covariogram solution weights as a function of distance. The weights decay much more rapidly than the covariogram that created them, making weight thresholding more enticing.

4.2 An Alternate View of W

We have posited that for some given covariogram over distance, $\mathbf{C} = C(\mathbf{D})$, that

$$\mathbf{W}^\star = \phi \exp[-\alpha \mathbf{D}] \approx \mathbf{C}^{-1}\mathbf{C}^\star \tag{4.14}$$

Earlier in this Chapter, we offered some simulated examples of the empiric cost associated with using $\mathbf{W}$ over $\mathbf{C}^{-1}\mathbf{C}^\star$. Here, we offer an alternate way of quantifying (4.14).

For simplicity, suppose we have a single interpolation site.

Suppose that

$$\mathbf{C}^{-1}\,\mathbf{c}^\star = \mathbf{w}^\star + \boldsymbol{\nu} \tag{4.15}$$

where $\boldsymbol{\nu}$ is a vector of errors. Then,

$$\mathbf{c}^\star = \mathbf{C}\mathbf{w}^\star + \mathbf{C}\boldsymbol{\nu} \tag{4.16}$$

We can think of the quantity $\mathbf{C}\mathbf{w}^\star$ as the correlation of the new site *implied* by $\mathbf{w}^\star$. Call this quantity $\mathbf{h}^\star$, and to ease notation, say $\boldsymbol{\eta} = \mathbf{C}\boldsymbol{\nu}$. We can then rewrite (4.16) as

$$\mathbf{c}^\star = \mathbf{h}^\star + \boldsymbol{\eta} \tag{4.17}$$

This implies that we can measure the proximity of $\mathbf{C}^{-1}\,\mathbf{c}^\star$ to $\mathbf{w}^\star$ by estimating the magnitude of $\mathbf{c}^\star - \mathbf{h}^\star$ using least squares (LS).

This approach offers two advantages, one is that we need no large matrix inversion. Second, we need locate only the WIDALS rate hyperparameter, α (the "flattening" hyperparameter, ϕ, precipitates from the LS solution). The key disadvantage is that this modeling approach is heuristic in the sense that while (4.13) measures error in response units, the LS solution of (4.16) measures correlation error. Nonetheless, a principal objective underlying the entirety of this submission

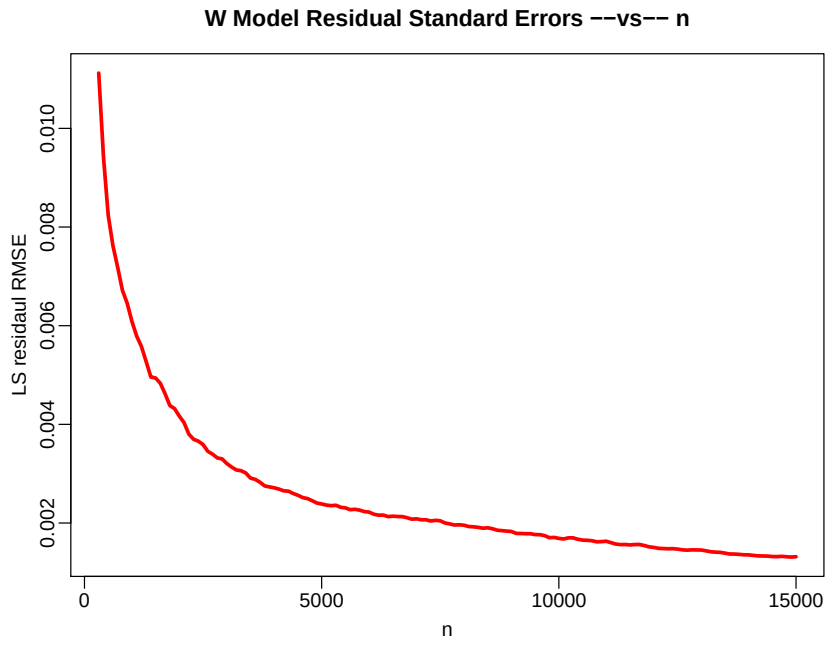

Figure 4.8: Infill $\mathbf{W}$ model error – unitless.

is that the WIDALS $\mathbf{W}$ can offer an acceptable approximation to the optimal normal system of equations solution. Figures 4.8 and 4.9, discussed presently, certainly promote this position.

We follow the general schema of our prior infilling scenarios. We define $[-1,1]\times[-1,1] \subset \mathbb{R}^2$. We place a single interpolation site at the origin. We have $\mathbf{C} = \exp[-2\mathbf{D}] + 0.01\,\mathbf{I}$. This time, however, we infill (100 supporting sites at a time up to 15,000) using the uniform density. For each addition, we locate α, fit LS, record the residual standard errors along with R^2. This is repeated 241 times, the standard errors and R^2's are averaged.

Figure 4.8 shows the LS error residuals, $n^{-1/2}\ \|\widehat{\boldsymbol{\eta}}\|$, as n grows. Figure 4.9 shows the amount of variance in $\mathbf{c}^\star$ explained its LS prediction, $\widehat{\mathbf{c}^\star}$. The figures suggest that under random uniform infilling with our chosen covariogram, $\mathbf{w}^\star$ becomes progressively more like $\mathbf{C}^{-1}\,\mathbf{c}^\star$.

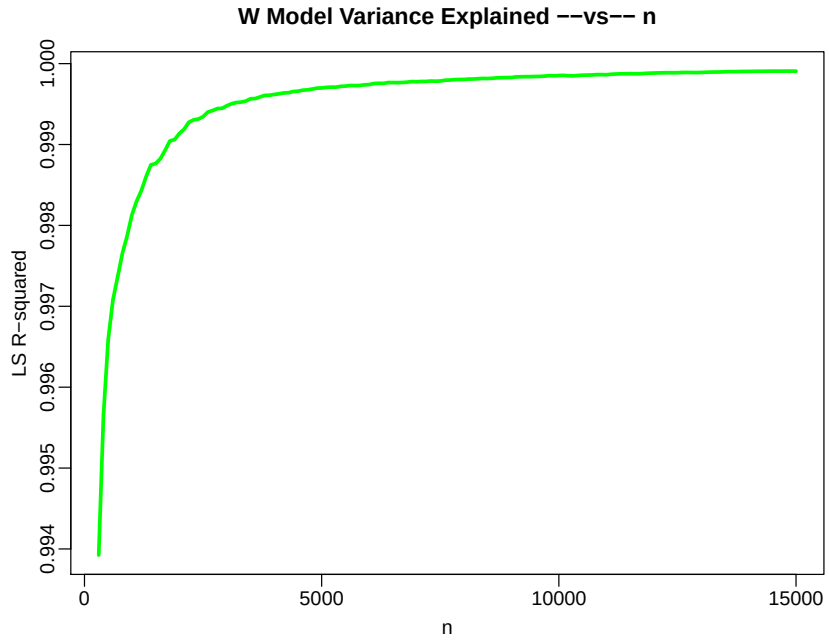

Figure 4.9: Infill fit quality of **W** model.

4.2.1 W versus Some Recent Methods

Contemporaneous with latter revisions of this document was the release of the excellent `R` package **LatticeKrig** by Nychka, Hammerling, Sain, and Lerud [NHS12]. While we have coded methods for Multi-Resolutional Spacial Models (MRSMs), **LatticeKrig** offers a ready invitation for comparison.

In [Ste02], Michael Stein offers "It is not clear how one could obtain useful and general results on the screening effect for any fixed set of observations." The context is the determination of the asymptotic properties of a Kriging predictor under arbitrary infilling. Let us take a moment to discuss, albeit heuristically, how spacial prediction of a fixed point is affected by progressively increasing supporting sites.

For the following, assume that our spacial support is some well-formed algebra providing measurability of distance between elements of an arbitrary subset.

Have $\mathbf{s}^{\diamond}$ be a collection of monitored sites in some bounded region of $\mathbf{\Omega} \subset \mathbb{R}^k$. Have $s^{\star}$ be a single unmonitored site residing in our region. Have $\boldsymbol{Z}(\mathbf{\Omega})$ be a zero-mean, Gaussian random field with known Covariogram, $C(\cdot)$.

The Kriging predictor, just like the GLS predictor will provide that $s^{\star} = s_1 \in \mathbf{s}^{\diamond} \Longrightarrow \widehat{z}(s^{\star}) = z(s_1)$.

While we certainly do not purport to advance Stein's baton in this paper, with this in mind, we can at least fashion a highly contrived scenario to serve as an intermediate to the larger challenge of obtaining "general results" for infilling. Suppose we locate the point in $\mathbf{s}^{\diamond}$ that is closest to $s^{\star}$, call that point s_1. Now suppose (holding all other sites in $\mathbf{s}^{\diamond}$ fixed) we slide s_1 progressively closer to $s^{\star}$ along path $\mathbf{p}^{s_1 \to s^{\star}}$.

If it happens to be the case that the function $f^{\mathbf{p}} = \left(z(\mathbf{p}^{s_1 \to s^{\star}}) - \widehat{z}(s^{\star})\right)^2$ is monotonic over $\mathbf{p}^{s_1 \to s^{\star}}$, then it must be non-increasing. In this sort of pseudo-infilling, we can invoke a simple *epsilon* demonstration of asymptotic optimality.[1]

Alas, it is unclear to me how one could fashion an analogous proof when $\mathbf{s}^{\diamond}$ is changing more generally.

While it is hoped that the previous discussion has some didactic value, it is presented at this particular point in this text to highlight an essential point in preparation for the following comparisons. Often prediction techniques that make use of spacial decomposition, including Fixed Rank Kriging (as in **LatticeKrig**), and wavelet transformations, offer the researcher the power to *zoom in* on the spacial domain for the sake of describing covariance structure and hence interpolating. This property, coupled with the fact that in the following comparisons $\mathbf{s}$ is fixed, means that these presently upcoming comparisons cannot be viewed as *competitive* in the usual sense of the word.

[1] One can be less sophomoric here and attempt to introduce the existence of some monotonic function, $g^{\mathbf{p}}$, such that $g^{\mathbf{p}} \geq f^{\mathbf{p}}$ over $\mathbf{p}^{s_1 \to s^{\star}}$.

4.2.1.1 MRSM: Fixed Grid Infill

In an prior simulation in §4.1.1.1, we fixed a single interpolation point, $s^\star$, then infilled by having supporting sites encroach on $s^\star$ as a more and more finely resolved grid pattern. The effect was to minimize the so-called borrowing power of $s^\star$, since the distances between $s^\star$ and $s^\diamond$ were dissimilar to the mutual distances in $s^\diamond$.

Here, we create a more canonical infilling scheme. As before, we start with our region $[-1,1] \times [-1,1] \subset \mathbb{R}^2$. We create $\mathbf{s}$ as a 140×140 evenly spaced grid in our region. Using the Matérn function with the shape parameter, $\kappa = 0.5$, and the rate parameter, $\alpha = 2$, we simulate numerous realizations of $\boldsymbol{Z}(\mathbf{s})$.

For each of these realizations of $\boldsymbol{Z}(\mathbf{s})$, we randomly remove 100 sites to which to interpolate, then progressively, randomly add the remaining sites.

For each addition, we fit our WIDALS $\mathbf{W}$ to the supporting sites using stochastic search, then use the resulting WIDALS hyperparameters to interpolate to our 100 sites, $\mathbf{s}^\star$. The resulting RMSE is recorded.

We likewise fit the MRSM models provided in **LatticeKrig** (though using the function's reported likelihood). When fitting a single-level model, we need locate three fitting parameters; for the 2-level model, we locate five fitting parameters.

The resulting RMSE performance is presented in Figure 4.10. The WIDALS $\mathbf{W}$ appears to produce a lower RMSE than the level-2 Fixed Rank Filter.

4.2.1.2 MRSM: Uniform Random Infill

Again we start with our region $[-1,1] \times [-1,1] \subset \mathbb{R}^2$.

For each simulation we randomly locate 20,000 sites in our region. We again use the Matérn function with the shape parameter, $\kappa = 0.5$, and the rate parameter, $\alpha = 2$, to simulate a realization of $\boldsymbol{Z}(\mathbf{s})$.

For each of these realizations of $\boldsymbol{Z}(\mathbf{s})$, we randomly remove 100 sites to which to interpolate, then progressively, randomly add the remaining sites.

For each addition, we fit our WIDALS $\mathbf{W}$ as in the previous Section. The resulting RMSE is recorded.

We likewise fit the MRSM models provided in **LatticeKrig** as in the previous Section.

The resulting RMSE performance is presented in Figure 4.11. The WIDALS $\mathbf{W}$ appears to produce a lower RMSE than the level-2 Fixed Rank Filter.

4.2.1.3 Haar Wavelets: Fixed Grid Infill

The infilling scenario is identical to that used above to compare MSRM.

Level-J Haar decomposition produces a design matrix with $1 + 2 \cdot (2^{J+1} - 1) + (2^{J+1} - 1)^2 = 4^{J+1}$ columns.

The resulting RMSE performance is presented in Figure 4.12. The WIDALS $\mathbf{W}$ appears to produce a lower RMSE than the level-4 Haar decomposition.

4.2.1.4 Computation Time

To make one pass over 20,000 locations takes about 2.5 seconds with FRK level-1, about 6 seconds with FRK level-2, and about 10 seconds with $\mathbf{W}$.

Total fitting times were about comparable. With $\mathbf{W}$, we can expedite the search by fixing the flattening hyperparameter, ϕ to 1, meaning that we need locate only the $\mathbf{W}$ rate hyperparameter. Fitting in this way takes about 5 minutes. Fitting the FRK level-1 model (3 hyperparameters), also takes about 5 minutes. Fitting the FRK level-2 model (5 hyperparameters), takes about 10 minutes.

Researchers with intimate knowledge of FRK will likely find easy shortcuts, thus reducing fitting time. WIDALS, *vis-a-vis*, offers substantial computational

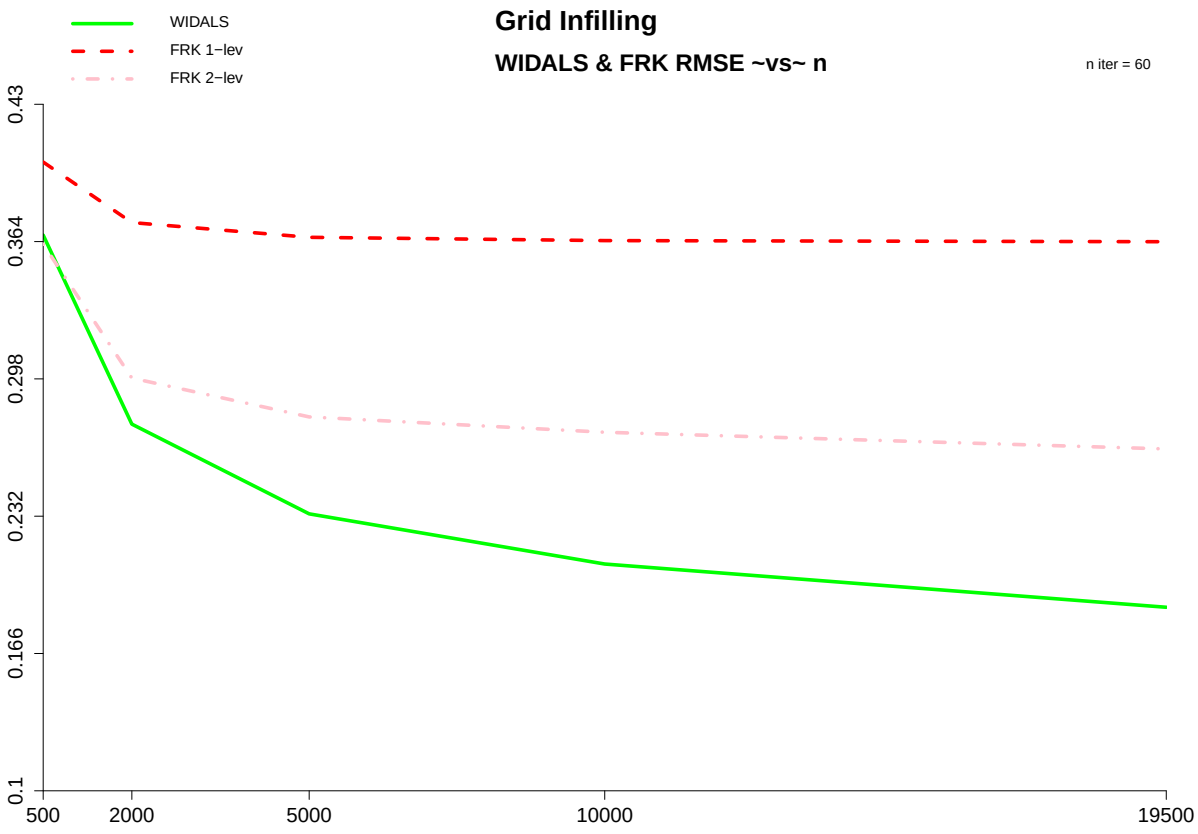

Figure 4.10: WIDALS **W** & FRK, Grid infilling.

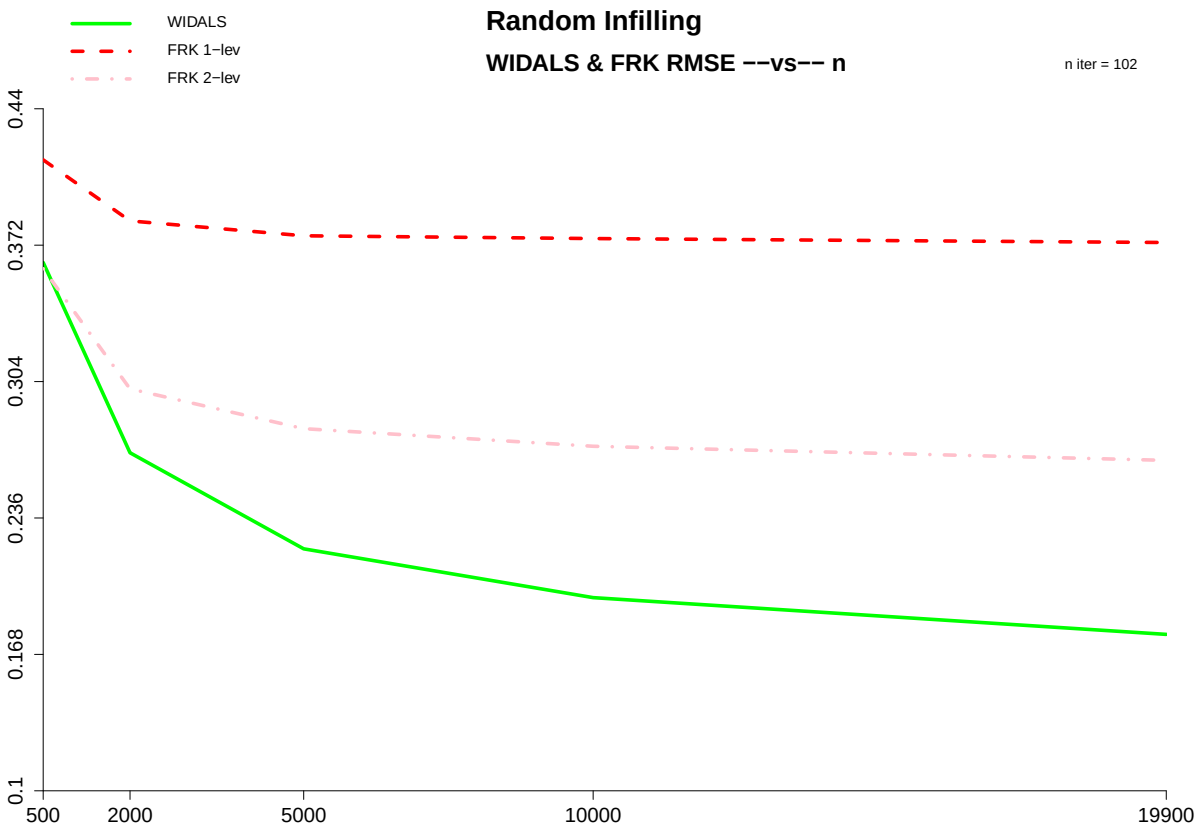

Figure 4.11: WIDALS **W** & FRK, Uniform random infilling.

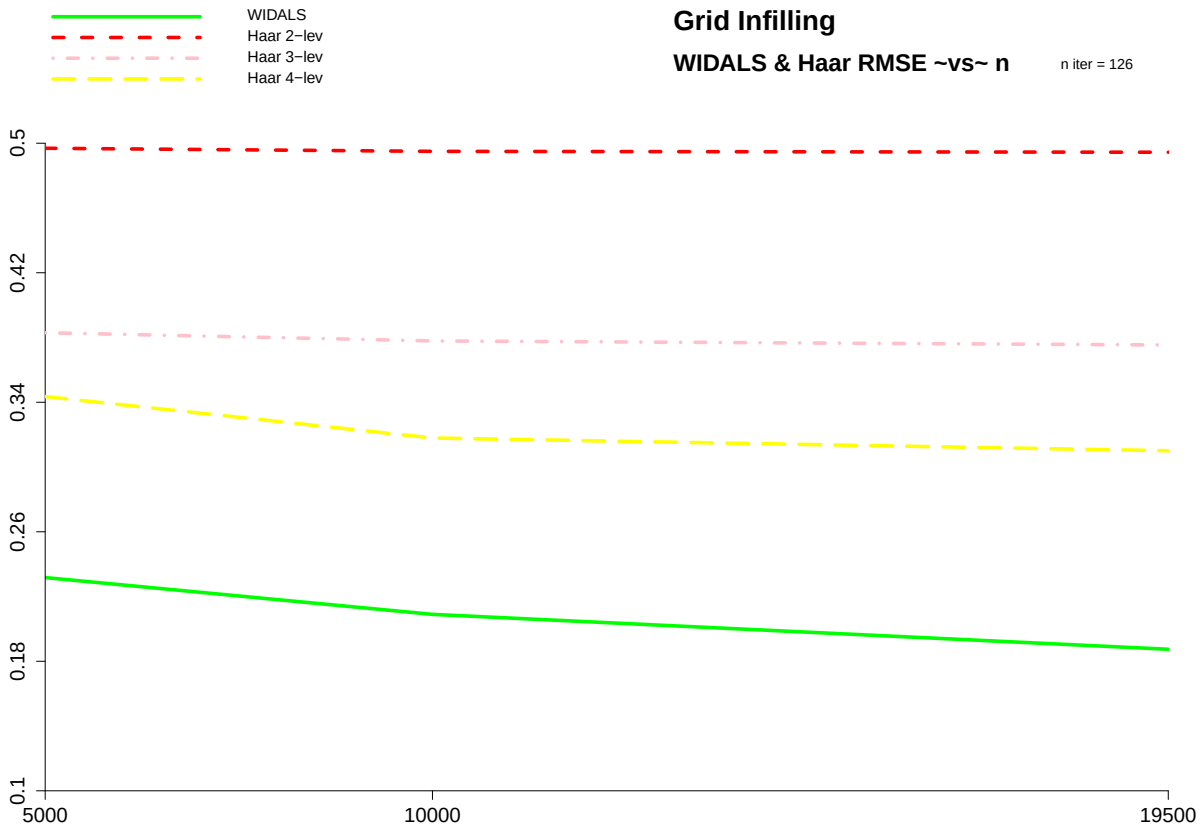

Figure 4.12: WIDALS **W** & Haar, Grid infilling.

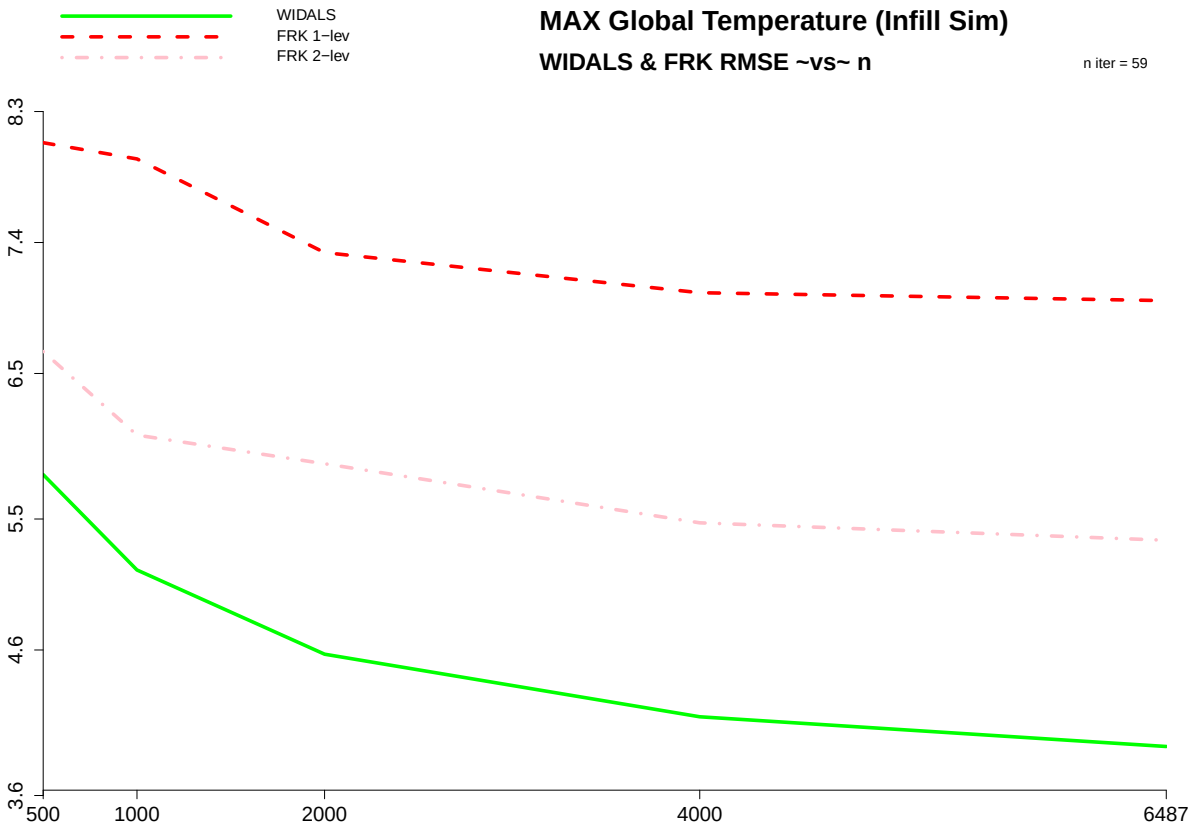

Figure 4.13: WIDALS **W** & FRK on demeaned, Global Daily Maximum Temperature (full data set).

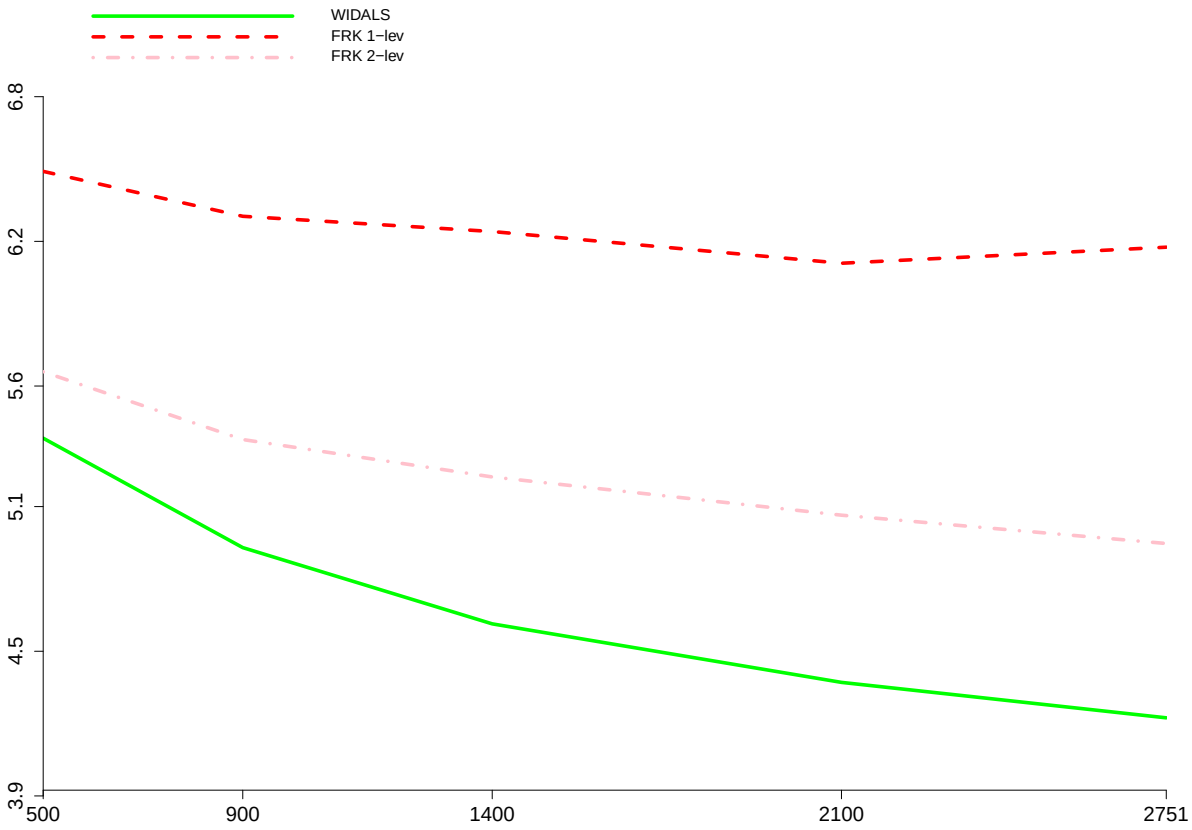

Figure 4.14: WIDALS **W** & FRK on demeaned, Global Daily Maximum Temperature (reduced data set).

mitigation when simultaneously fitting many realizations, §4.1.3.1.

4.2.1.5 MRSM: NCDC Maximum Daily Temperature

Lastly here, we compare FRK with the WIDALS **W** on real data. This data is described in some detail in §7.3, where we culminate this submission with a rather large-scale imputation-interpolation exercise.

For the current comparison, we use Global Maximum Daily Temperature for the date 2011-03-10 (this was one of the dates for which the fewest data were missing). The data were preprocessed by demeaning using the following covariates: a constant term, sensor Elevation, incident solar area (ISA), and the product of Elevation and ISA (*please see* (7.1)-(7.3) for calculation of ISA, and the adjacent text for further details).

We simulate infilling in precisely the way described above: 100 interpolation sites are randomly removed, then supporting sites randomly added, the resulting interpolation RMSE recorded. This is repeated, the RMSEs averaged. Mean RMSEs on are shown in Figure 4.13.

It is understood that **LatticeKrig** makes use of Euclidean distance. To mitigate distance distortion occurring near the Earth's poles, and to also eliminate the wrap-around effect of the sphere of the Earth, we utilized sensors whose latitude were between -50 and 50, and whose longitude was greater than -30, and less than 150 — that is, the Eastern Hemisphere. Resulting mean RMSEs for this reduced data set are shown in Figure 4.14.

CHAPTER 5

Model Fitting

5.1 Locating Hyperparameters, Generally

Not just with the simulations and examples presented here, but on hundreds of data sets, we have had great success using the Metahueristic Stochastic-Search (MSS) to locate our WIDALS hyperparameters, $\Pi^{\text{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\}$, detailed shortly in §5.4.2[1] Notice that all sensible interpretations of these values requires them to be positive: ρ is defined in ALS as a ratio of variances, λ is our covariance regularizer, $(-\alpha)$ is our distance coefficient operated on by the exponential function to create the WIDALS weights, γ is our temporal *distance*, and our weight flattening parameter, ϕ, is virtually always near 1 (in the absence of any spacial or spacio-temporal correlation, ϕ will tend to zero). For this reason, we have used the log-normal density as our search distribution.

While MSS is somewhat computationally inefficient compared to some other popular fitting algorithms, such as descent algorithms, it does possess a couple attractive features.

The biggest attraction of MSS is that it adapts trivially to handle an arbitrary cost function. This may have particular appeal in ecological prediction, since the *cost* of miss-predicting some measurement will vary on context. Regional attainment of airborne contaminates (ACs) is one example. Ozone levels for persons

[1]Actually, over the span of our research, we have made use of a number of search methods, including the more general Metropolis-Hastings, which has value in cases where the hyperparameters vary, say, by spacial location or time.

with sensitivities is another.

Another attraction of MSS is that it so easily fits into the parallel computing paradigm (local searches can be passed to threads).

Fitting has been accomplished using the function `MH.snow` in the `R` package **widals** [Zes12b] inside the parallel functionality provided by the package **snowfall**.

5.2 Model Selection: Selecting H

ALS, WIDALS, and FRKALS require known regressors, $\mathbf{H}$, for the estimation of the mean function, $\mathbf{Hb}$. Selection of covariates to include (as columns) in $\mathbf{H}$ for our purposes lends itself to techniques generally common to regression. Be reminded that our performance criterion involves testing over history by either forecasting to all locations at time t, or predicting at time t to locations deliberately removed. As such, testing is a form of cross-validation, so that one means (perhaps the primary means) of assessing candidate covariates is to simply select the $\mathbf{H}$ that minimizes testing cost. We augment this with a brief summary for the construction of *standard errors* for the resulting partial slopes, $\mathbf{b}$.

Of course, invoking standard errors implies inference, which in turn requires numerous model and sampling assumptions (see [Ber04] for a thorough treatment). For our purposes, the fact that we intend to deal with large n and potentially large τ, skewness in the response will likely not too greatly distort the usual interpretation of the resulting standard errors (i.e., equating the SE with the width of one standard deviation of a close-to-Gaussian sampling distribution of $\boldsymbol{b}$).

Consider the positive scalar ρ from (2.37). This entity can be thought of as (an estimate of) the signal-to-noise ratio (SNR) of the deterministic component of the generative system to be solved [McC05].

Suppose the SNR has been set to zero. Then the gain sequence, $(\mathrm{g}_0, \mathrm{g}_1, ...)$ becomes the harmonic sequence $\left(1, \frac{1}{2}, \frac{1}{3}, ...\right)$. In the mean function estimation case, at time t the slope estimate, $\mathbf{b} = \mathbf{L}_{\mathrm{HH}}^{-1}\ \mathbf{l}_{\mathrm{Hz}}$ is simply an evenly weighted least squares estimate. The so-called *effective sample size* of the process at time t, then, is $g_t^{-1} = t$. This same reasoning can be extended for any value of ρ [McC05]. If, say, $\mathrm{g}_t = 0.01$, then the observed covariances at time t are providing a weighted adjustment of 1/100 towards the current estimates of $\mathbf{L}_{\mathrm{HH}}^{-1}$ and $\mathbf{l}_{\mathrm{Hz}}$, and so the effective sample size at time t is 100.

Preferably, t_1 will be at some point beyond *steady state*. Just as in the general parlance of filtering, steady state refers to convergence of the gain, g. Just to get an idea, if $\rho = 1$, then $\mathrm{g} \to 0.618$ at about $t = 9$; if $\rho = 0.01$, then $\mathrm{g} \to 0.0951$ at about $t = 90$.

For example, to estimate standard errors across all n sites, define

$$\mathbf{\Pi}_t := \mathbf{L}_{\mathrm{HH}\,t}^{-1}\ , \qquad \widehat{\sigma}_t^2 := (t - t_1)^{-1}\, n^{-1} \sum_{k=t_1}^{t} \left\| \mathbf{z}_k - \widehat{\mathbf{z}}_k \right\| \tag{5.1}$$

So now a vector containing effective standard errors of each partial slope (with itself) can be created in the usual way.

$$\mathsf{SE}[\boldsymbol{b}_t] = \widehat{\sigma}_t \cdot \sqrt{\mathrm{g}_t} \cdot \sqrt{\mathsf{diag}[\mathbf{\Pi}_t]} \tag{5.2}$$

For example, suppose $\tau = 300$, $n = 5000$. For some candidate $\mathbf{H}$, we locate our hyperparamters according to §5.1 by testing over, say $t_1 = 100$ to τ. For any t_t in the testing range, we can observe how many respective standard errors away each of the p entries of $\mathbf{b}$ depart from zero.

5.3 Interpreting WIDALS Hyperparamers

Fitting data with WIDALS requires locating 4 or 5 hyperparameters, $\Pi^{\mathrm{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\}$ (note that γ serves no purpose when we predict using only lag-0

information). Table 5.1 distills their essence.

Name	Symbol	∼Domain	Meaning
Signal-to-noise ratio (SNR)	ρ	$[0, \infty]$	When $\rho = 0$, g_t becomes the harmonic sequence; when $\rho = \infty$, then $g_t = 1$ for all time.
Regularizer	λ	$(0, \infty)$	When λ is large, 1) response predictions are pulled towards zero, and 2) weakly contributing covariates become less influential in prediction.
Spacial rate	α	$(0, \infty)$	(Used to construct WIDALS weights, (4.8)) When α is small, responses at spacially distant locations covary; when α is large, (stochastic) covariation only occurs between sites close to one another.
Temporal distance constant	γ	$(0, \infty)$	When γ is small, responses at temporally distant locations covary; when γ is large, (stochastic) covariation is weak across time.
Weight coefficient	ϕ	$(0, 1]$	(Used to construct WIDALS weights, (4.9)) When $\phi = 0$, all temporal-spacial response covariation has been accounted for by the mean function, $\mathbf{H}_t\mathbf{b}_t$. When $\phi = 1$, system residuals, $\mathbf{z}_t - \mathbf{H}_t\mathbf{b}_t$, covary across space, or space-time.

Table 5.1: The role and domain of our 5 WIDALS hyperparameters.

5.3.1 A Little About λ, H

The engineering literature heralds λ, the *diagonal loader*, but often merely as a means of assuring — and maintaining numeric stability during — matrix inversion. In statistics, we view λ, the *regularizer*, as, for one thing, an ally in model parsimony, since turning λ up will suppress the magnitude of partial slopes of those covariates whose partial correlation among the other covariates is relatively small in magnitude. Both disciples trivially understand that (for any bounded

covariance matrix) as $\lambda \longrightarrow \infty$, $\|\mathbf{b}\| \longrightarrow \mathbf{0}$. This property has particular saliency in the context of working with space-time data. In this arena, the researcher's incredulity over model assumptions amplifies. For example, if one is imputing missing values, it is realistically too much to suspect that the values are missing at random; moreover, the very location of monitored sites is likely not random. So, if we center our response (making the global mean 0), and we additionally globally center $\mathbf{H}_t$, then increasing λ provides a means of *regressing* — in the Galtonian sense of regress — our predictions. That is, we can mitigate, and even eliminate, runaway predictions — predictions that stray from the response domain.

If we call these exogenous covariates for our monitored sites $\mathbf{H}_t^{\diamond}$, and those for the sites to which we wish to interpolate, $\mathbf{H}_t^{\star}$ (see the bottom of Table 1.1 for a reminder about the superscripts), we globally standardize $\mathbf{H}_t^{\diamond}$. Using these p means and p standard deviations, we standardize $\mathbf{H}_t^{\star}$. If online implementation is desired, these two objects can be easily standardized in real time, though the details are omitted here.

5.4 WIDALS Fitting, with an Example

At this point, let us solidify our discussion by walking through a scenario. Imagine we have in hand historical space-time data, $\mathbf{Z}$ $(\tau \times n)$. Each column represents data from a unique, fixed spacial location. Each row represents a time point. The larger goal may include predicting future values at unmonitored locations (called interpolation). However, as a precursor — and this virtually universal in statistical modeling in general — we need to *retrospectively* apply some mathematical solution or *model* to our data in order to, among other things, get some sense of the quality of our predictions (e.g., RMSE prediction error), decide what exogenous predictors (i.e., $\mathbf{H}_t$) will help reduce this prediction RMSE, and, of course, to locate the solution hyperparameters.

When interpolating, it is required that exogenous variables be available for new locations. In other words, if we want to interpolate low-level atmospheric Ozone at Fenway Park, and we've include Elevation as a predictor, then at some point we will need to acquire the elevation at Fenway Park. Of course, $\mathbf{H}_t$ can also house spacial or spacial-temporal transformations (e.g., radial bases expansions), which trivially accommodate arbitrary locations.

For the sake of simplicity, we shall assume the mean function is spacially constant, so that $\mathbf{H}_t = \mathbf{1}$ for all time. Also, let us assume there's no temporal correlation between our data, so that, from (4.7), $\mathbf{D}^{\diamond\star} = \mathbf{D}^{\Omega}$ is $n \times n$.

Since our ultimate goal is future interpolation, we need to retrospectively *simulate* "future" interpolation. Our modeling, then, will take the form of temporal training-testing with spacial-k-fold cross-validation. Suppose $n = 1000$, and $\tau = 200$. For this make-believe scenario, we shall set our temporal training range to be $t \in 30 : 180$, and our testing range to be $t \in 181 : 200$. We now randomly bin up our 1000 sites into 10 groups of 100.

Recall that our WIDALS tuning hyperparameters for this scenario are $\Pi^{\text{WID}} = \{\rho, \lambda, \alpha, \phi\}$ (we do not need γ since $\mathbf{D}^{\diamond\star} = \mathbf{D}^{\Omega}$). And, as we have throughout this submission, we assume these are time and space invariant, meaning that our full model can be fit with four scalar values.

Call $\mathbf{s}^{\star}_{(i)}$ our ith collection of 100 sites, and call $\mathbf{s}^{\diamond}_{(i)}$ the remaining 900.

In similar fashion, call $\mathbf{Z}^{\diamond}_{(i)}$ our data with the columns corresponding to $\mathbf{s}^{\star}_{(i)}$ removed.

Call $\mathbf{H}^{\star}_{(i)}$ our covariate matrix with rows corresponding to $\mathbf{s}^{\star}_{(i)}$; $\mathbf{H}^{\diamond}_{(i)}$ contains covariates associated with $\mathbf{s}^{\diamond}_{(i)}$.

5.4.1 Spacial-k-Fold Cross-Validation

We now outline the spacial-k-fold cross-validation routine:

For some fixed Π^{WID},

for each $i \in 1:k=1:10$

we use Π^{WID}, $\mathbf{Z}^{\diamond}_{(i)}$, $\mathbf{H}^{\diamond}_{(i)}$, and $\mathbf{H}^{\star}_{(i)}$ in (4.1)-(4.3) to "interpolate" $\mathbf{Z}^{\star}_{(i)}$

End i loop.

We now collect together all ten $\mathbf{Z}^{\star}_{(i)}$'s, and, for example, calculate the RMSE (i.e., the square-root of the mean square distance between these predictions and the actual data).

End routine.

5.4.2 Locating Π^{WID}, by Example

Recall for this scenario that we've set our training range to be $t \in 30:180$, so we can start by sub-setting our original data to include only rows 1 through 180.

Start with some initial value for Π^{WID}. For example, have $\Pi^{\text{WID}}_0 = \{1,1,1,1\}$. We now create a collection of candidate hyperparameters in the following way.

For each of these four values, we draw several random values from under a log-normal density, centered at the log of the current value (in this initial case, that's $\log[1] = 0$).

Each of these candidate hyperparameterizations are passed to the Spacial-k-Fold Cross-Validation routine described above. The one that results in the lowest *training* RMSE (i.e., the lowest RMSE over the training range), is kept, that is, used as centers for the next random draw of candidate values.

This is iterated until the training RMSE converges.

Finally, for *testing*, the Π^{WID} that minimizes the training RMSE is passed one last time to Spacial-k-Fold Cross-Validation, but this time the RMSE over the *testing range* is recorded.

It is common practice to have the standard deviation of the log-normal density

progressively shrink with more and more iterations.

5.4.3 An Example Using Simulated Data

For the sake of example, we borrow a simulated data set from our upcoming Chapter devoted to solution comparisons. We have $\mathbf{Z}$ be a realization of `sim-ss-1` (see §6.2.1 and Table 6.1). Sites are located on a 14×14 grid for a total of 196 sites. Training range is $t \in 50:250$, testing range is $t \in 251:297$. We use 14-fold validation for fitting.

We fit the simulated data using the previously described method; however, at every training RMSE improvement, we additionally record the prediction residuals over the testing range.

For the sake of illustration, we deliberately start with a poor choice for our initial hyperparameters: $\Pi_0^{\text{WID}} = \{0.001, 0.001, 0.01, 0.1\}$. Also, as has been common practice fitting WIDALS, each iteration is composed of two searches, one for the ALS hyperparameters, (ρ, λ), and one for the weights hyperparameters, (α, ϕ).

Figure 5.1 shows these testing residuals over actual data values for each of 24 training-RMSE-reducing fits. Early fits tended to under-estimate large response values, and over-estimate small response values. The diagnostic implication is that the mean function is being poorly tracked. This sort of residual pattern would occur, for example, in regression, if the true slope was 2, but the model was erroneously fitted with a slope estimate of 1.

Table 5.1 tracks the hyperparameter values along with the corresponding testing RMSE for each training-RMSE-reducing fit. Notice that, at least as far as the printed decimal precision permits, the testing RMSE is non-increasing (recall that we are fitting based on training RMSE). This suggests that Π^{WID} is relatively time-invariant.

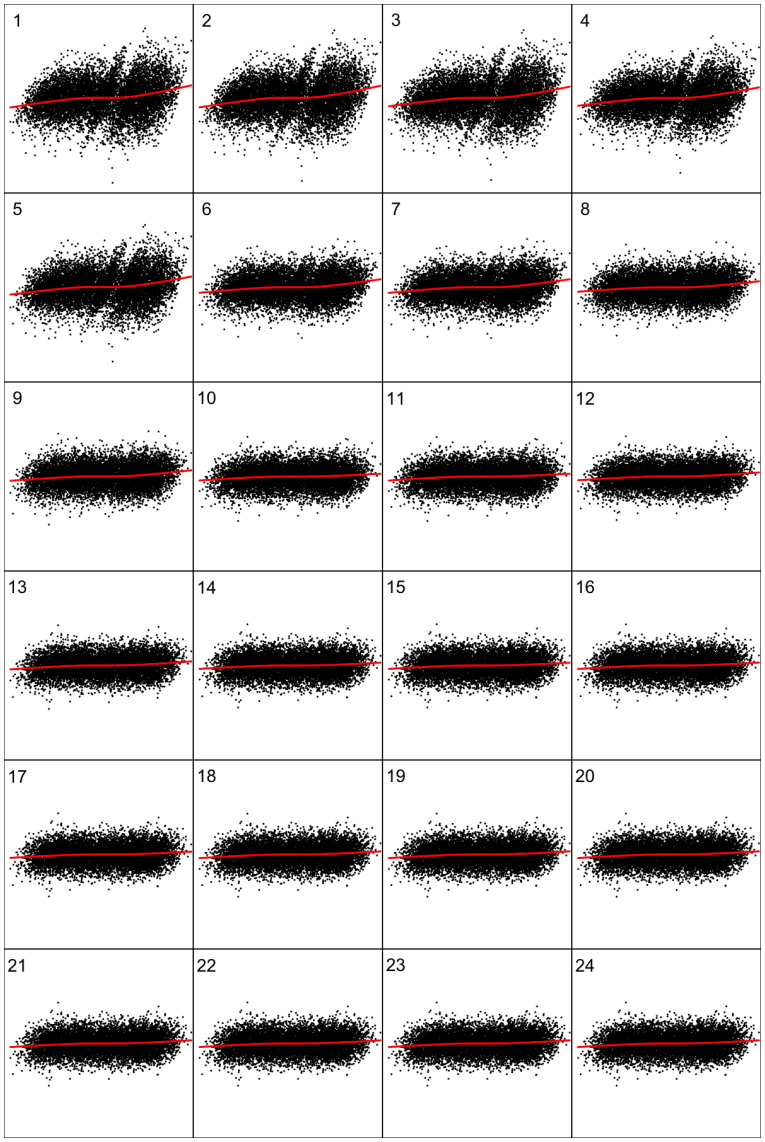

Figure 5.1: An Example: WIDALS residuals over actual values for progressive Stochastic-Search fits. Abscissa range is -25 to 22, ordinate range is -10 to 10. The red path is a lowess smoother.

Search	ρ	λ	α	ϕ	Testing RMSE
1	6.0720	0.0006	0.0100	0.1000	2.0102
2	6.0720	0.0006	0.3464	0.1159	1.9613
3	23.7775	0.0000	0.3464	0.1159	1.9457
4	23.7775	0.0000	8.6175	0.1512	1.8033
5	56.2187	0.0001	8.6175	0.1512	1.8024
6	56.2187	0.0001	23.8566	0.5103	1.4076
7	36.8107	0.0004	23.8566	0.5103	1.4076
8	36.8107	0.0004	23.2854	0.7374	1.2715
9	17.6185	0.0028	23.2854	0.7374	1.2713
10	17.6185	0.0028	1.0200	1.2304	1.1814
11	22.7377	0.0556	1.0200	1.2304	1.1813
12	22.7377	0.0556	2.6499	0.9649	1.1267
13	4.3601	0.0846	2.6499	0.9649	1.1246
14	4.3601	0.0846	2.2242	1.0824	1.1180
15	3.3597	0.1562	2.2242	1.0824	1.1180
16	3.3597	0.1562	2.1846	1.0789	1.1177
17	3.0787	0.1656	2.1846	1.0789	1.1177
18	3.0787	0.1656	2.1095	1.0598	1.1174
19	2.8240	0.1562	2.1095	1.0598	1.1173
20	2.8240	0.1562	2.1709	1.0559	1.1170
21	2.6999	0.1519	2.1709	1.0559	1.1169
22	2.6999	0.1519	2.2399	1.0565	1.1168
23	2.6857	0.1523	2.2399	1.0565	1.1168
24	2.6857	0.1523	2.2598	1.0535	1.1167

Table 5.2: WIDALS hyperparameters as located by stochastic search fitting.

CHAPTER 6

Prediction Performance Using Simulation

Simulation provides for comparison of cost, but more importantly offers an estimate of cost variation across different realizations of a particular system. Theoretical knowledge that solution "A" produces minimal expected cost is not comparable to empiric evidence that the cost produced by solution "A" only beats that produced by solution "B" 52% of the time.

We now compare RMSE performance of the Kalman Filter, ALS, FRKALS, and WIDALS on 6 systems. As is always the case with such empiric demonstrations, meaningful interpretation depends on several assumptions and certain particulars. Firstly, while the purpose of this submission is to promote WIDALS, generally, no special effort was made to hand pick systems to this end. However, application of the Kalman Filter to two of the systems (System 4 and System 5) involves gross misidentification — to a degree that the RMSE of the Kalman Filter cannot be interpreted as a likely consequence of a competent researcher; these two systems are meant to highlight the inherent ability of WIDALS to track temporally correlated measurement shocks. Application of the Kalman Solution to Systems 3 and 6 is far more believable, the degree of misidentification much more subtle.

6.1 Simulation

We generate data through the following state-space system:

$$\boldsymbol{\beta}_t = \mathbf{F}\,\boldsymbol{\beta}_{t-1} + \boldsymbol{\nu}_t \tag{6.1}$$

$$\boldsymbol{z}_t^T = \mathbf{H}_t\,\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \tag{6.2}$$

with $\boldsymbol{\nu}_t \sim \mathcal{N}[\,\mathbf{0}, \mathbf{Q}\,]$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}[\,\mathbf{0}, \mathbf{R}\,]$.

Our choice for our generative system for the sake of comparison offers ready interpretation, and also allows for comparison against the celebrated Kalman Filter.

6.1.1 The Kalman Filter

From an engineering perspective, there exist two exhaustive and authoritative texts covering the larger class of dynamic systems, [Say03] and [Hay02]. From a statistics perspective, the equally comprehensive and desirable [WH97] can be consulted, while [SS06] offer a compact discussion.

In statistics, the system (6.1) and (6.2) is known as a type of "time-varying parameters" model. Notice that this system looks much like a regression model where $\mathbf{H}$ plays the role of "design matrix," and $\boldsymbol{\beta}_t$, the partial slopes — or the model *parameters*.

Given $\{\mathbf{F}, \mathbf{Q}, \mathbf{H}_t, \mathbf{R}, \boldsymbol{\beta}_0\}$, the optimal $L2$ estimator for the latent state, $\boldsymbol{\beta}_t$, comes by way of the recursive process

$$\widehat{\boldsymbol{\beta}}_t^{(-)} = \mathbf{F}\,\widehat{\boldsymbol{\beta}}_{t-1}^{(+)} \tag{6.3}$$

$$\mathbf{P}_t^{(-)} = \mathbf{F}\mathbf{P}_{t-1}^{(+)}\mathbf{F}^T + \mathbf{Q} \tag{6.4}$$

$$\mathbf{K}_t = \mathbf{P}_t^{(-)}\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_t^{(-)}\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{6.5}$$

$$\mathbf{P}_t^{(+)} = \big(\mathbf{I} - \mathbf{K}_t\mathbf{H}\big)\mathbf{P}_t^{(-)} \tag{6.6}$$

$$\widehat{\boldsymbol{\beta}}_t^{(+)} = \widehat{\boldsymbol{\beta}}_t^{(-)} + \mathbf{K}_t\left(\boldsymbol{z}_t^T - \mathbf{H}\,\widehat{\boldsymbol{\beta}}_{t-1}\right) \tag{6.7}$$

an example of the Kalman Filter (KF). The symbol, $\widehat{\boldsymbol{\beta}}_t^{(-)}$, called the *a priori* estimator, is our prediction of the random variable $\boldsymbol{\beta}_t$ prior to observing $\mathbf{z}_t$. The symbol $\widehat{\boldsymbol{\beta}}_t^{(+)}$, called the *a posteriori* estimator, is our prediction of the random variable $\boldsymbol{\beta}_t$ after observing $\mathbf{z}_t$. Since $\mathbf{H}$ is linear, and since both $\boldsymbol{\nu}_t$ and $\boldsymbol{\varepsilon}_t$ are each uncorrelated across time, the best $L2$ estimator of a yet unmeasured $\boldsymbol{z}_t$ is $\mathbf{y}_t = \mathbf{H}\widehat{\boldsymbol{\beta}}_t^{(-)}$. Or equivalently,

$$\mathsf{E}[\, \boldsymbol{z}_t^T \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ..., \mathbf{z}_1 \,] = \mathbf{H}\widehat{\boldsymbol{\beta}}_t^{(-)} \tag{6.8}$$

Viewing the system from the state space,

$$\mathsf{E}[\, \boldsymbol{z}_t^T \mid \mathbf{z}_t, \mathbf{z}_{t-1}, ..., \mathbf{z}_1 \,] = \mathbf{H}\widehat{\boldsymbol{\beta}}_t^{(+)} \tag{6.9}$$

However, viewing the system from the measurement space, for a new location, $\omega^\star$,

$$\mathsf{E}\left[\, z_t(\omega^\star)^T \mid \mathbf{z}_t, \mathbf{z}_{t-1}, ..., \mathbf{z}_1 \,\right] = \mathbf{H}^\star \widehat{\boldsymbol{\beta}}_t^{(+)} + \left(\mathbf{z}_t - \mathbf{H}^\diamond \widehat{\boldsymbol{\beta}}_t^{(+)}\right) \left[\mathbf{R}^{\diamond\diamond}\right]^{-1} \mathbf{r}^{\diamond\star} \tag{6.10}$$

Notice that we have used random variables to describe the process (6.3)-(6.7), i.e., we use $\boldsymbol{z}_t$ and $\widehat{\boldsymbol{\beta}}_t$ rather than $\mathbf{z}_t$ and $\widehat{\boldsymbol{\beta}}_t$ respectively. This is done to highlight the fact that $\mathbf{P}_t$, the error matrix, a player with a lengthy storyline, is indifferent to whatever particular data realization $\mathbf{Z}$ of $\boldsymbol{Z}$ we process. Centrally,

$$\mathbf{P}_t^{(-)} = \mathsf{E}\left[\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(-)}\right)\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(-)}\right)^T\right] \tag{6.11}$$

and

$$\mathbf{P}_t^{(+)} = \mathsf{E}\left[\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(+)}\right)\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(+)}\right)^T\right] \tag{6.12}$$

measure the expected quality of our *a priori* and *a posteriori* predictions respectively. The inverse, $\mathbf{P}_t^{-1}$ may rightly be called the system's (or solution's) *precision* matrix.

6.2 Comparisons

While sometimes data arise through a known, well-defined system, data in nature usually do not. If we wish to forecast or predict, say, carbon monoxide concentrations over Boston from historical data, we do not choose a state-space model with the belief that we need only locate the correct hyperparameters to achieve minimal prediction error. Rather, we choose our model believing that it defines a rich enough class of relationships that truth and our fitted model will be close enough so that we may in some way profit from our effort. Empiric comparison between the performance of candidate solutions against ground truth offers a nice window by which we might view this profitability, not only because we can compare a typical cost between solutions, but because we can do so relative to the variation in cost associated with multiple realizations from the same system. Additionally, comparisons are often used as a forum by which different solutions may compete. Implicit in such presentations is that the constructed data in some way emulate something akin to a real dataset with which the audience may at some point come into contact. Certainly, the logic goes, if solution A is easier to implement than solution B, and provides similar quality predictions, the reader should turn to solution A in real life. This is, of course, contentious on its face, since there will always be some lack of consensus over what fundamental assumptions are reasonable, e.g., what generative systems are worthy of consideration, and what candidate solutions are realistic.

For the sake of comparison, we choose 10 hyperparameterizations of (6.1)-(6.2). **Importantly**, in what follows, we divide these 10 systems into 2 groups of 5. This is so because each group of 5 is identical to the other except for the Matern covariogram shape hyperparameter which, in the first set, group "A", is set to $\kappa = 0.5$, and in the second set, group "B", set to $\kappa = 2$. There is nothing particularly noteworthy about our choices; they are intended to be somewhat

generic. In effort to pique the imagination, though, we construct three of each of each five for the sake of miss-identification of the Kalman solution. Two have temporally colored noise, one has a (nearly) spacially constant mean function.

6.2.1 Simulated Systems

In each of the five, we have our spacial support be $14 \cdot 14 = 196$ evenly spaced gridded sites in the domain $\Omega = [0, 10] \times [0, 10] \subset \mathbb{R}^2$.

In each of the five we have $\mathbf{F} = \mathrm{F}\,\mathbf{I}$ and $\mathbf{Q} = \mathrm{Q}\,\mathbf{I}$, $\mathrm{F} = 0.999$, $\mathrm{Q} = 0.1$.

In each of the five we have $\mathbf{R} = \mathrm{R}\,C(\mathbf{D}) + \sigma_\varepsilon^2\,\mathbf{I}$, with $\mathrm{R} = 7$, $\sigma_\varepsilon^2 = 0.5$, except that in two, `sim-1-dt3` and `sim-1-dt6`, the measurement errors are temporally correlated, i.e., looking at 6.2, $\boldsymbol{\varepsilon}_i \not\perp \boldsymbol{\varepsilon}_k, \forall i, k$. To get a visual sense of this temporal coloration, see the right-most panel of Figures 6.4-6.5 for Group A, and Figures 6.10-6.11 for Group B.

In each case, $C(\cdot)$ is Matern, the (negative) rate parameter, d-alpha for each can be found in Table 6.1. For each of the five systems, we use a different shape parameter. Heat maps (196×196) of our spacial covariation, $\mathbf{R}$, can be found, for group A, in Figures 6.2-6.5, and 6.8-6.11 for group B.

In each excepting `sim-ss-1-MIDH` we have $\mathbf{H}$ be a 2-D cosine tensor expansion in Ω. For example, the ith row of $\mathbf{H}$ (i.e., the bases tranformation for site i) equals $\cos[x_i \cdot \mathbf{u}_x] \otimes \cos[y_i \cdot \mathbf{u}_y]$, where $\otimes$ signifies the tensor product, and we have set $\mathbf{u}_x = (0.005\pi,\, 0.15\pi)$, and $\mathbf{u}_y = (0.05\pi,\, 0.2\pi)$. So, $\mathbf{H}$ is 196×4.

A snapshot of this mean function under simulation at three time points can be seen in Figure 6.3.

For `sim-ss-1-MIDH`, we have $\mathbf{u}_x = (0.005\pi)$, and $\mathbf{u}_y = (0.05\pi)$ — that is, the mean surface, $\mathbf{H}\boldsymbol{\beta}$, would appear nearly constant in the x direction. In *all* the solutions we will explore for this generative system, $\mathbf{H}$ is set (erroneously) to be constant, i.e., an $n \times 1$ matrix filled with 1's.

For each we create 30 examples (replications) of length $\tau = 300$. We set our training range to $t \in 50:249$, and test over $t \in 250:297$.

Simulations were accomplished using the `R` package **SSsimple** [Zes12a].

6.2.2 Simulating Temporal Correlation of Observation Noise

We employ Gibbs sampler to generate temporally correlated observation noise [RDS96, Hoc09].

$$\mathbf{D}^{(0;1)} = \sqrt{\begin{pmatrix} \mathbf{D}^{\Omega} & \mathbf{D}^{\Omega} \\ \mathbf{D}^{\Omega} & \mathbf{D}^{\Omega} \end{pmatrix} + \begin{pmatrix} \mathbf{D}^{(0)} & \mathbf{D}^{(1)} \\ \mathbf{D}^{(1)} & \mathbf{D}^{(0)} \end{pmatrix}} \tag{6.13}$$

$$\mathbf{C} = C\left(\mathbf{D}^{(0;1)}; \alpha, \kappa\right) \tag{6.14}$$

The objects from (6.13) are identical, or are direct analogues of those presented in (4.7) — as such, recall, $\mathbf{D}^{(0)}$ is the empty matrix. $C(\cdot)$ is the Matern covariogram. Have $\mathbf{C}_O$ and $\mathbf{C}_D$ respectively be an off-diagonal and on-diagonal quadrant of $\mathbf{C}$. Set $\mathbf{C}_R = \mathbf{C}_D - \mathbf{C}_O\mathbf{C}_D^{-1}\mathbf{C}_O$. Now,

$$\boldsymbol{\mu}_{\varepsilon\, t} = \mathbf{C}_O\mathbf{C}_D^{-1}\left(\boldsymbol{\varepsilon}_{t-1} - \boldsymbol{\mu}_{\varepsilon\, t-1}\right) \tag{6.15}$$

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}[\boldsymbol{\mu}_{\varepsilon\, t}, \mathbf{C}_R] \tag{6.16}$$

with $\boldsymbol{\mu}_{\varepsilon\, 0} = \boldsymbol{\varepsilon}_0 = \mathbf{0}$. So that for any positive, finite value in the constant temporal distance matrix $\mathbf{D}^{(1)}$, $\boldsymbol{\varepsilon}_{t-1} \not\perp \boldsymbol{\varepsilon}_t$. The error draws for (6.16) were acomplished with the `R` package **mvtnorm**.

6.2.3 System Identification & Solutions

Our objective here is to get some flavor of how, or to what degree, using the (much less computationally intensive) WIDALS deteriorates a prediction quality against something close to an optimal solution. This is not intended as a foray into the

expansive, and potentially contentious, arena of *System Identification* (SID) *per se*. For this reason, our SID implementations as applied to our simulated models are either trivial or wrong, without pretense.

Recall that we start with 10 generative models — the five shown in Table 6.1 for each of two values of the Matern shape parameter, $\kappa \in \{0.5, 2\}$, used to construct the system covariance matrix, $\mathbf{R}$.

For each, we entertain two predictive scenarios: one-step-ahead forecasting, and interpolation. In forecasting, we use lag-1, or, lag-1 and lag-2 information to forecast to time t. That is, we use $\mathbf{z}_{t-1}$ or $(\mathbf{z}_{t-2}, \mathbf{z}_{t-1})$ to predict $\boldsymbol{z}_t$. In interpolation, we use 14-fold sitewise cross-validation. That is, we cleave $\mathbf{Z}(\boldsymbol{\omega})$ by a fixed set of sites, $\boldsymbol{\omega}_1$, to create $\mathbf{Z}(\boldsymbol{\omega}_1)$ and $\mathbf{Z}(\boldsymbol{\omega}_2)$, where $\boldsymbol{\omega}_1 \cup \boldsymbol{\omega}_2 = \boldsymbol{\omega}$ and $\boldsymbol{\omega}_1 \cap \boldsymbol{\omega}_2 = \emptyset$. So that we use $\mathbf{z}_t(\boldsymbol{\omega}_2)$, or, $\big(\mathbf{z}_{t-1}(\boldsymbol{\omega}_2), \mathbf{z}_t(\boldsymbol{\omega}_2)\big)$ to predict (interpolate) $\mathbf{z}_t(\boldsymbol{\omega}_1)$.

This means that we have 20 general prediction scenarios. Boxplots of solution RMSEs (30 replications) for all 20, for a select handful of specific solutions, is provided in Figures 6.6-6.13.

Solution labelled "SS QRF" utilizes Stochastic-Search to locate six scalar hyperparameters, $\{\mathrm{Q}, \mathrm{R}, \alpha, \kappa, \sigma_\varepsilon, \mathrm{F}\}$, running over the Kalman Solution (6.3)-(6.7). The spacial mean function, $\mathbf{H}$, was set to truth.

Solution labelled "SS RLL" utilizes the base `R` function `optim` running over the log-likelihood function for $\mathbf{R} \,|\, \mathbf{H}, \mathbf{F}, \mathbf{Q}$ to locate $\{\mathrm{R}, \alpha\}$. The resulting best hyperparameter set was passed to the Kalman Solution to obtain the solution RMSE.

Solution labelled "R true" passes the true hyperparameter set to the Kalman Solution to obtain the solution RMSE.

Solution labelled "WIDALS 1 L1" uses $\mathbf{z}_{t-1}$ to forecast $\boldsymbol{z}_t$, or $\mathbf{z}_t(\boldsymbol{\omega}_2)$ to interpolate $\mathbf{z}_t(\boldsymbol{\omega}_1)$. The five non-negative scalar hyperparameters, $\Pi^{\mathrm{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\}$, from §4.0.3 are located by stochastic-search.

Solution labelled "WIDALS 1 L2" uses $(\mathbf{z}_{t-2}, \mathbf{z}_{t-1})$ to forecast $\boldsymbol{z}_t$, or $\big(\mathbf{z}_{t-1}(\boldsymbol{\omega}_2), \mathbf{z}_t(\boldsymbol{\omega}_2)\big)$ to interpolate $\mathbf{z}_t(\boldsymbol{\omega}_1)$. We use stochastic-search to locate Π^{WID}.

Solution labelled "GLM 1 L1" is analogous to "WIDALS 1 L1", except that the GLM solution for the residual smoothing replaces that used in WIDALS. That is, we replace $\mathbf{W}$ from (4.9) with $C(\mathbf{D}^{\diamond\times\star})\, C(\mathbf{D}^{\diamond\times\diamond})^{-1}$ (same as (4.4)). This solution requires 6 hyperparemeters, ρ and λ for the ALS part, and γ the temporal distance constant for the distance matrices, and α, κ, and σ for the Matern covariogram, $C(\cdot)$.

Solution labelled "GLM 1 L2" is likewise analogous to "WIDALS 1 L2".

Solution labelled "FRKALS" uses $1+4+16 = 21$ resolution points, $196/4 = 49$, covariance smoothing points, and the radial bases transform to construct $\mathbf{S}$ all as described by Cressie and Johannesson in [CJ08].

Early in our analyses we also employed the sub-optimal sub-space method to locate the full complement of hyperparameters, $\mathbf{H}, \mathbf{F}, \mathbf{R}, \mathbf{Q}$ as detailed in [Lju98] and implemented in the package **SSsimple**. These hyperparameters, when used in the Kalman solution occasionally produced decidedly inferior RMSEs, so these results were omitted.

6.2.3.1 Computational Expense

The Kalman solution requires τ inversions of an $n \times n$ matrix (6.5).

The GLM solution requires a single $kn \times kn$ matrix inversion, where k is the number of lags used.

6.2.4 Results Discussion

Figure 6.6 shows the distributions of 1-step-ahead forecast solution RMSEs when the system spacial covariance is the Matern covariogram acting on euclidean dis-

tance with shape parameter, $\kappa = 0.5$ (same as exponential covariogram) (this is Group A). As in all the comparisons offered in this Chapter, the RMSEs are pairwise associated, i.e., a system realization is simulated, then solved by the various methods. This means that we must be careful in our interpretation of the resulting distributions. For example, in the leftmost plot, using System 1 (see Table 6.1), labelled `sim-ss-1`, we cannot say that the WIDALS lag-2 solution (5th boxplot distribution from left) beat the Kalman solution fitted with truth ("R-true") about 40 percent of the time. In fact, R-true beat WIDALS 1 L2 in 26 of the 30 simulations. In these boxplot representations, we can get a sense of how the RMSEs of the various solutions compare against the total variation of the RMSE across random realizations from the same generative system.

Keep in mind that in forecasting, the Kalman solution employs the *a priori* states (aka, partial slopes in a regression context) to predict the ensuing observations. The observation covariance structure, $\mathbf{R}$, while used to predict these states, (6.5), it does not play into the prediction of the ensuing observations since the Kalman solution, as given, assumes observations are uncorrelated across time.

The context of the boxplots as presented invokes a scenario in which two researches each have a different draw from the same system, one uses one solution, the other, another solution. In this way, the boxplot distributions can be regarded as independent.

In System 3, labelled `sim-1-dt6`, the observation shocks are slightly colored across time. The degree of this coloration can be seen visually in Figure 6.5 rightmost panel. Notice the faint patterning of lightness (compare with Figure 6.4, where the observation shocks are temporally uncorrelated). While the Kalman solutions (SS QRF, SS RLL, R true) are correctly identified under the constraint of the system (6.1)-(6.2) under the usual assumption of uncorrelated measurement shocks (so in a sense, the label R true is not a misnomer), the system, in the functional sense, has been miss-IDed. Here, the WIDALS solutions both out

perform the erroneous Kalman solution. This disparity in performance is greatly exaggerated when the strength of the temporal coloration increases, as shown in `sim-1-dt3` (see Figure 6.5, rightmost panel, for a visual sense of the strength of this coloration).

Perhaps most curious of all our comparisons occurs when the smooth spacial structure (the surface *mean function*, **H**, in a regression context) has been very slightly miss-IDed, e.g., the rightmost panel in Figure 6.6. Recall that in truth, **H** was set to be a cosine expansion with a very long spacial periodicity. While, as we have stressed, this submission does not attempt to engage the voluminous details of system identification, it is so that such subtle deflections in **H** can offer challenges in detection. As mentioned above, all eight solutions (incorrectly) have **H** be constant. Said in short, WIDALS survived this miss-ID far better than did the Kalman solution.

Figure 6.7 shows our eight solutions working on Group A using interpolation. Recall that, as mentioned, *honest* performance assessment requires emulating the absence of sites. As described earlier, we use 14-fold cross-validation over space to create these RMSEs. Here, there are three major takeaways. First, FRKALS suffers. Estimating the covariance structure using the computationally facile matrix product $\mathbf{SKS}^T$ (see [CJ08], section 2.2) greatly penalizes the resulting interpolated prediction compared with the other solutions. Second, the WIDALS solutions compete closely with the Kalman solutions. Third, the Kalman solutions here are much more resilient to miss-IDing than when forecasting.

Figure 6.12 is analogous to Figure 6.6, the difference is that the Matern shape parameter has been set to $\kappa = 2$. When $\kappa > 1/2$, the covariogram gains an inflection point — covariance declines more slowly as distance increases away from zero. The picture here is very similar to that of Figure 6.12, except that WIDALS's performance is slightly degraded. This property, by the way, has come to be well known by us in researching WIDALS over numerous datasets: when

the empiric covariance is well fit by a covariogram with a second derivative that increases for some distance at and beyond the origin, the WIDALS prediction quality degrades slightly.

Figure 6.13 is similarly analogous to Figure 6.7. And, as in the previous forecasting case, the inflected covariogram slightly degrades the WIDALS prediction against the Kalman solutions. Nonetheless, if we view these RMSEs as independent across solutions (*a la* the scenario described above with two researchers each in possession of different realizations from the same system), in *every case* presented, the Kalman solution will best a WIDALS solution only about, say, 20 percent of the time — the 3rd quartile of the WIDALS RMSE about abuts with the median RMSE of the Kalman solution.

6.3 Systems

Sys #	Sys Name	d-alpha	**R** scale	Var **Z**	Var **Y**	Miss IDed?
1	sim-ss-1	0.40	7.00	31.00	24.00	—
2	sim-ss-2	1.00	7.00	32.00	24.00	—
3	sim-1-dt6	0.40	7.00	34.00	24.00	slight
4	sim-1-dt3	0.40	7.00	34.00	24.00	severe
5	sim-ss-1-MIDH	0.40	7.00	8.00	0.00	slight

Table 6.1: Our 5 simulated systems.

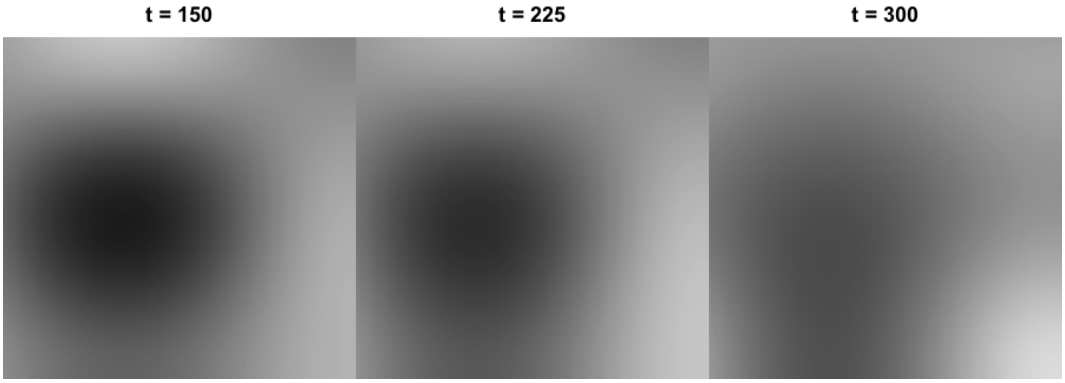

Figure 6.1: Three snapshots of the mean-function surface use in Systems 1-4 taken form one of the simulations.

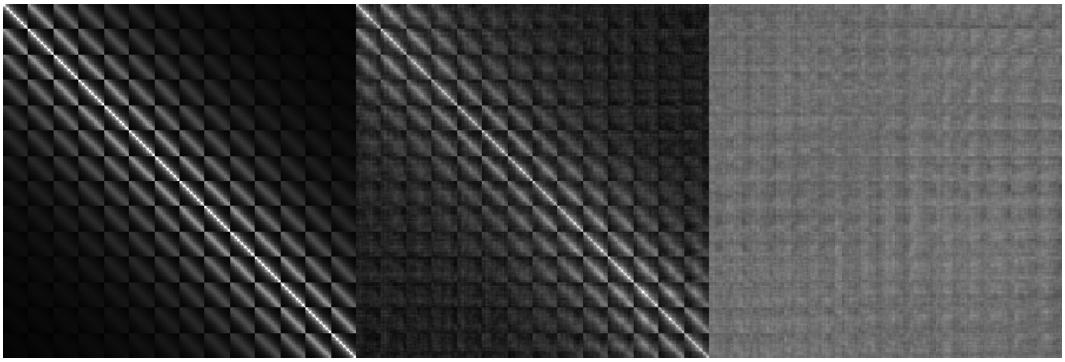

Figure 6.2: True cov, empiric cov, lag 1 cov of System 1, group A ($\kappa = 0.5$).

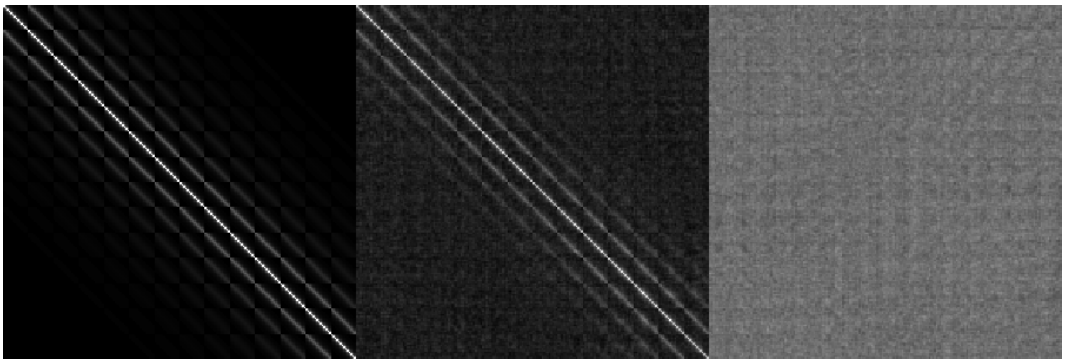

Figure 6.3: True cov, empiric cov, lag 1 cov of System 2, group A ($\kappa = 0.5$).

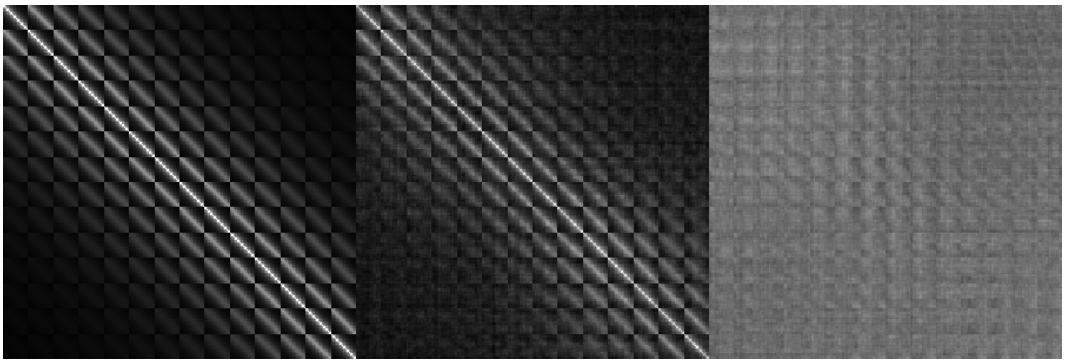

Figure 6.4: True cov, empiric cov, lag 1 cov of System 3, group A ($\kappa = 0.5$).

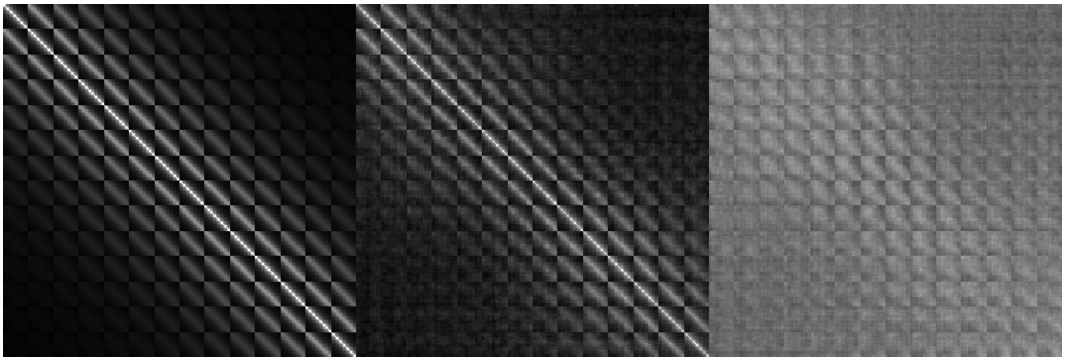

Figure 6.5: True cov, empiric cov, lag 1 cov of System 4, group A ($\kappa = 0.5$).

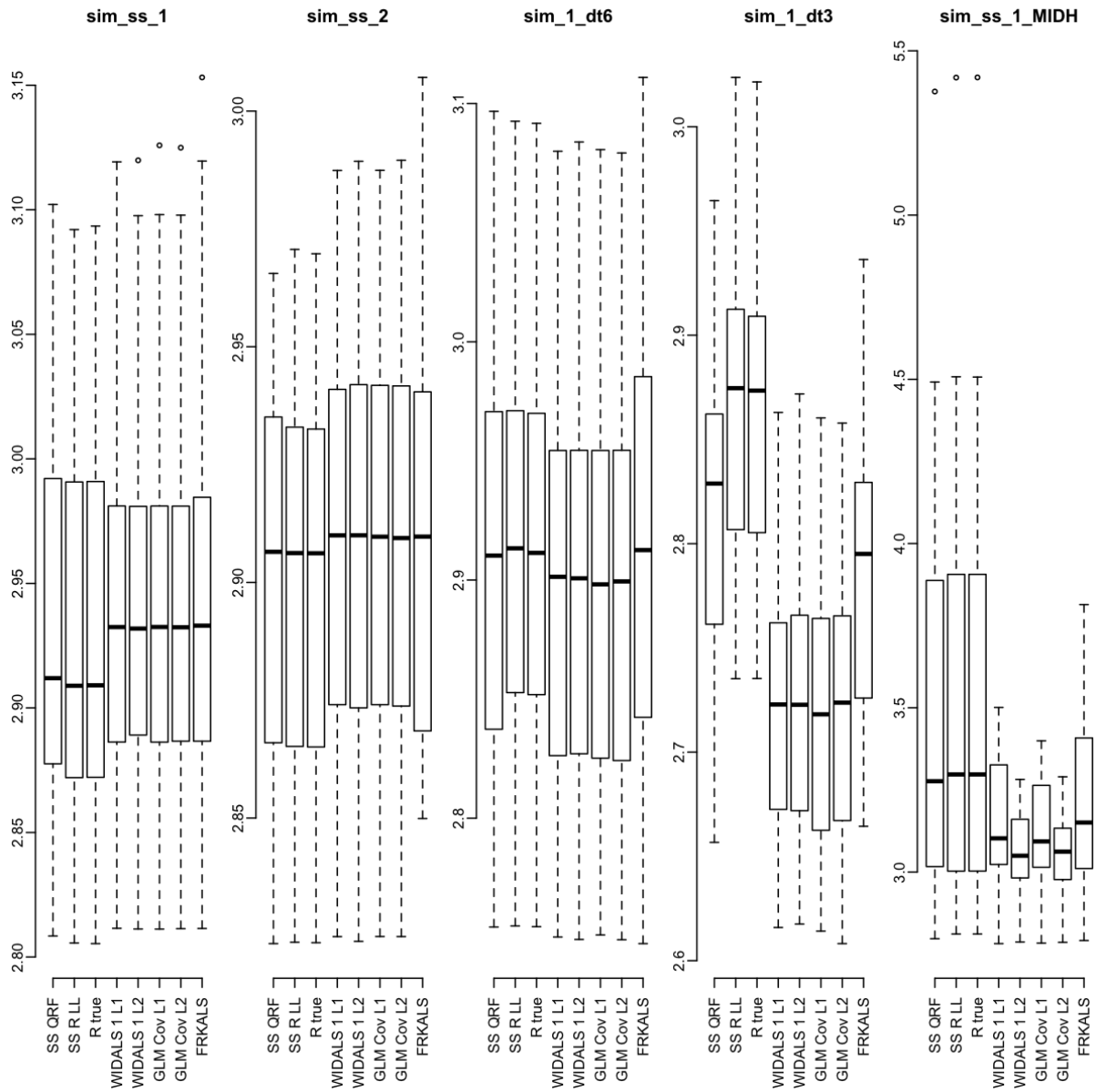

Figure 6.6: RMSE forecast performance. kappa=0.5

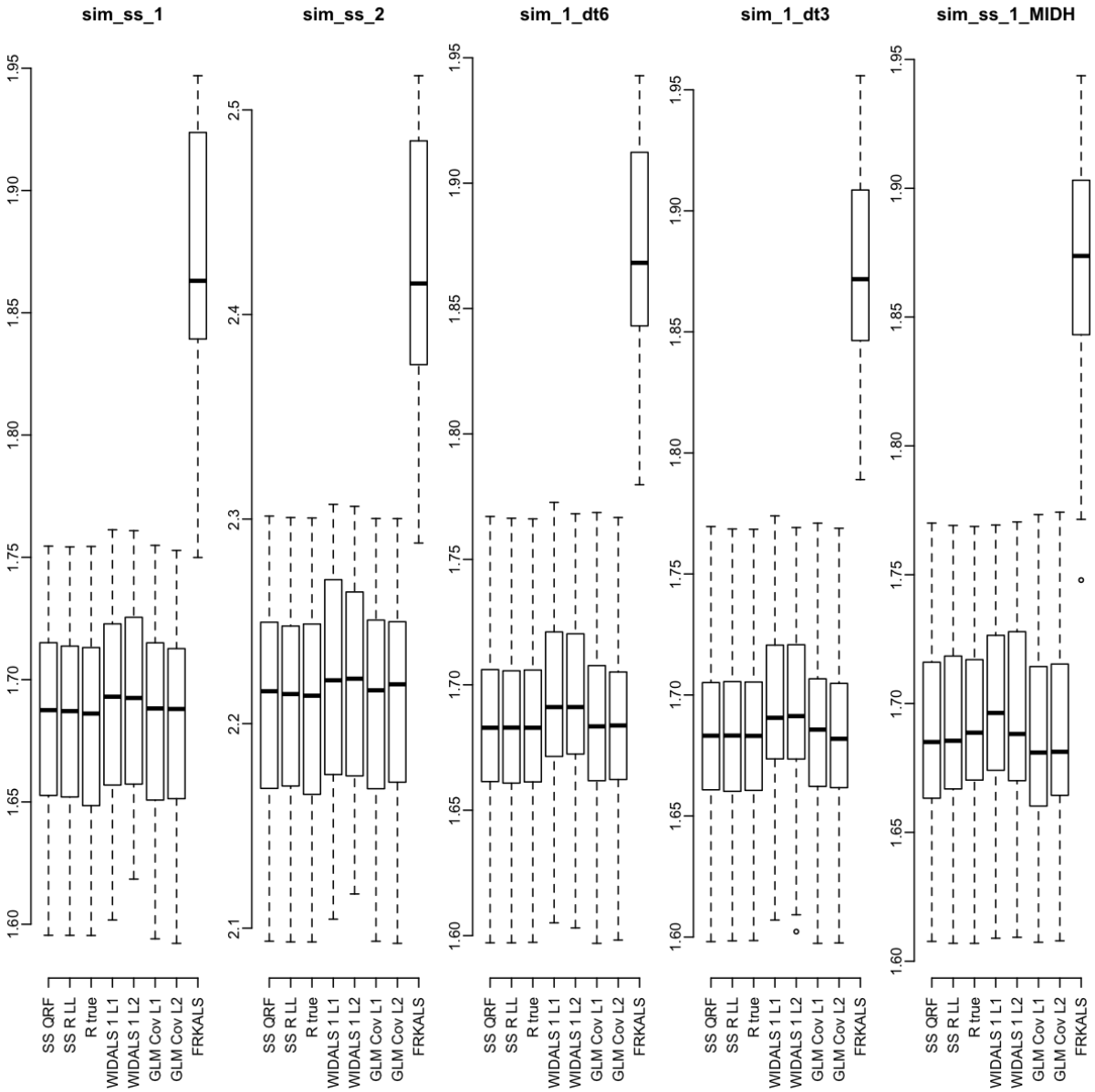

Figure 6.7: RMSE interpolation performance. kappa=0.5

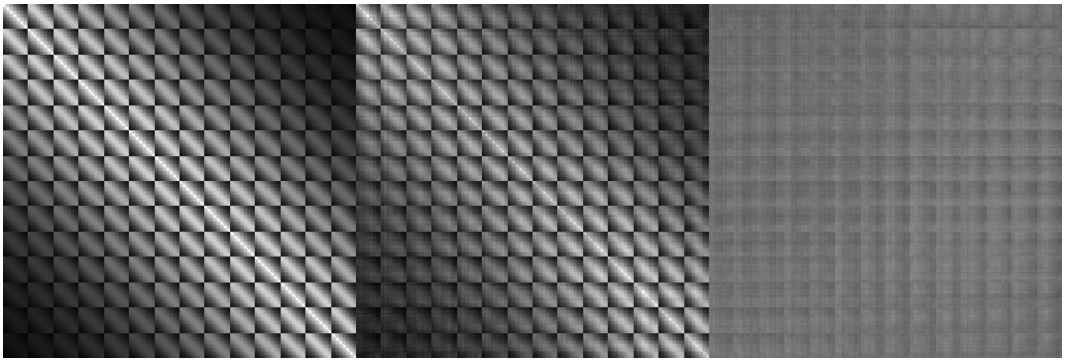

Figure 6.8: True cov, empiric cov, lag 1 cov of System 1, group B ($\kappa = 2.0$).

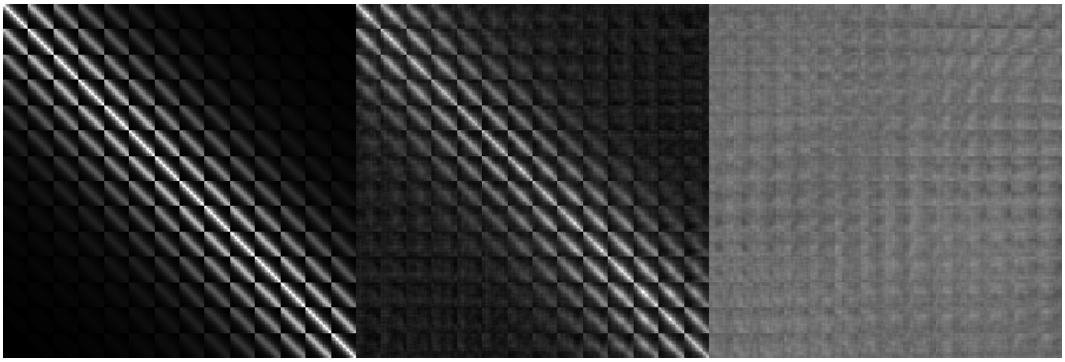

Figure 6.9: True cov, empiric cov, lag 1 cov of System 2, group B ($\kappa = 2.0$).

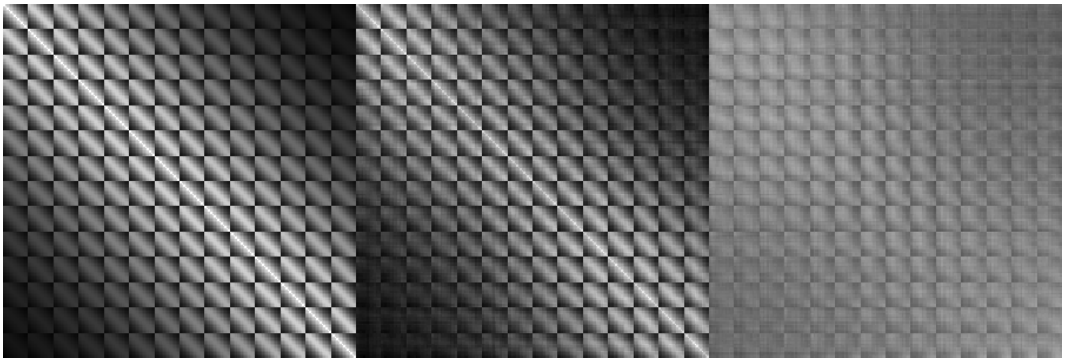

Figure 6.10: True cov, empiric cov, lag 1 cov of System 3, group B ($\kappa = 2.0$).

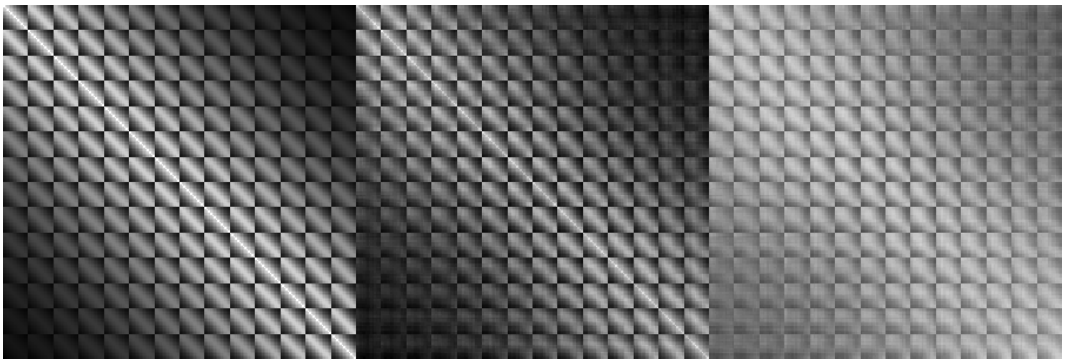

Figure 6.11: True cov, empiric cov, lag 1 cov of System 4, group B ($\kappa = 2.0$).

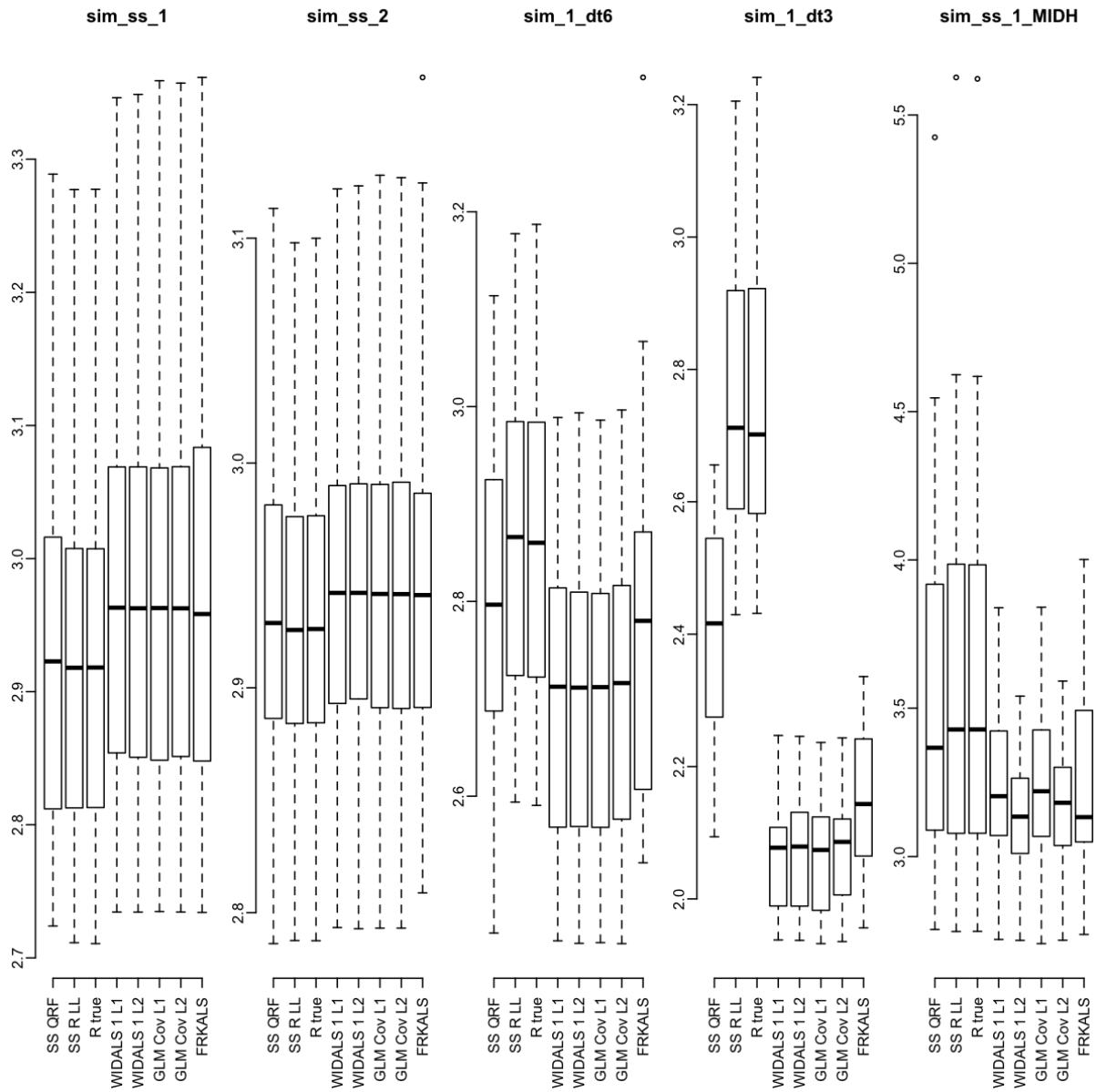

Figure 6.12: RMSE forecast performance. kappa=2

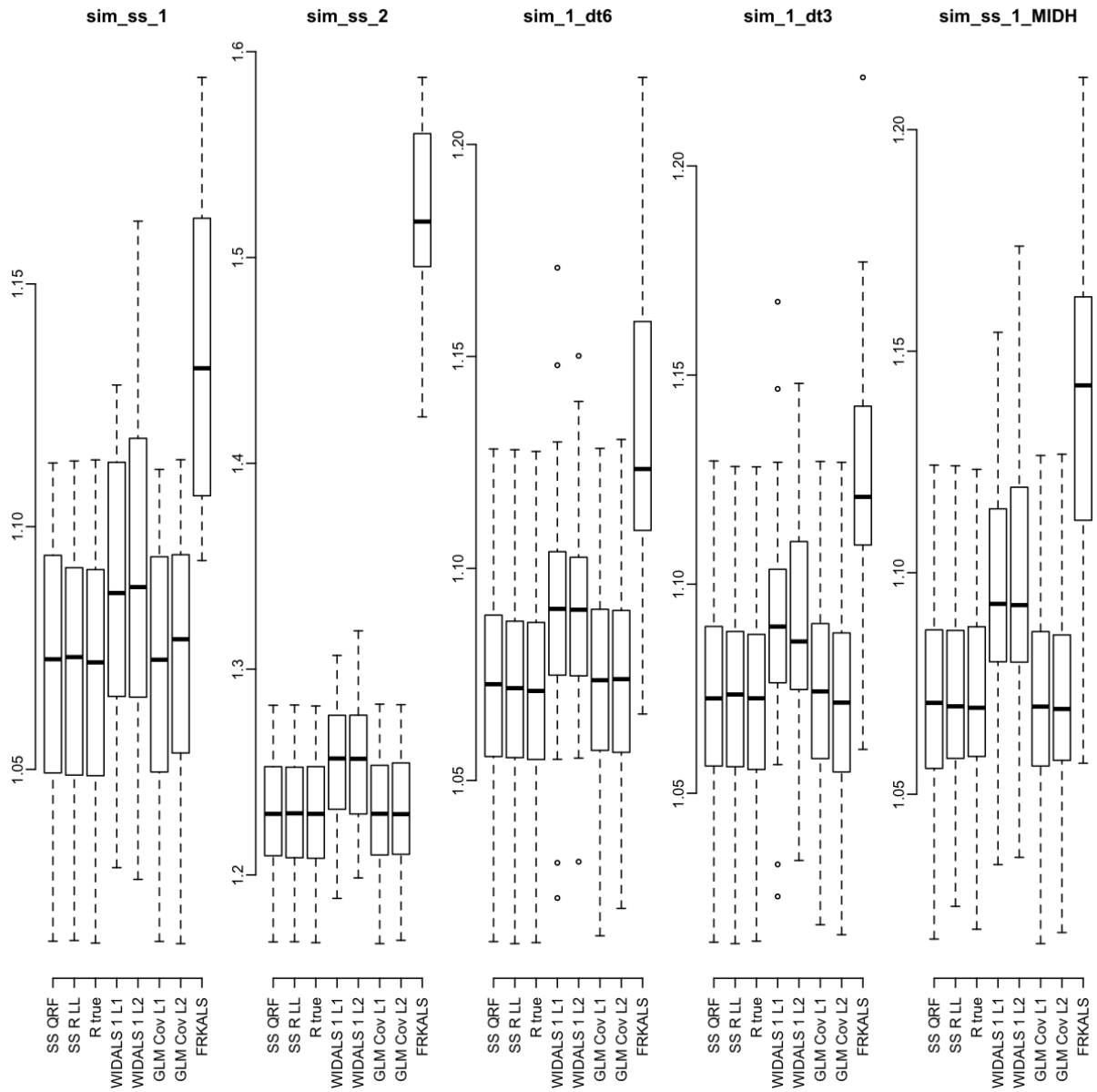

Figure 6.13: RMSE interpolation performance. kappa=2

6.4 Full System Infilling

In §4.1.1 we examined the cost of using the WIDALS $\mathbf{W}$ against an optimal solution, given the true system covariogram, in a static setting under some infilling scenarios. In the immediately preceding Section we compared the cost of WIDALS solutions against the Kalman Filter, though for relatively small data simulations. Here, we shall conclude our investigation into comparisons under spacial infilling.

As done previously, our spacial domain is $\mathbf{\Omega} = [-1, 1] \times [-1, 1]$.

We make use of three possible parameterizations of our simulated systems, and three infilling scenarios. We employ two different mean functions, a constant, (6.17), and an order-7, somewhat complex mean function, (6.18). In the latter case, as well as using a constant diagonal state variance matrix, $\mathbf{Q} = 0.1\,\mathbf{I}$, we offer a version that possesses rather strong mutual covariation between the 7 state variables, as shown in Figure 6.15. Note that, e.g., from (6.18), $\sin[14\mathbf{x}]$ will make about four-and-a-half cycles over $[-1, 1]$.

$$\mathbf{H} = \mathbf{1} \tag{6.17}$$

$$\mathbf{H} = \left[\mathbf{1}, \mathbf{x}^T, \mathbf{y}^T, \sin[14\mathbf{x}]^T, \sin[14\mathbf{y}]^T, \cos[14\mathbf{x}]^T, \cos[14\mathbf{y}]^T \right] \tag{6.18}$$

As with our spacial region, our system covariogram, should be familiar to the reader, $C(\mathbf{D}) = \exp[-2 \cdot \mathbf{D}] + 0.01\,\mathbf{I}$, as shown in Figure 6.14.

In every case, our temporal domain is $t \in 1 : 140$, and our training-testing range is $71 : 140$.

Cross Infilling: We place a single site at the origin to which we will interpolate. We then infill in a cross pattern (negative integer powers of 1.1 along each axis). Under the given system parameters, at certain intervals during addition of supporting sites we calculate the theoretical Kalman MSE. For each we fit WIDALS over the support, interpolate to the single fixed site, record this empiric MSE. This

is repeated through 100 data replications. Infilling results for all three proffered parameterizations is given in Figure 6.16. Two prominent features jump out at a short glance. First, the MSE ratio is generally greater when $\mathbf{H}$ is complex, and second, most importantly, in each case there appears a general tendency of the MSE to unity as the spacial domain becomes more populated with supporting sites.

Spiral Infilling: We place a single site at the origin to which we will interpolate. We then infill as we did in the cross pattern (negative integer powers of 1.2 along each axis, this time). However, the angle w.r.t. the origin slowly increases as sites are added. Under the given system parameters, for each addition we calculate the theoretical Kalman MSE. For each we fit WIDALS over the support, interpolate to the single fixed site, record this empiric MSE. This is repeated through 100 data replications. Infilling results for constant $\mathbf{H}$ and complex $\mathbf{H}$ with state cross-correlation are given in Figure 6.17. Here, the MSE ratio appears indifferent to the choice between our two system measurement functions. As with cross infilling above, the MSE ratio tends to unity, albeit a little more slowly.

Random Infilling: We randomly (uniform) select 100 sites to which we interpolate. We then randomly (uniform) add 1000 support sites, in turn, up to a total of 12000 supporting sites. Under the given system parameters, for each addition we calculate the theoretical Kalman MSE. For each we fit WIDALS over the support, interpolate to the 100 fixed sites, record this empiric MSE. This is repeated through 20 unique distribution of locations and a single data replication for each. Infilling results for constant $\mathbf{H}$ and complex $\mathbf{H}$ with state cross-correlation are given in Figure 6.18. 70% confidence bands have been added to the plots. In this scenario, the result is far less convincing, although one does get a sense that the MSE ratio is progressively decreasing.

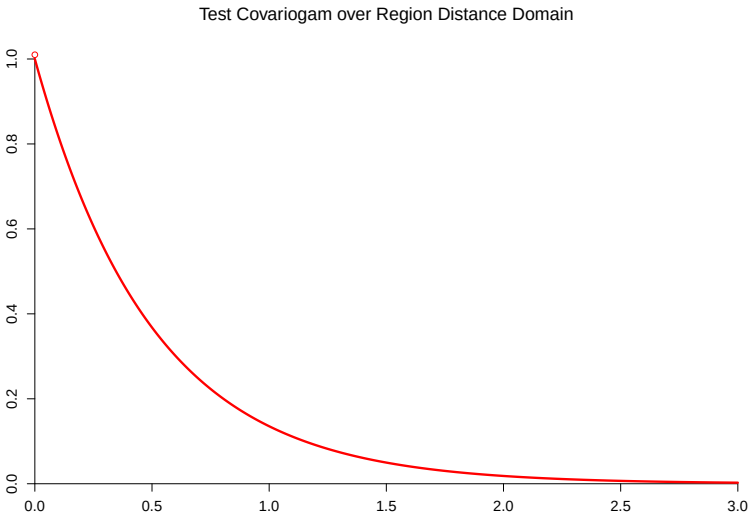

Figure 6.14: System Covariogram.

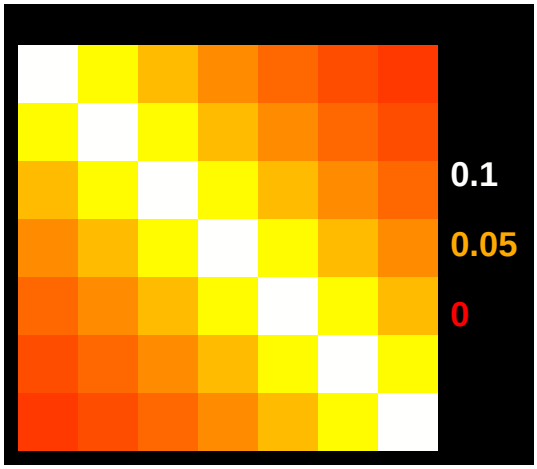

Figure 6.15: System Q.

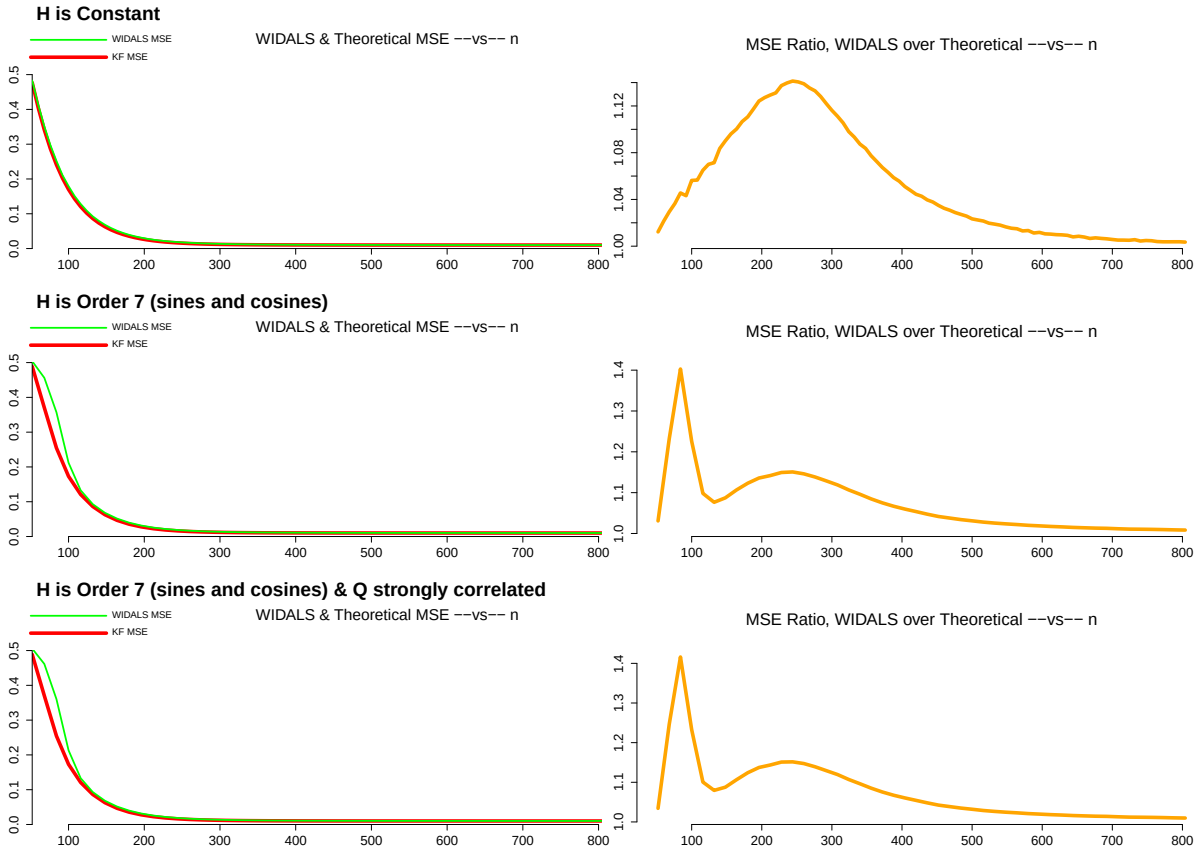

Figure 6.16: WIDALS vs KF, Cross infilling.

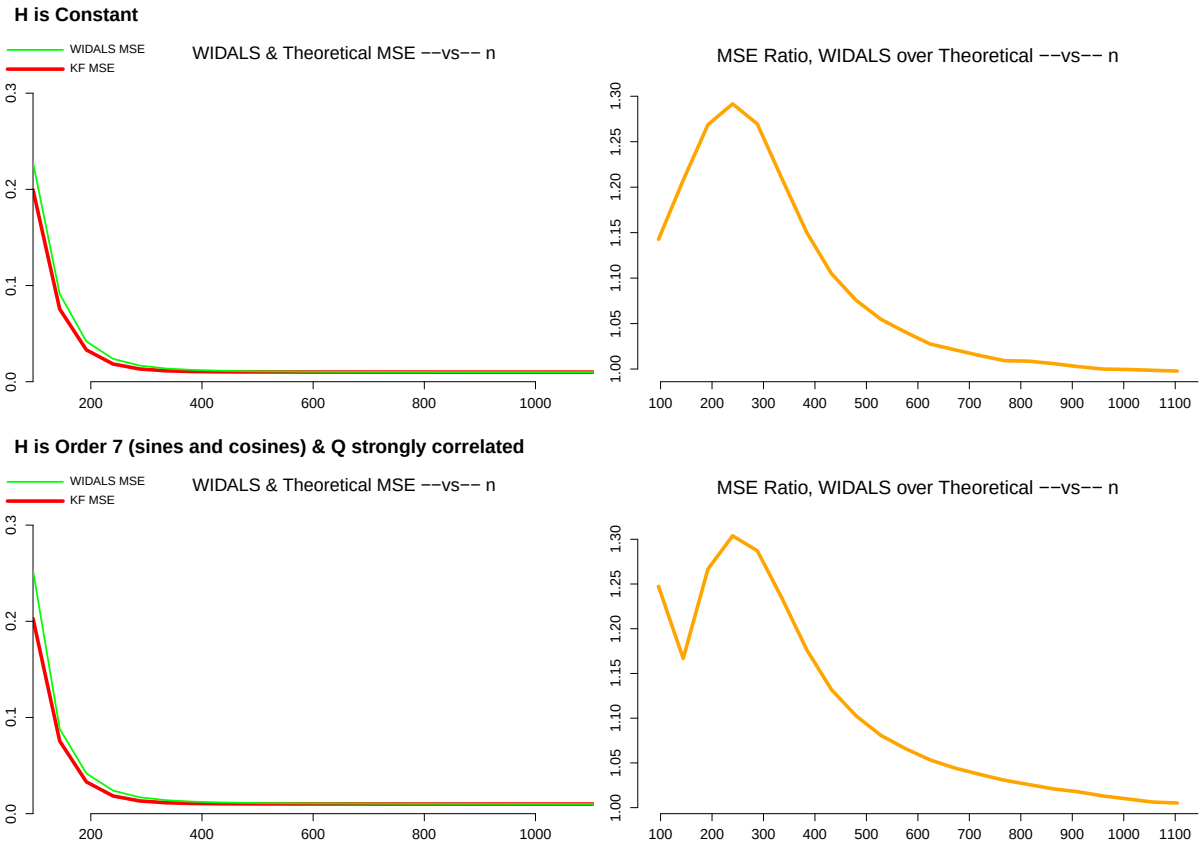

Figure 6.17: WIDALS vs KF, Spiral infilling.

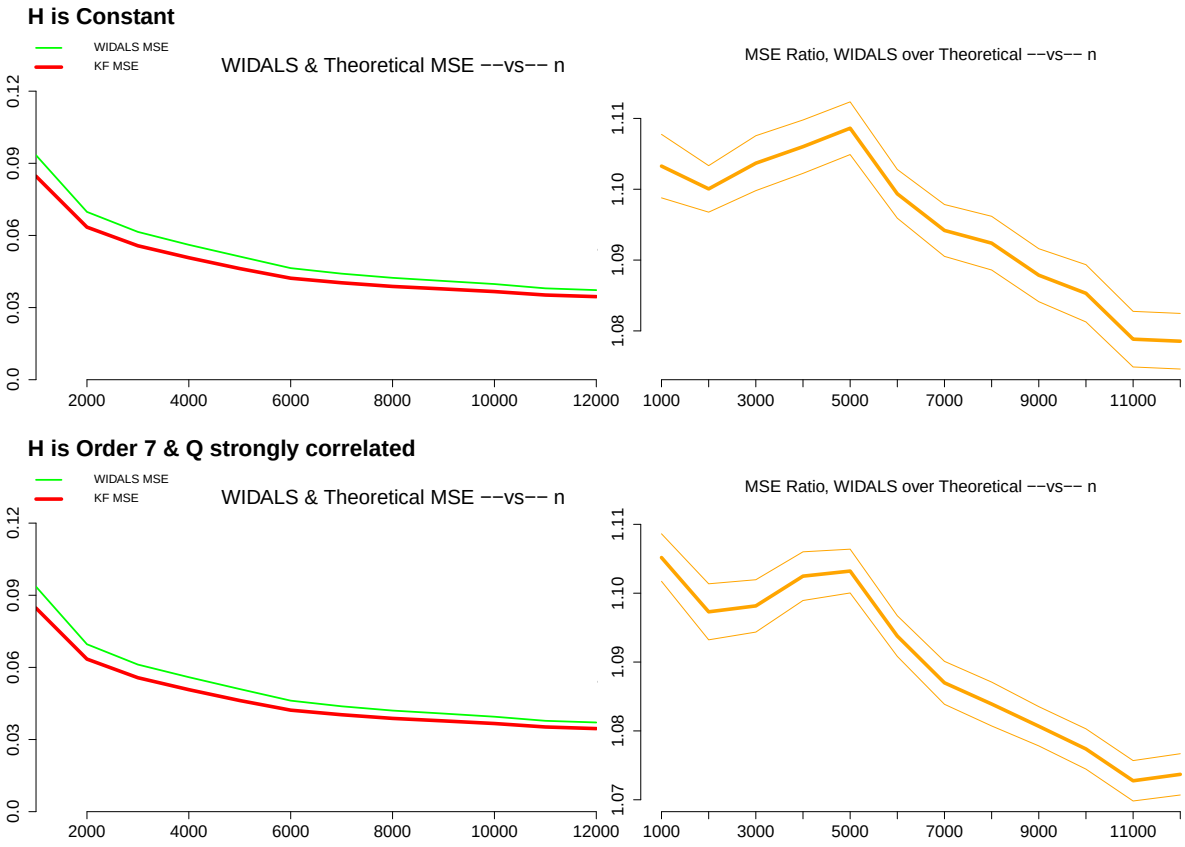

Figure 6.18: WIDALS vs KF, Uniform Random infilling.

6.4.1 Comparison Concluding Comments

The presented results are a relatively small snapshot from a much larger body of work. We will conclude this Section with some general observations. When the interpolation point(s) has *borrowing power* from the supporting sites, i.e., the distribution of distances between $\mathbf{s}^\star$ and $\mathbf{s}^\diamond$ is similar to that of the mutual distances in $\mathbf{s}^\diamond$, AND, when even a tiny nugget effect is present, it appears that the WIDALS over Optimal MSE will tend towards unity — although, in some cases, much more slowly than in others.

When the spacial support is very large, say 10,000 or more, solution tractability requires computational modifications to the canonical Kalman Filter. In our work, we made use of two adaptations, both available in a newer release of **SSsimple** (v0.6.0). One involves using the Woodbury identity for the inversion of $\mathbf{HPH}^T+\mathbf{R}$ required for the Kalman solution, the other makes use of finding the system steady-state so to obviate the inversion and associated matrix products as soon as possible over the time domain. Notice the small "hiccup" at $n = 12000$ in the lower MSE plot from Fig. 6.18. As a point of disclosure, other random uniform infill simulations suggested a departure from the downward trend of the MSE ratio for very large n. There is some suggestion that this departure may result, perhaps in part, from the presently described adaptations to the Kalman Filter. Future investigations into WIDALS's performance should explore this possibility.

CHAPTER 7

Applications & Examples

7.1 Introducing: "Phyning"

We use the term *Phyning* in the present work to refer to a process of *recursive fitting-and-imputation over multiple contemporaneous data sets where the responses may be correlated across data sets.*

Supose we have two data sets, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ that are row-wise associated, i.e., responses are collected (or can be viewed as having been collected) within the same time intervals, e.g., daily.

The respective sites, $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ may be such that $\boldsymbol{\omega}_1 \cap \boldsymbol{\omega}_2 = \emptyset$, or $\boldsymbol{\omega}_1 \cup \boldsymbol{\omega}_2 = \boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$, or anything in between.

For the sake of the following description, assume the sites are the same, $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$. This is done only to ease notation.

Have $\widetilde{\mathbf{Z}}_1$ be a prediction of $\mathbf{Z}_1$.

Have $\overset{\circ}{\mathbf{Z}}_1$ be the result of filling missing values in $\mathbf{Z}_1$ with $\widetilde{\mathbf{Z}}_1$.

Have $\mathbf{H}_t^{\overset{\circ}{\mathbf{Z}}_1} = \overset{\circ}{\mathbf{z}}{}_{1\,t}^{T}$ be the $n \times 1$ matrix filled with $\overset{\circ}{\mathbf{z}}_{1\,t}$.

Have $\mathbf{H}_t^{\boldsymbol{\Omega}}$ be system covariates.

Now, start by having $\widetilde{\mathbf{Z}}_1$ be the global mean of $\mathbf{Z}_1$. Likewise for $\widetilde{\mathbf{Z}}_2$.

Create $\overset{\circ}{\mathbf{Z}}_1$ and $\overset{\circ}{\mathbf{Z}}_2$.

Create $\mathbf{H}_t = \left[\mathbf{H}_t^{\boldsymbol{\Omega}}, \mathbf{H}_t^{\overset{\circ}{\mathbf{Z}}_2}\right]$

Loop the following:

Use MH over WIDALS to search out Π_1^{WID} using $\mathbf{H}_t$ and $\overset{\circ}{\mathbf{Z}}_1$.

If a local search improves the RMSE, recreate $\overset{\circ}{\mathbf{Z}}_1$.

Create $\mathbf{H}_t = \left[\mathbf{H}_t^{\mathbf{\Omega}}, \mathbf{H}_t^{\overset{\circ}{\mathbf{Z}}_1}\right]$.

(The following three steps are the same as the previous three, but for the sake of $\mathbf{Z}_2$)

Use MH over WIDALS to search out Π_2^{WID} using $\mathbf{H}_t$ and $\overset{\circ}{\mathbf{Z}}_2$.

If a local search improves the RMSE, recreate $\overset{\circ}{\mathbf{Z}}_2$.

Create $\mathbf{H}_t = \left[\mathbf{H}_t^{\mathbf{\Omega}}, \mathbf{H}_t^{\overset{\circ}{\mathbf{Z}}_2}\right]$.

End loop.

The loop is run until the RMSEs converge. Of course, one can deliberately remove data from the data sets to get a more genuine sense of the quality of the imputations at large.

As is obvious, Phyning appears a monumental computational campaign. With WIDALS, however, phyning over numerous data sets containing millions of datum is easily in reach.

7.2 Mini Comparison Using Daily Climate Data

We apply Phyning to 6 contemporaneous data sets supplied by the National Climatologic Data Center (NCDC). We perform this both using WIDALS and also a simple Kalman solution. Each data set is $\tau \times n = 1016 \times 74$ (all sharing the same 74 sites). Data are recorded daily, running from 2009-01-01 to 2011-10-13. Some detail is offered in Table 7.1. The column "n CV NA" gives the number of observations that had been present in the original data but were deliberately removed for the sake of measuring performance. The column "n Total NA" gives the total number of missing observations (including the ones deliberately removed).

We construct our mean function covariates, $\mathbf{H}$, in the following way. For each time t, $\mathbf{H}_t$ is 74×6. The first column is constant. The second column contains elevation of sensor. The third and fourth columns contain the sine-cosine annual seasonal transformation. The fifth column contains Earth surface area incidental to the sun (the "incident surface area", ISA, commonly used to calculate incident solar energy),

$$\psi = 23.5 \sin[\, \delta(t)\, 360/365.25\,] \tag{7.1}$$

$$\eta = 90 - \arccos\big[\, \cos[y] \cos[\psi] + \sin[\psi] \sin[y]\,\big] \tag{7.2}$$

$$A\,(\delta(t), y) = \sin[\,\eta\,] \tag{7.3}$$

where δ is the number of days past the Spring equinox (March 20th), and where y is latitude in degrees. The sixth column contains the interaction (the mathematical product) of elevation and the ISA.

Notice, by the way, that the second column is only a function of space (elevation is spacially *separable* from time). The third and fourth columns are exclusively functions of time. The fourth and fifth columns are *non-separable* with respect to space-time. When dealing with large data sets, the computational memory burden can be greatly mitigated by keeping these separate (this is the approach used in **widals**).

The WIDALS solution used lag-0 information. Fitting the WIDALS solution, then, required locating four scalar, non-negative hyperparameters (since we use only lag-0 information, γ, our measure of temporal distance, is not required).

The Kalman solution used the *a posteriori* state estimate. For simplicity in IDing, $\mathbf{F}$ and $\mathbf{Q}$ were set to be diagonal constant. $\mathbf{R}$ was taken to be exponential over distances with a nugget effect. Fitting the Kalman solution, then, required locating five scalar, non-negative hyperparameters.

To ease visual comparison of the resulting RMSEs, all our observations were globally standardized. Precipitation and Visibility were preprocessed by applying

a square root transformation, then standardizing.

The result of phyning (9 passes) is provided in Figure 7.1. Two things impress us immediately. For one, neither WIDALS nor KF performed well predicting Precipitation and Visibility. Both these variables tend to act locally, and neither are obviously correlated with our chosen exogenous regressors, $\mathbf{H}_t$. Secondly, WIDALS and KF had virtually identical performance predicting temperature, both daily minimum and daily maximum. In both cases, the RMSEs are about 20% of the global standard deviation. Overall, this matchup is a virtual tie, with WIDALS besting KF predicting Standard Temperature Pressure, while KF performed better predicting Sea Level Pressure.

visib is maximum daliy prcp is total daily

NCDC Name	Long Name	Units	n Total NA	n CV NA
PRCP	Precipitation	inches	51485	110
MAX	Max Temperature	deg F	451	337
STP	Stnd Temperature Pressure	millibars	495	352
SLP	Sea Level Pressure	millibars	484	362
MIN	Min Temperature	deg F	496	370
VISIB	Visibility	miles	723	358

Table 7.1: NCDC Data info with missing entry counts for mini data sets (74 sites).

7.3 A More Adventurous Application of WIDALS

7.3.1 Preliminaries

This Section is the *magnum opus* of this present work. We use the same variables as in the previous comparisons, although we preserve their units — we center them, but do not divide through by their standard deviation. For Precipitation,

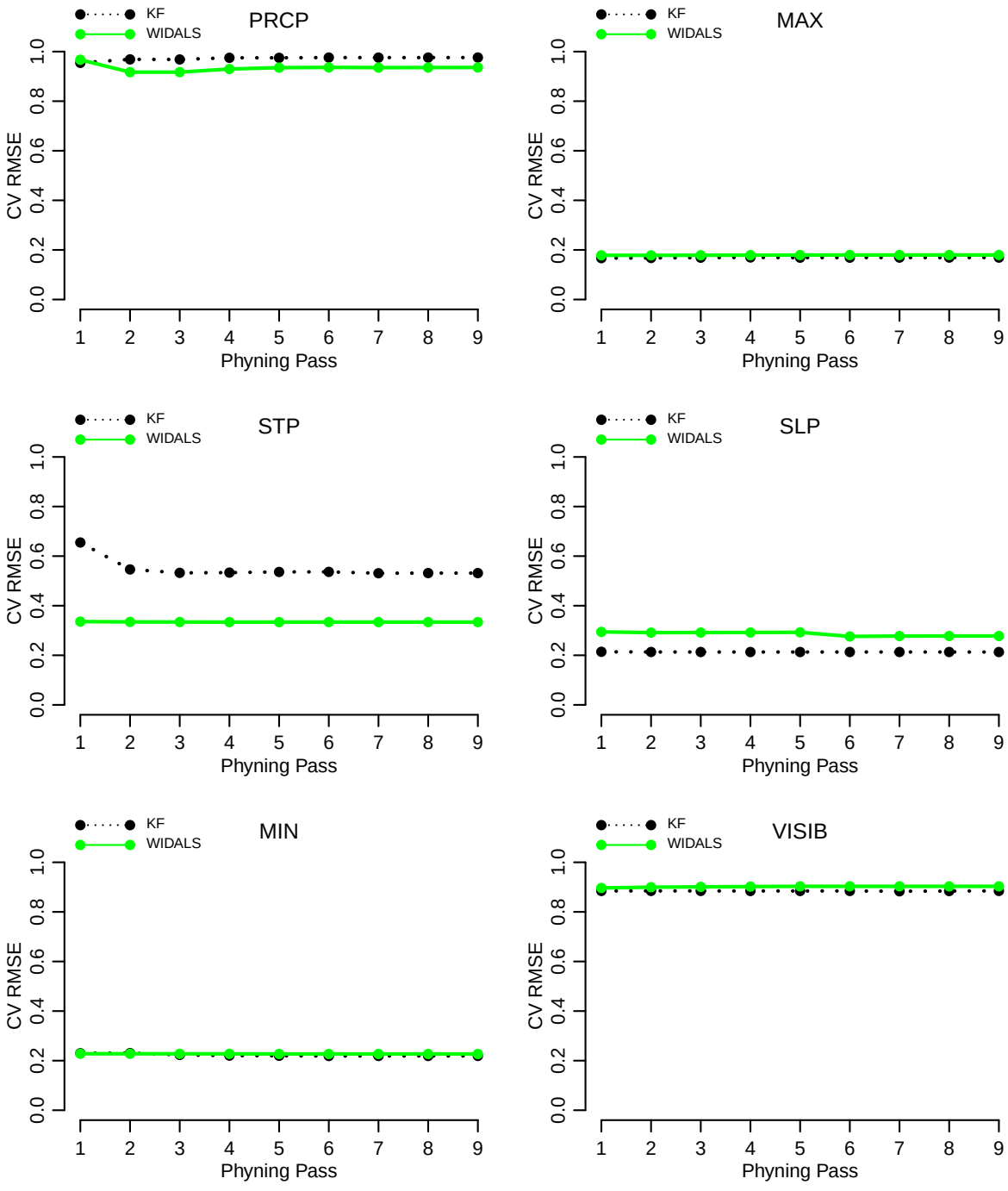

Figure 7.1: Imputation CV performance on tiny NCDC data.

we impute to 8386 sites; for the other 5, we impute to a union of 11669 sites, $\boldsymbol{\omega}^{\diamond}$ — of course, $\boldsymbol{\omega}^{\text{PRCP}} \subset \boldsymbol{\omega}^{\diamond}$, Table 7.3. Figure shows $\boldsymbol{\omega}^{\diamond}$.

Figure 7.5 shows both the observed densities and imputed densities of our 6 variables (note that Precipitation and Visibility are square-rooted to reign in right skewedness). For the moment, let us ignore the wide green path, and imagine we are at the inception of our project performing univariate preliminaries. Daily Precipitation, as expected, has a sharp mode at zero. Maximum Temperature is slightly asymmetric, with a mode at about 90 degrees Fahrenheit, with a sharp decline in temperatures greater than 90. Standard Temperature and Pressure has a range of about 550 to 1060, with a sharp mode at about 1020 millibars, with a quick decline in values greater than the mode. Sea Level Pressure possesses hints of normality, with a mode at about 1020 millibars. Daily Minimum Temperature suggests bi-modality, with some left-skewedness. Visibility reveals multi-modality.

Since our sights our set on imputing missing values across multiple data sets, it is sensible to create partial correlations between these variables and prospective exogenous covariates. Table 7.2 is the result of calculating partials on those *ca.* 74 thousand observations where values are present across all six variables. The essence can be distilled fairly easily. Precipitation is something of an outcast, being only very weakly related to the other 5 climate variables, as well as the three exogenous covariates ("elev"; "isa," the incident solar area; "isaxelev," the product of elevation and ISA). Visibility as well, though it is weakly related to temperature. Our two temperature variables are mutually related, and both are related to both pressure variables. Interestingly, the three exogenous covariates have but a modest partial relationship with our 6 climate variables.

Despite the likely weak contribution of PRCP and VISIB during imputation, we include them anyway for the sake of demonstration.

7.3.2 Imputation

The phyning imputation performed on the tiny data subsets in §7.2 is identically repeated, though on a grander scale. The CV performances can be found in two places, Figures 7.4 and 7.7. The latter zooms out to permit comparison with imputation using ALS alone.

Favorable testimony for WIDALS the quality of the WIDALS imputation can be found in two places. First, in Figure 7.5. With the exception of Precipitation, and perhaps Visibility, the distribution of the values imputed by phyning over WIDALS closely trace the values observed in the data. In fact, in the case of STP, the two densities trace virtually identical paths. Second, in Table 7.10, where we have our final fit statistics for both WIDALS and ALS alone. These values derive from deliberately removed data, the numbers of which can be found in the right column of Table 7.3. Without euphemism, both ALS and WIDALS blew up attempting to impute the extremely sparse Precipitation data set. Otherwise, our results are rather attractive. With Maximum Daily Temperature, the WIDALS MSE is about one-third that of ALS. With STP, the WIDALS MSE is about 70% that of ALS, with both methods explaining away about 99% of the total global variance. In the case of SLP, the WIDALS MSE is about one-seventh that of ALS. With Minimum Daily Temperature, the WIDALS MSE is less than one-half that of ALS. Finally, WIDALS explained about 4 times the global variance in Visibility than ALS.

The inability of either solution to work effectively on Precipitation is not particularly surprising. The data *is* extremely sparse, but moreover recall our earlier investigation on the observed partial correlations in Table 7.2, which speaks for Precipitation's relative independence from its brethren variables. It may be something of a curiosity that Visibility, also something of a renegade, though to a lesser degree, was as amenable to imputation as it was. Looking at the last column of

Table 7.2, it would seem an unlikely guess that — when everything is said and done — we were able to explain away almost 40% of Visibility's global variance.

7.3.3 Last Pass ALS Standard Errors

The ALS effective "standard errors" extracted from the WIDALS solution, as described in §5.2, after the final phyning pass are given for each of our six variables in Tables 7.4-7.9. We should be reminded that the "se's" reported here are somewhat *ad-hoc*. Without insulting the reader's intellect, we recall from elementary statistics that a standard error is the width of one standard deviation of the sampling distribution of the statistics under consideration — here, the partial slopes induced by ALS. While their construction presented at the end of §5.2 is certainly both intuitive and, to a point, mathematically rigorous, their value is inextricably tied to the solution's signal-to-noise ratio, ρ. However, ρ is a *free* hyperparameter — the researcher can crank the knob any which way he/she wishes. Even more, while no details are presented here, often with ALS, a dramatic change in ρ will have only an infinitesimal effect on the solution RMSE. With this said, we shall focus attention on the partial slopes, which have a much more visceral, and concrete interpretation.

Recall that our deterministic predictor matrix, $\mathbf{H}_t$ has been standardized throughout phyning, and also, that we centered our responses. This allows for at-a-glance interpretation of the resulting partial slopes.

Table 7.5 shows the partial slopes for Daily Maximum Temperature (MAX). Note the constant term is relatively close to zero. The dominant contributors are incident solar area (ISA) and the cohort variable Daily Minimum Temperature (MIN). All else equal, an increase in one standard deviation of ISA is associated with an average increase in about 7 degrees Fahrenheit in MAX (if we are at the average elevation, which is recorded as zero in $\mathbf{H}$, then the interaction between

ISA and Elevation does not contribute to MAX). All else equal, if MIN was one standard deviation higher today than it was 30 days ago, then we would expect MAX to be about 10 degrees higher than it was then.

Partial slopes for Standard Temperature and Pressure (STP) are given in Table 7.6. Again, the constant term is fairly close to zero. For the sake of illustration, consider two locations, ω_1 and ω_2, with the associated STP responses called z_1 and z_2. Suppose that the elevation of ω_2 is one standard deviation greater than that of ω_1, and also suppose that ω_2 is closer to the equator than ω_1 such that its ISA is one standard deviation greater. For convenience's sake, suppose that the resulting ISA-Elevation interaction for ω_2 is two standard deviations greater than that of ω_1. Then, everything else held constant, we approximately expect $z_1 = \mathrm{z}_2 + 1 \cdot (-72.5) + 1 \cdot (-0.1) + 2 \cdot (18.4) = \mathrm{z}_2 - 36$.

SLP, Table 7.7, is dominated by Elevation, STP, and MIN.

MIN, Table 7.8, is dominated by ISA and MAX.

VISIB, Table 7.9, is predicted, though weakly, by SLP and MIN. Recall that Visibility is under the square-root transform. All else the same, a two standard deviation increase in MIN is associated with an average visibility decrease of 0.51 $\sqrt{\text{miles}}$.

7.3.4 Global Interpolation

We finally use our filled data sets to extrapolate to 41,437 locations evenly spaced (approximately) on Earth's surface. Elevations for these points were acquired through the Google elevation API [Goo11] (Figure 7.3). First, each filled variable was refitted using MH to locate Π^{WID}. Note that $\mathbf{H}_t$ did not contain any response variables. We then constructed $\mathbf{H}_t^\star$, and finally interpolate our $n^\star \cdot t = 41437 \cdot 1016 = 42099992$ hat values.

Figure 7.6 shows the distribution of our interpolated values and original actual

values. The overall relationship between our actual and fitted values seems generally favorable. Of course, we do not expect anything close to perfect similarity — lurking variables present at $\boldsymbol{\omega}^{\star}$ will not be the same as those at $\boldsymbol{\omega}^{\diamond}$. Or, put another, more direct way, neither $\boldsymbol{\omega}^{\star}$ nor $\boldsymbol{\omega}^{\diamond}$ can be expected to be representative of $\boldsymbol{\Omega}$.

In MAX, STP, SLP, MIN, the fitted values have more pronounced nodes. Along the way during phyning and interpolation, the spiky, multi-modality of VISIB got smoothed. Interestingly, in MIN, a hint of the bi-modality in the original data is manifest in the extrapolated values.

Quite possibly the most dramatic testimony for WIDALS is revealed in Figure 7.8. By the time WIDALS has finished imputation through phyning, the devil-may-care spacial covariation in, e.g., Maximum Daily Temperature, has all but vanished.

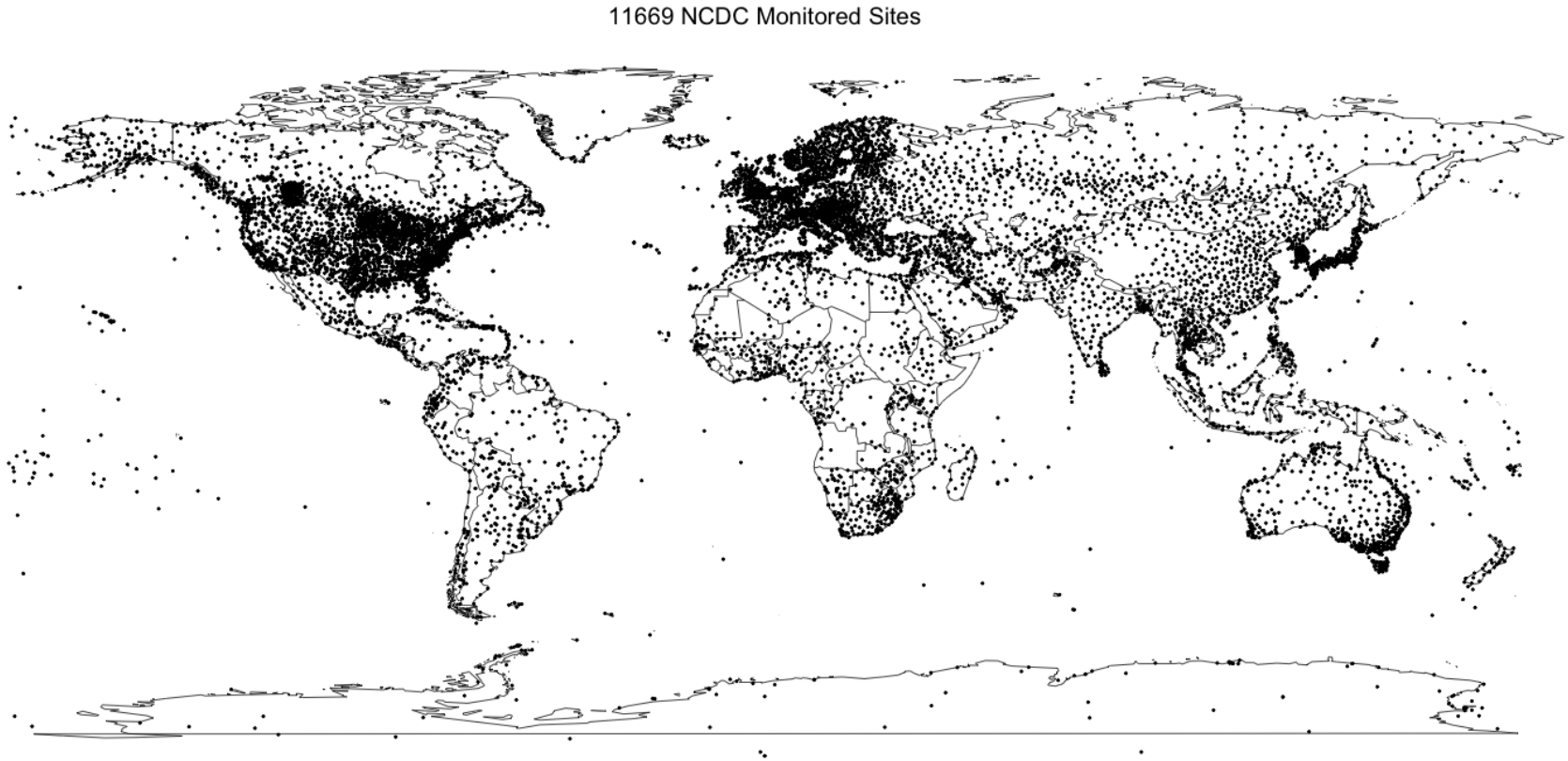

Figure 7.2: Loaction of 11669 NCDC climate sites.

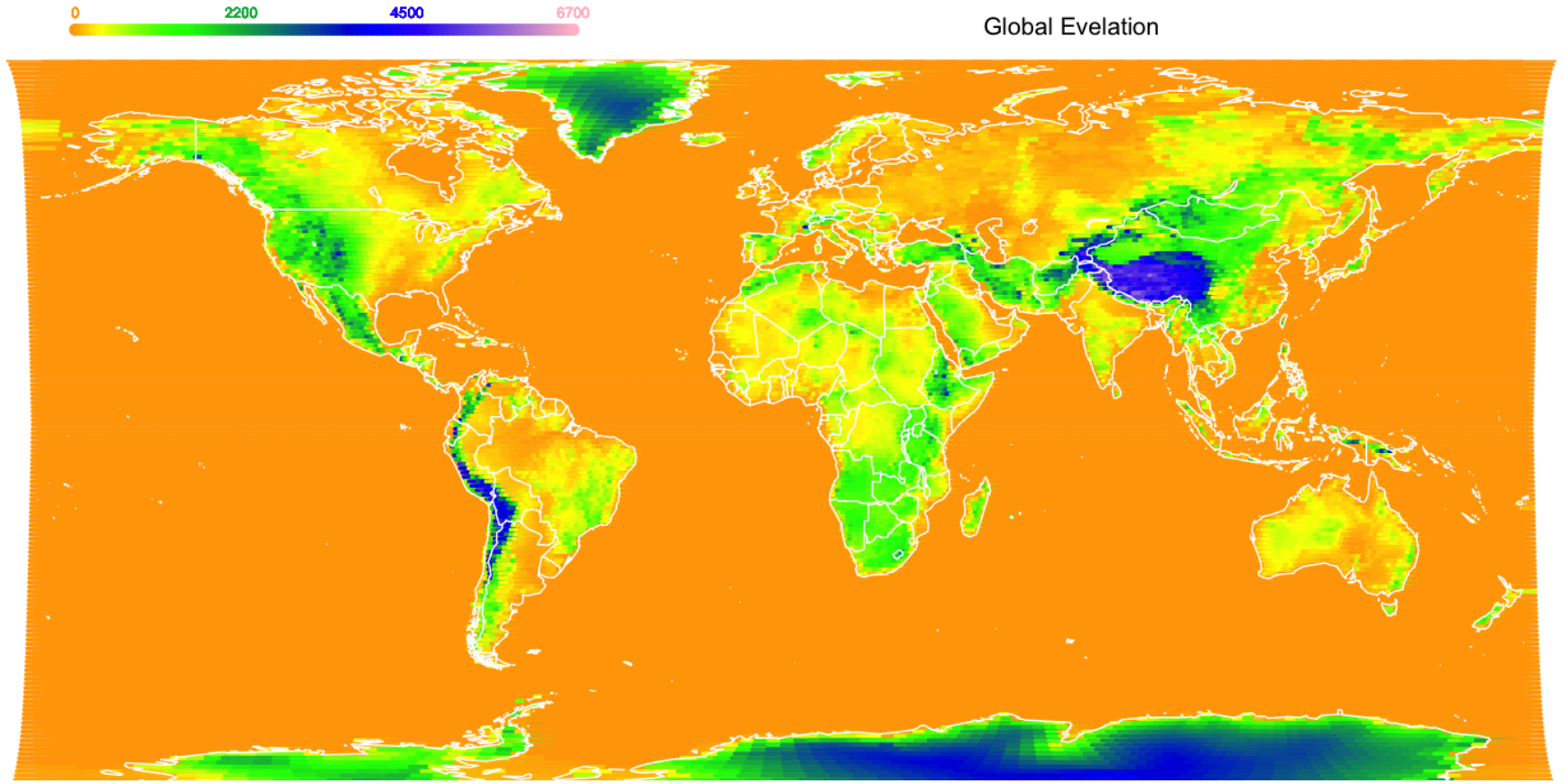

Figure 7.3: Global Elevation map, from Google Elevation API (units in meters).

	elev	isa	isaxelev	prcp	max	stp	slp	min	visib
elev	1.000	−0.423	0.948	−0.010	−0.038	−0.044	0.046	0.029	−0.026
isa	−0.423	1.000	0.484	0.084	0.171	0.043	−0.094	−0.024	0.069
isaxelev	0.948	0.484	1.000	0.012	0.024	0.042	−0.034	−0.019	0.018
prcp	−0.010	0.084	0.012	1.000	0.004	0.021	−0.038	−0.003	0.027
max	−0.038	0.171	0.024	0.004	1.000	−0.350	0.172	0.915	0.113
stp	−0.044	0.043	0.042	0.021	−0.350	1.000	0.265	0.360	0.049
slp	0.046	−0.094	−0.034	−0.038	0.172	0.265	1.000	−0.300	0.030
min	0.029	−0.024	−0.019	−0.003	0.915	0.360	−0.300	1.000	−0.186
visib	−0.026	0.069	0.018	0.027	0.113	0.049	0.030	−0.186	1.000

Table 7.2: The Partial Correlations between predictors/data using 74390 data points — the total number of space-time entries where values are present across all 6 climate variables.

7.3.5 Video Snapshots

For MAX, Figures 7.9-7.11 show raster maps of our predicted values at the turn of the years 2009 and 2010. A keen eye will spot, among plenty of other things, a polar blast blowing down through Canada into the U.S. Midwest.

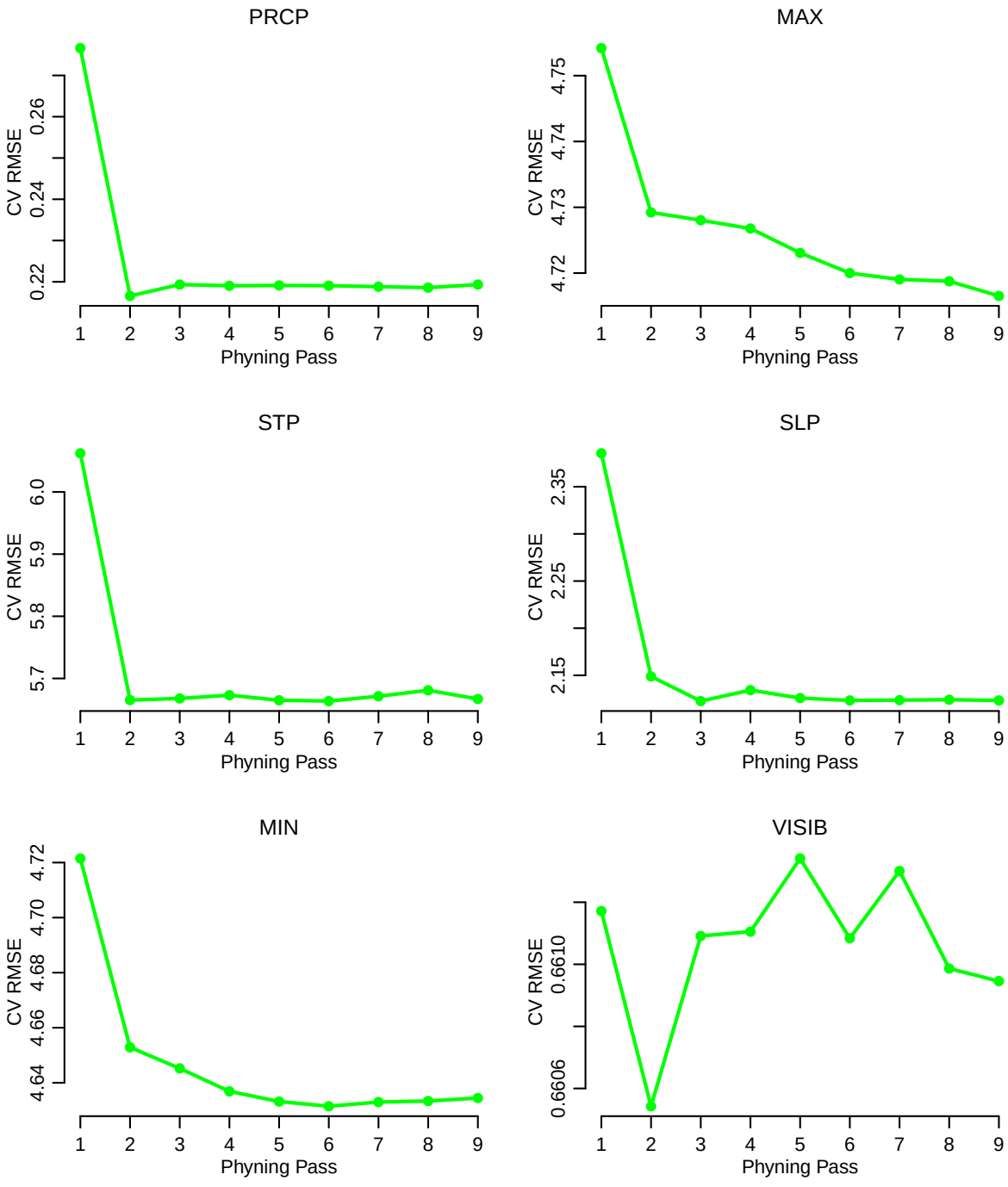

Figure 7.4: Imputation CV performance on Large NCDC data.

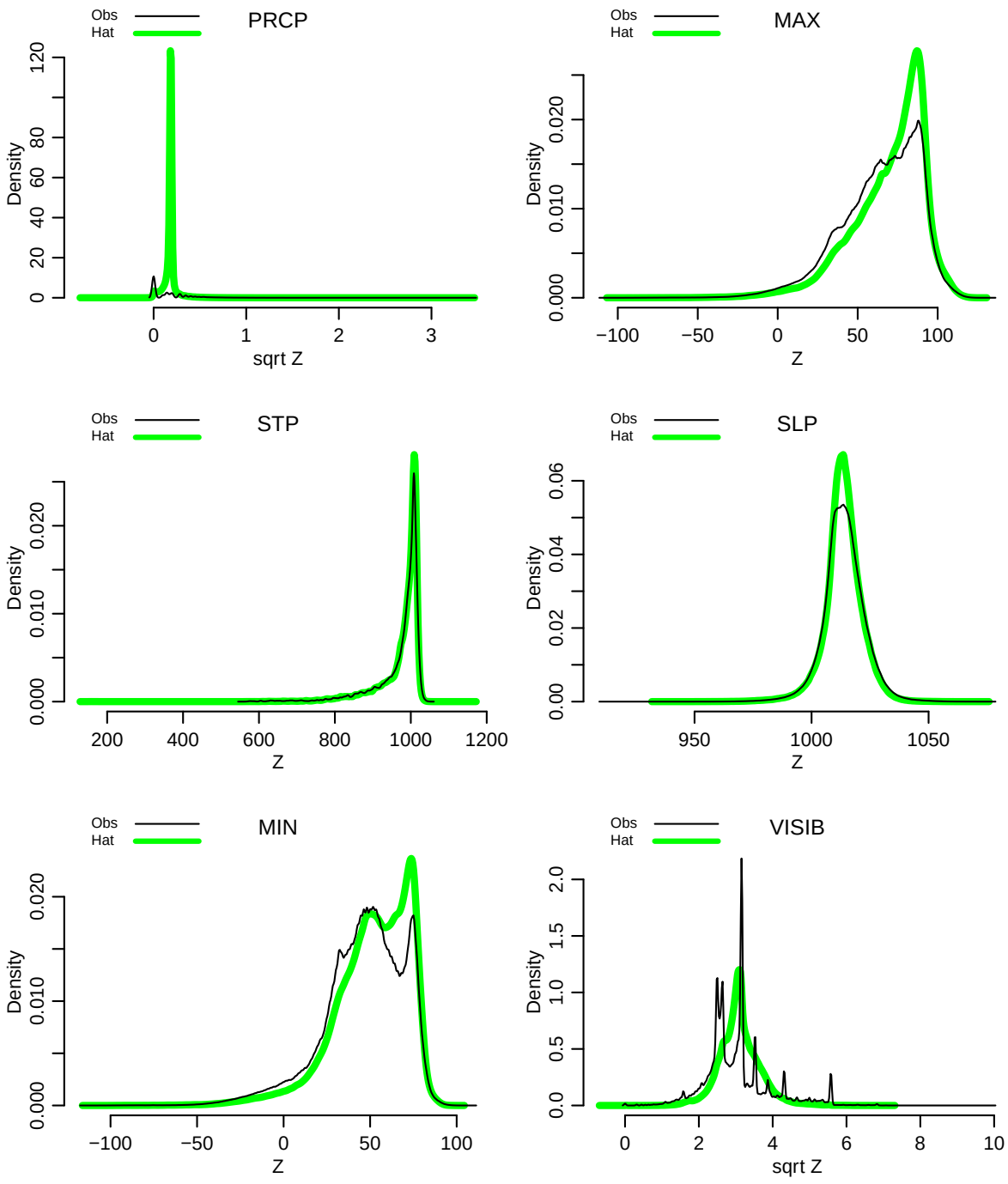

Figure 7.5: Distributions of actual values and values predicted by phyning — Large NCDC data.

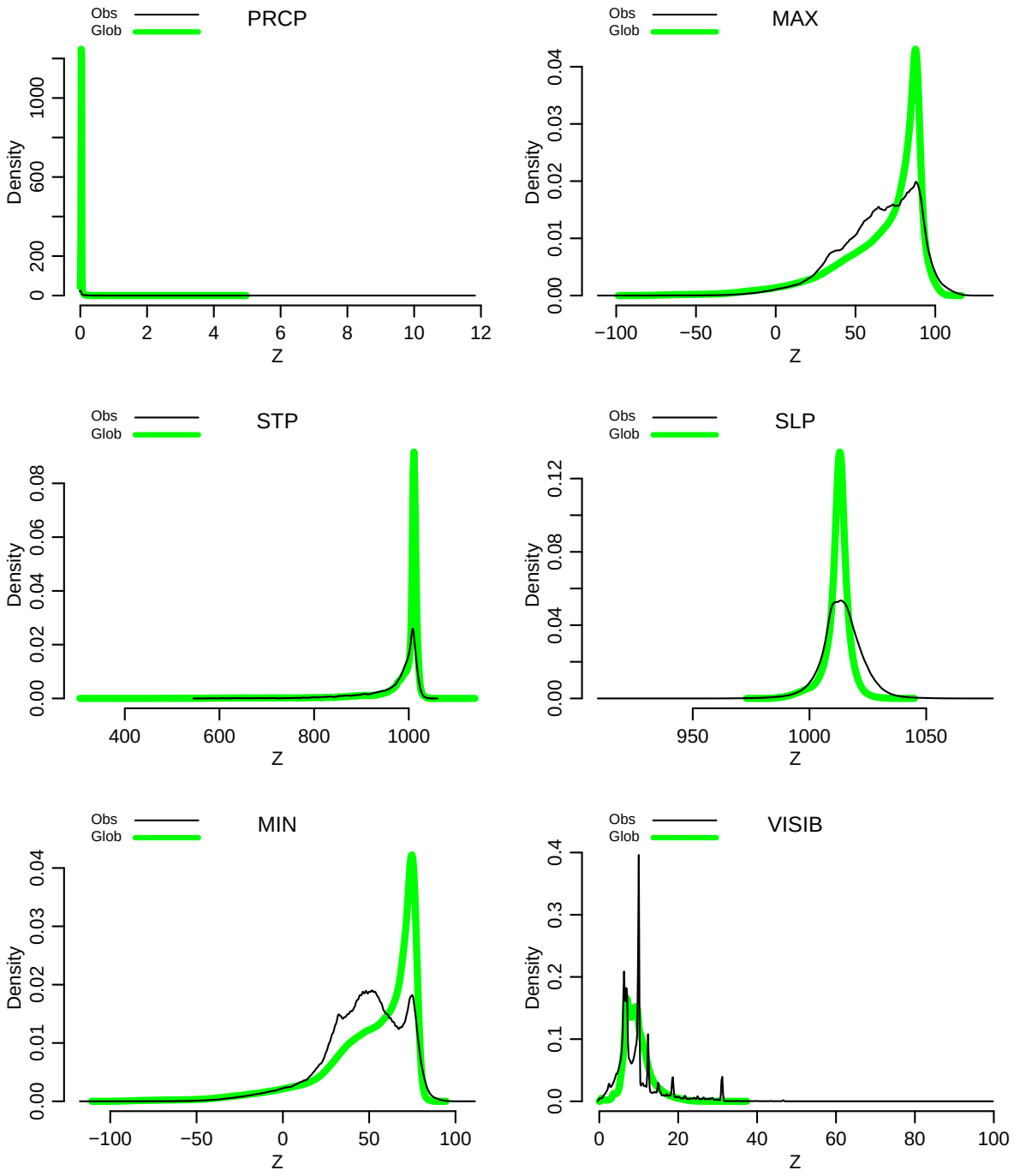

Figure 7.6: Distributions of actual values and WIDALSs globally extrapolated values — Large NCDC data. Note that PRCP and VISIB have been returned to their original units.

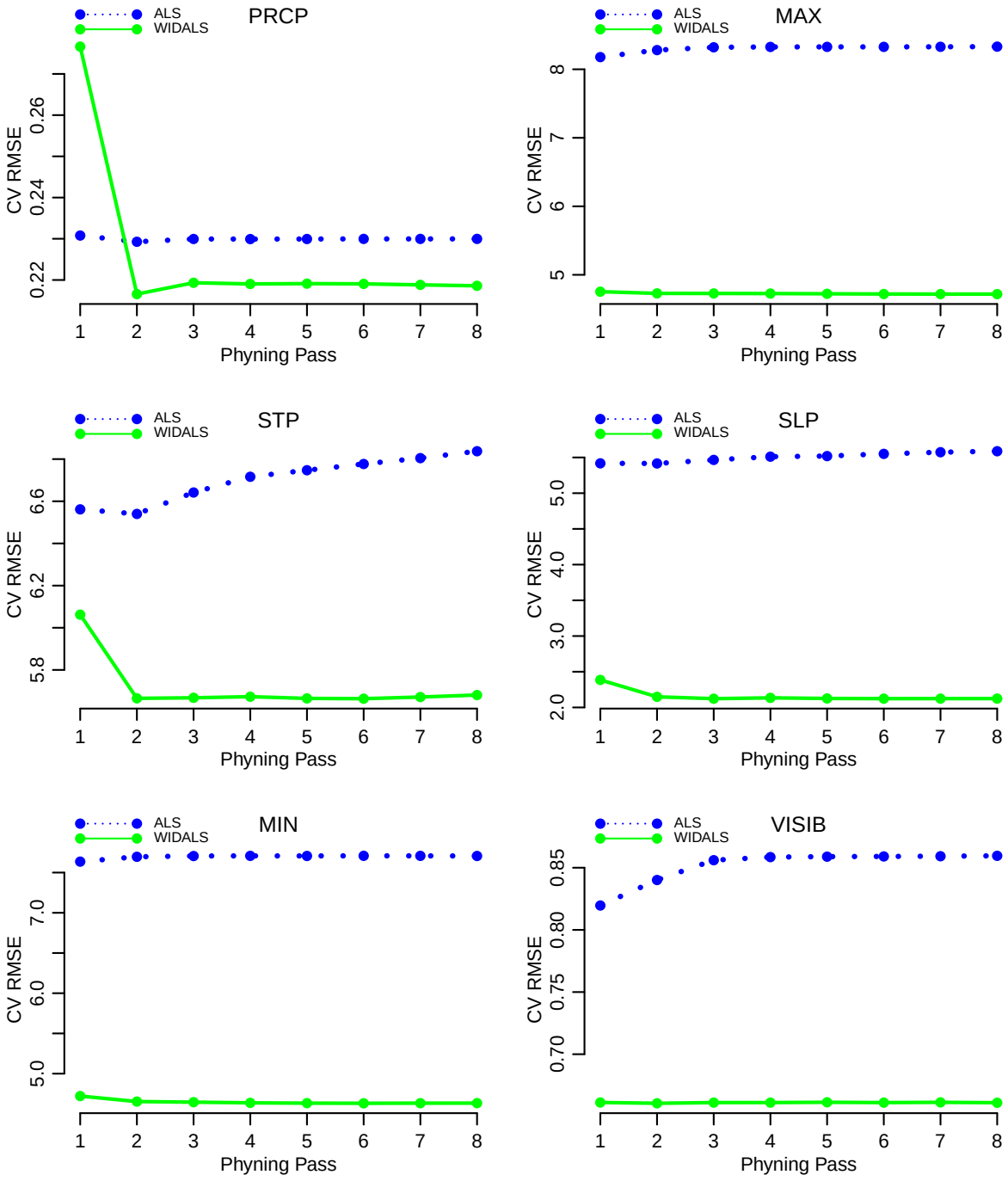

Figure 7.7: Imputation CV performance of WIDALS and ALS on Large NCDC data.

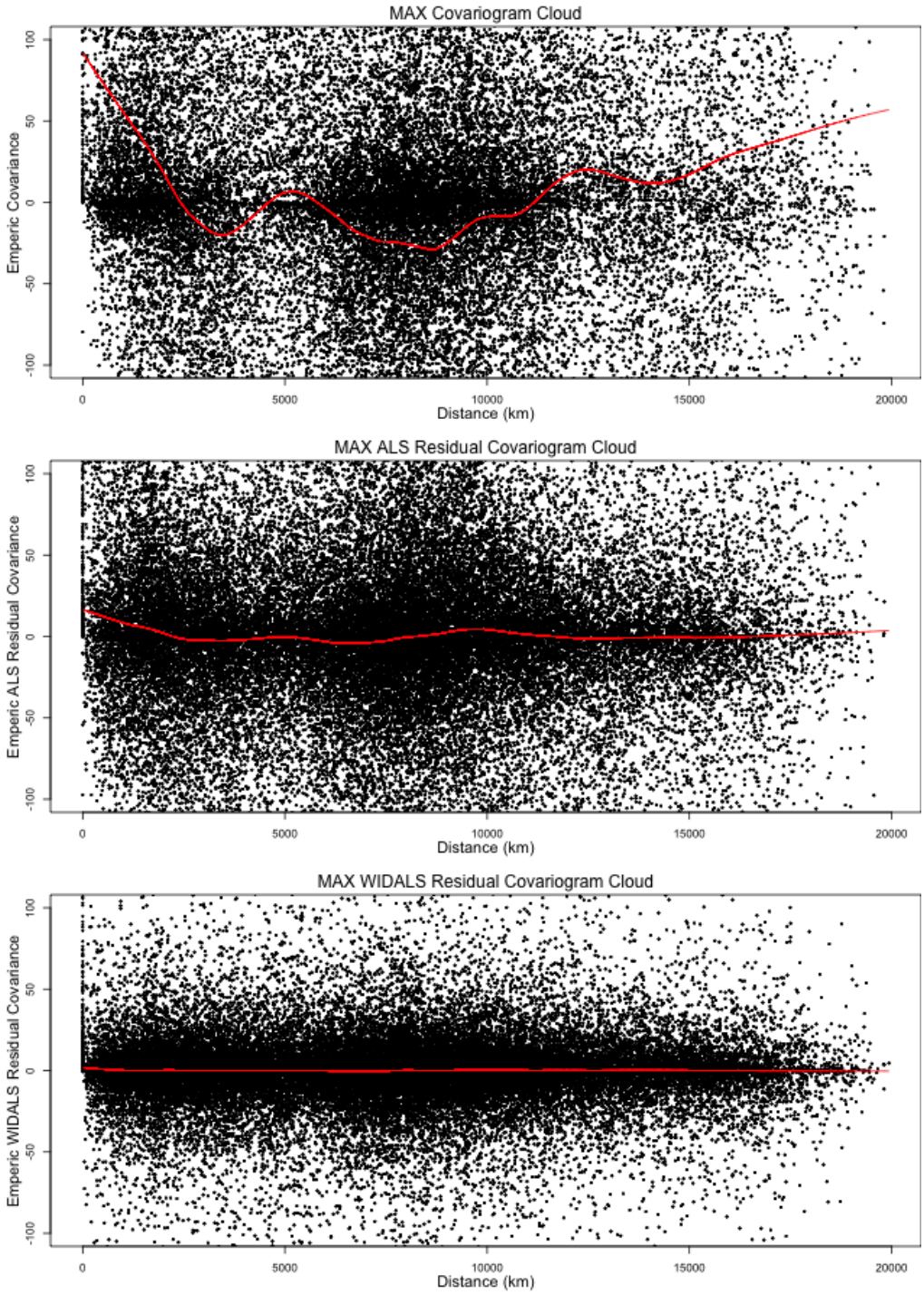

Figure 7.8: From a sample of CV sites-days. Covariogram cloud for MAX (top); ALS residuals covariogram cloud (middle); WIDALS residuals covariogram cloud (bottom). The WIDALS mimic residual smoothing function, **W**, has all but stripped away spacial covariation.

NCDC Name	n Sites	Units	n Total NA	n CV NA
PRCP	8386	inches	8249023	1391
MAX	11669	deg F	2013050	47439
STP	11669	millibars	6973999	23472
SLP	11669	millibars	5489088	30738
MIN	11669	deg F	2012237	47088
VISIB	11669	miles	4322650	35818

Table 7.3: NCDC Data info with missing entry counts for large data sets (11669 sites).

	est/prediction	als.se
const	−1.296489	0.45974
elevation	−0.017802	0.00665
sinet	−0.878586	0.30774
cosinet	0.104614	0.04399
ISA	0.013302	0.00143
ISA x elev	0.008040	0.00367
MAX	−0.020050	0.00217
STP	−0.008355	0.00464
SLP	−0.003850	0.00102
MIN	0.004851	0.00220
VISIB	−0.010901	0.00069

Table 7.4: Phyning last pass Standard Errors, PRCP

	est/prediction	als.se
const	1.1752	0.00731
elevation	−0.5174	0.01018
sinet	−2.2067	0.00672
cosinet	−1.5681	0.00813
ISA	6.9059	0.00893
ISA x elev	1.1405	0.00963
PRCP	−0.2841	0.00654
STP	0.4921	0.01007
SLP	−0.3467	0.00726
MIN	10.0075	0.00890
VISIB	−0.3504	0.00698

Table 7.5: Phyning last pass Standard Errors, MAX

	est/prediction	als.se
const	1.183508	0.78075
elevation	−72.520405	0.16820
sinet	−1.709145	0.56046
cosinet	0.425691	0.91856
ISA	−0.097295	0.09099
ISA x elev	18.444199	0.17977
PRCP	−0.407095	0.04721
MAX	−4.088403	0.13213
SLP	9.651354	0.04675
MIN	3.815448	0.13283
VISIB	0.005892	0.04585

Table 7.6: Phyning last pass Standard Errors, STP

	est/prediction	als.se
const	−0.1263	0.00202
elevation	4.0642	0.00554
sinet	−0.3290	0.00219
cosinet	1.1442	0.00239
ISA	1.8097	0.00362
ISA x elev	1.3604	0.00456
PRCP	−0.2689	0.00204
MAX	−0.1154	0.00431
STP	6.2719	0.00517
MIN	−3.0223	0.00427
VISIB	0.6133	0.00210

Table 7.7: Phyning last pass Standard Errors, SLP

	est/prediction	als.se
const	0.50812	0.00278
elevation	−1.26071	0.00347
sinet	−1.71248	0.00279
cosinet	−1.57619	0.00296
ISA	5.46759	0.00320
ISA x elev	0.01371	0.00341
PRCP	0.46787	0.00278
MAX	8.08464	0.00316
STP	1.33421	0.00345
SLP	−1.74512	0.00282
VISIB	−1.45398	0.00281

Table 7.8: Phyning last pass Standard Errors, MIN

	est/prediction	als.se
const	0.009507	0.01929
elevation	0.009617	0.01769
sinet	−0.013700	0.01422
cosinet	0.003564	0.02254
ISA	−0.079800	0.01138
ISA x elev	0.005436	0.01550
PRCP	−0.076599	0.00711
MAX	0.028615	0.01427
STP	−0.004810	0.01581
SLP	0.126330	0.00730
MIN	−0.258056	0.01428

Table 7.9: Phyning last pass Standard Errors, VISIB

	PRCP	MAX	STP	SLP	MIN	VISIB
Total Variance	0.0725	584.2371	3822.1845	78.8103	524.6468	38.3920
WIDALS CV MSE	0.0858	22.2873	32.3704	4.5165	21.4639	23.9033
WIDALS CV R^2	−0.1826	0.9619	0.9915	0.9427	0.9591	0.3774
ALS CV MSE	0.0742	69.4047	46.7444	31.0284	59.3806	34.8967
ALS CV R^2	−0.0232	0.8812	0.9878	0.6063	0.8868	0.0910

Table 7.10: Final Assessment of imputation quality. Amount of variance explained by WIDALS, and ALS alone, on deliberately removed values. (PRCP and VISIB returned to their original units.)

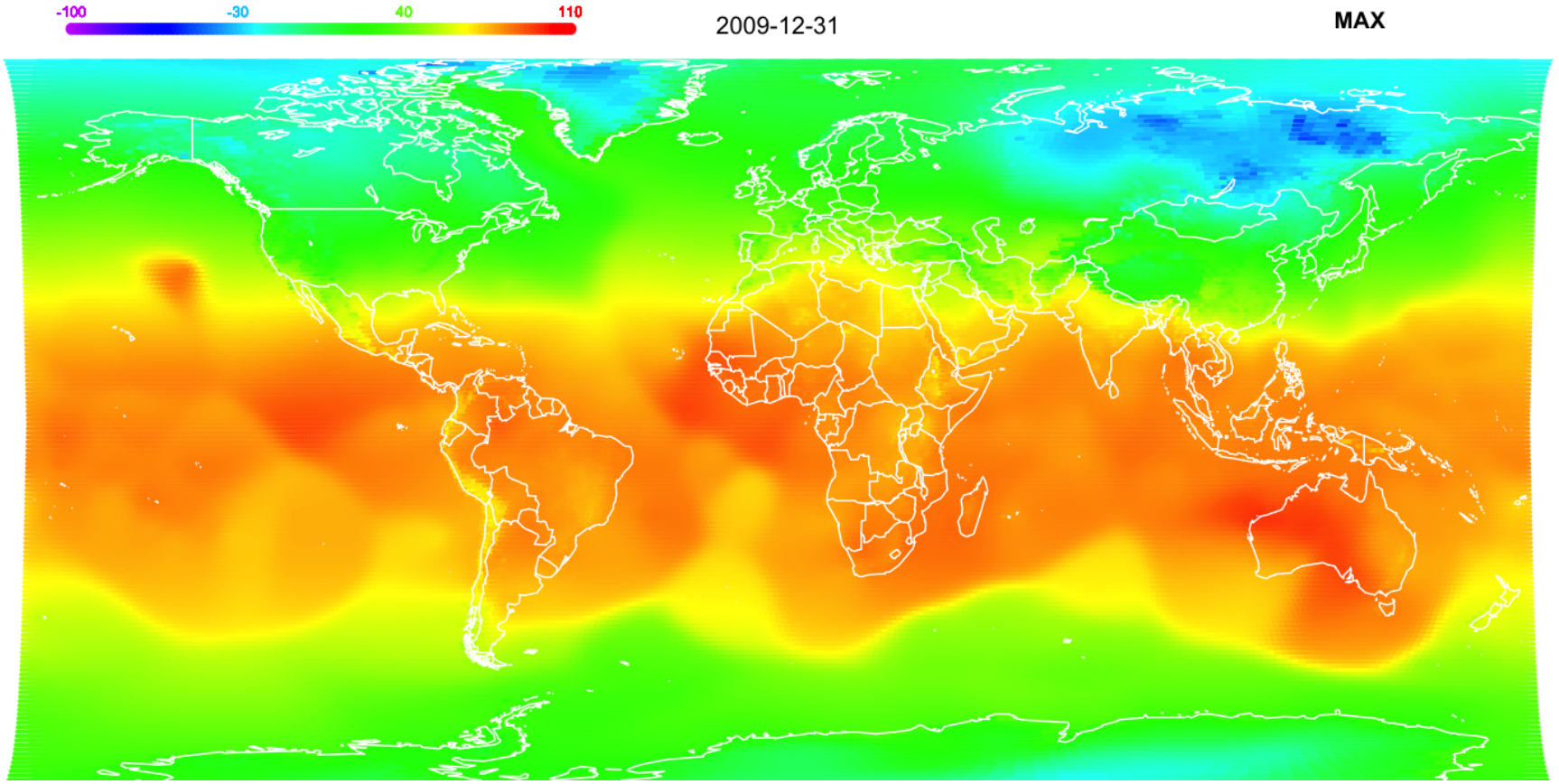

Figure 7.9: Snapshot of globally imputed MAX daily Temperature, using WIDALS.

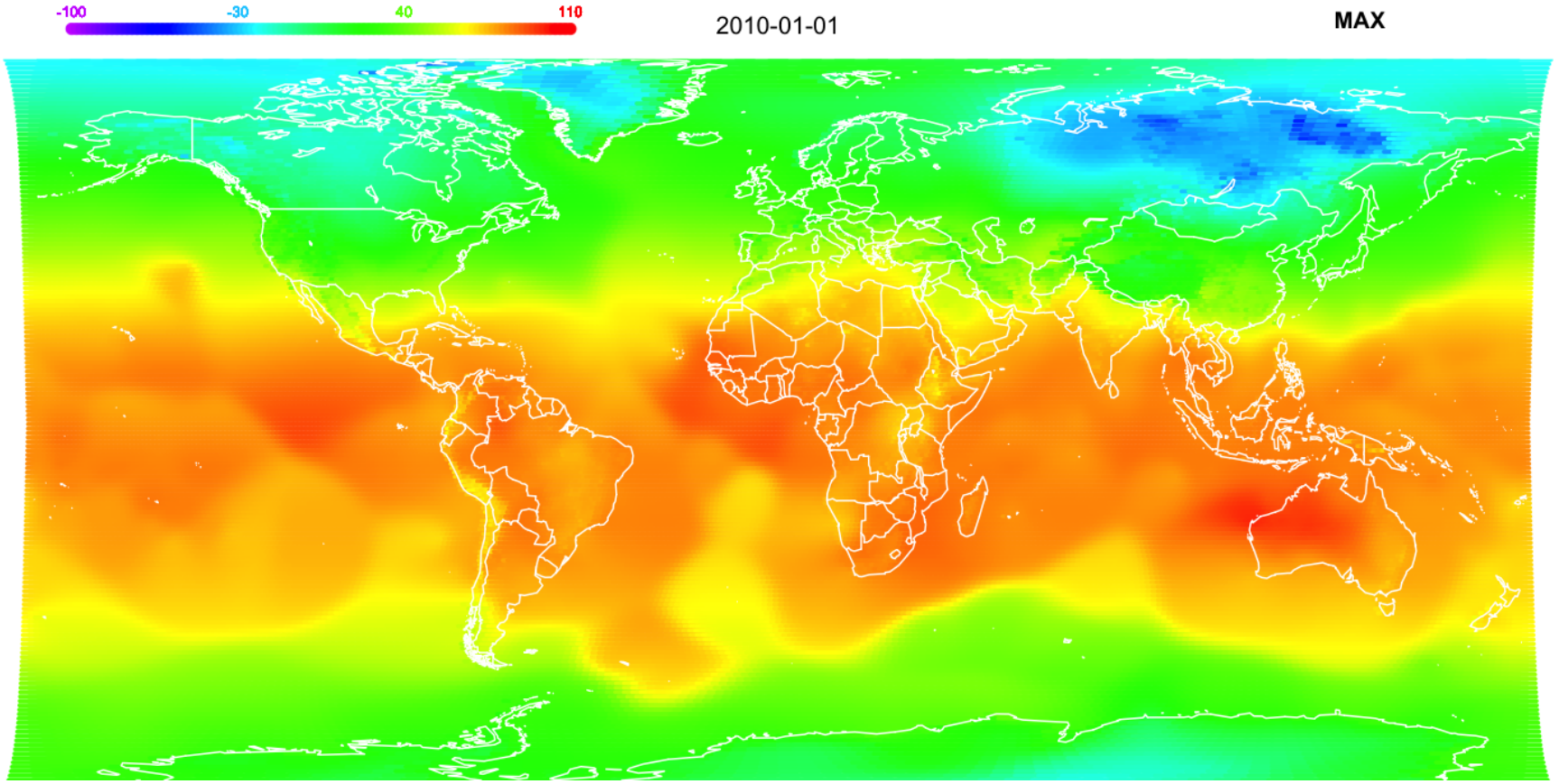

Figure 7.10: Snapshot of globally imputed MAX daily Temperature, using WIDALS.

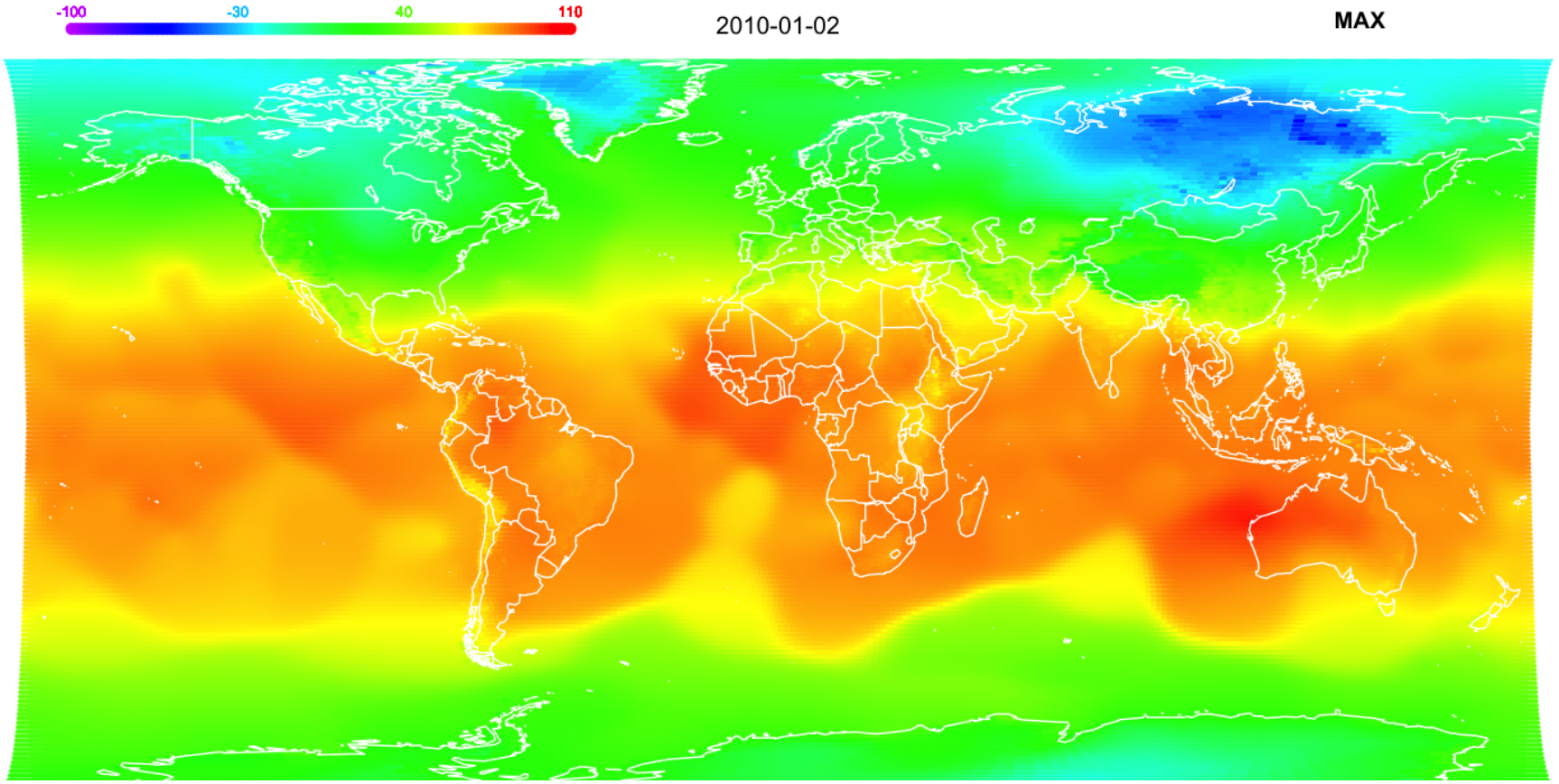

Figure 7.11: Snapshot of globally imputed MAX daily Temperature, using WIDALS.

For MIN, Figures 7.12-7.14 show raster maps of our predicted values for the same three days. The downward migration of the polar cold-front is present.

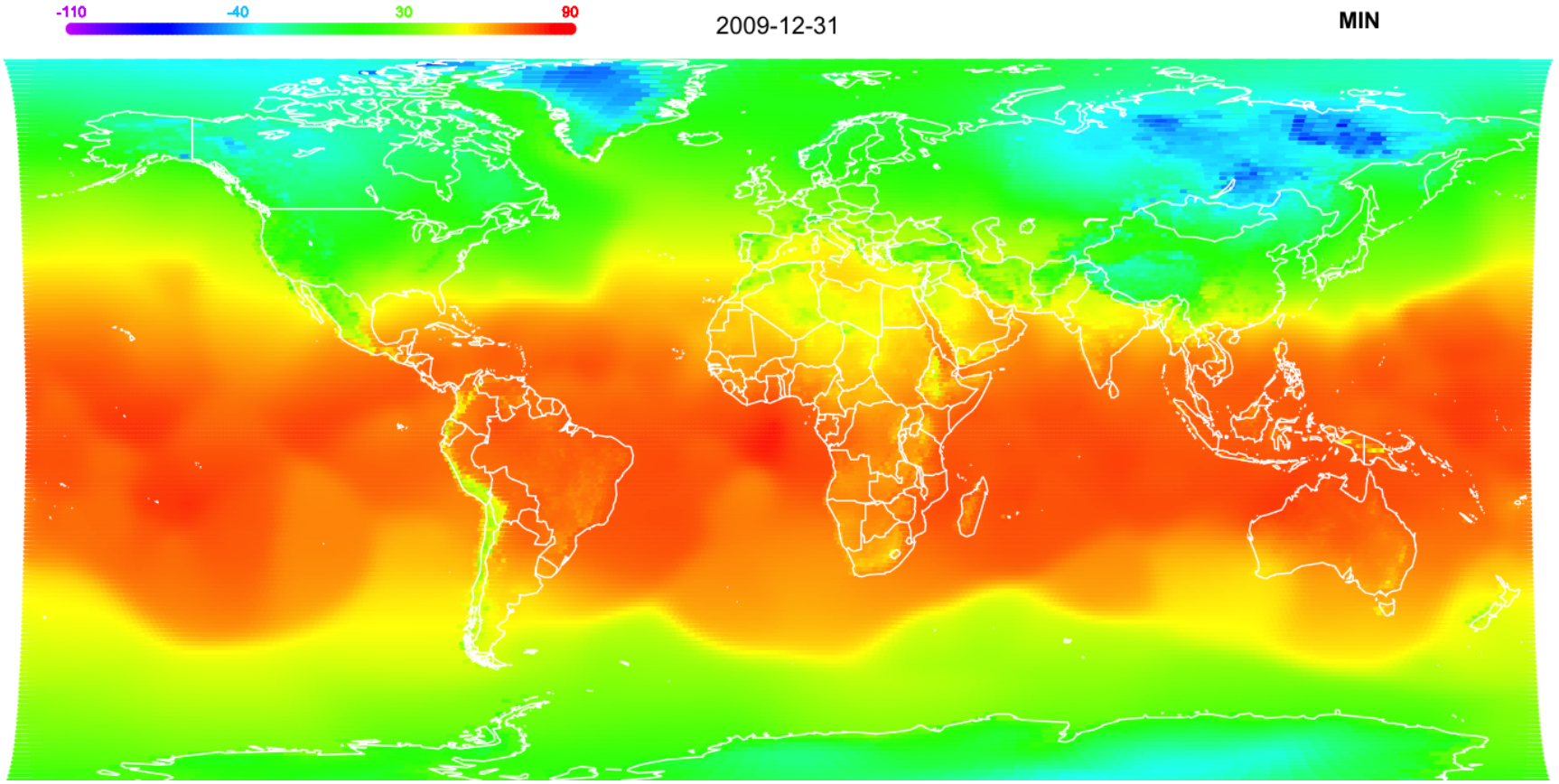

Figure 7.12: Snapshot of globally imputed MIN daily Temperature, using WIDALS.

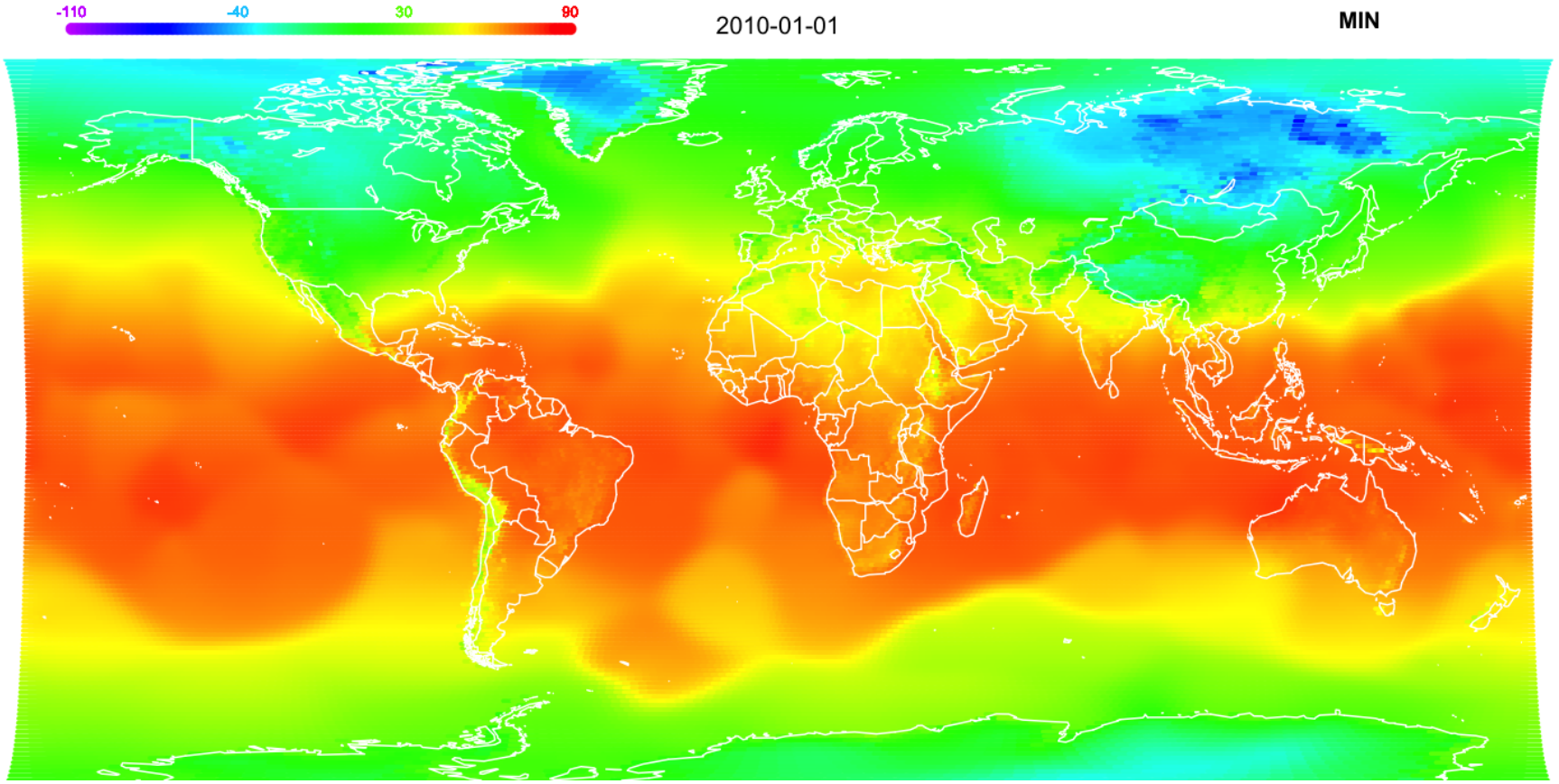

Figure 7.13: Snapshot of globally imputed MIN daily Temperature, using WIDALS.

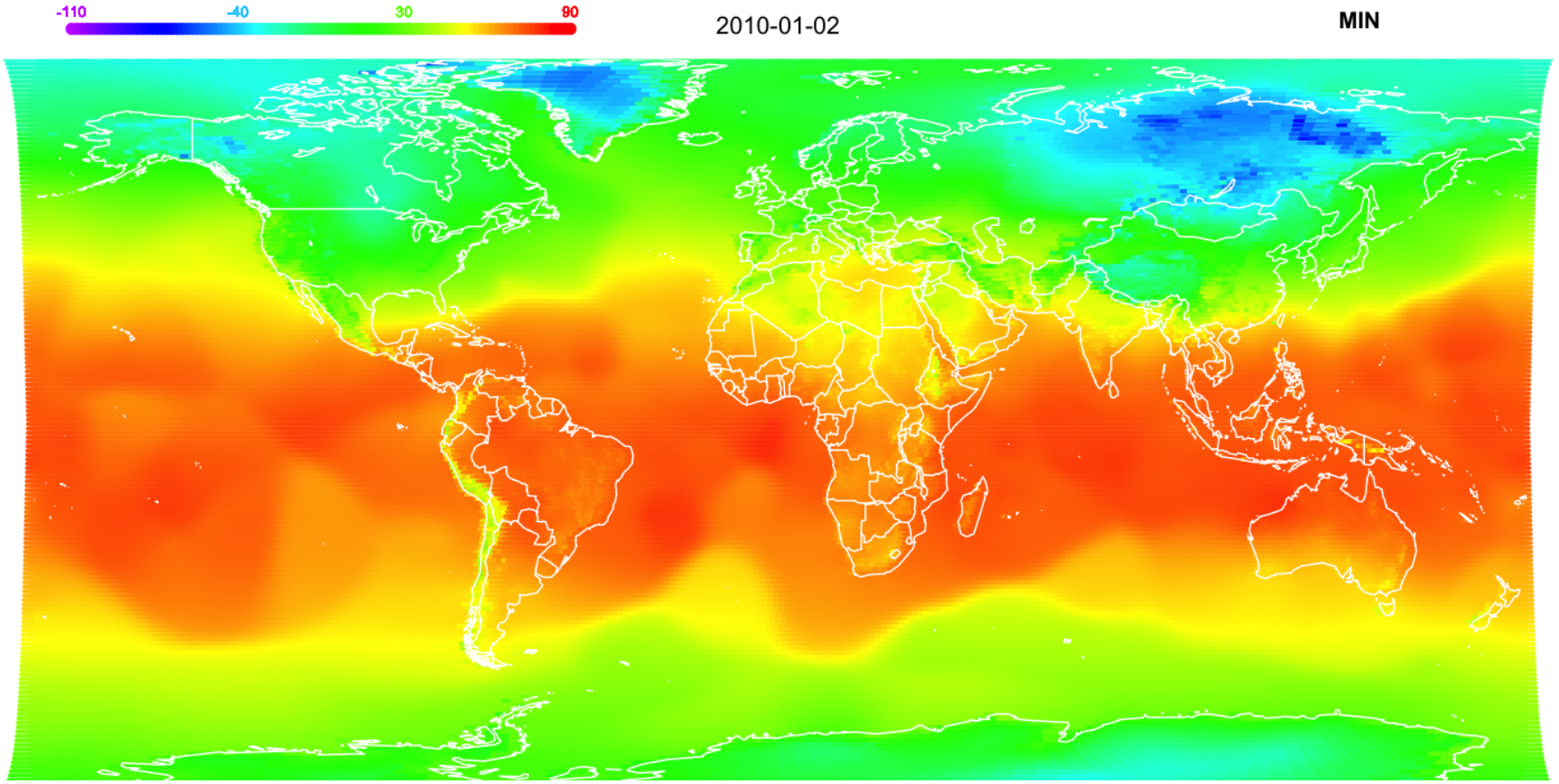

Figure 7.14: Snapshot of globally imputed MIN daily Temperature, using WIDALS.

For STP, Figures 7.15-7.17 show raster maps of our predicted values for the same three days. STP is fairly time-invariant.

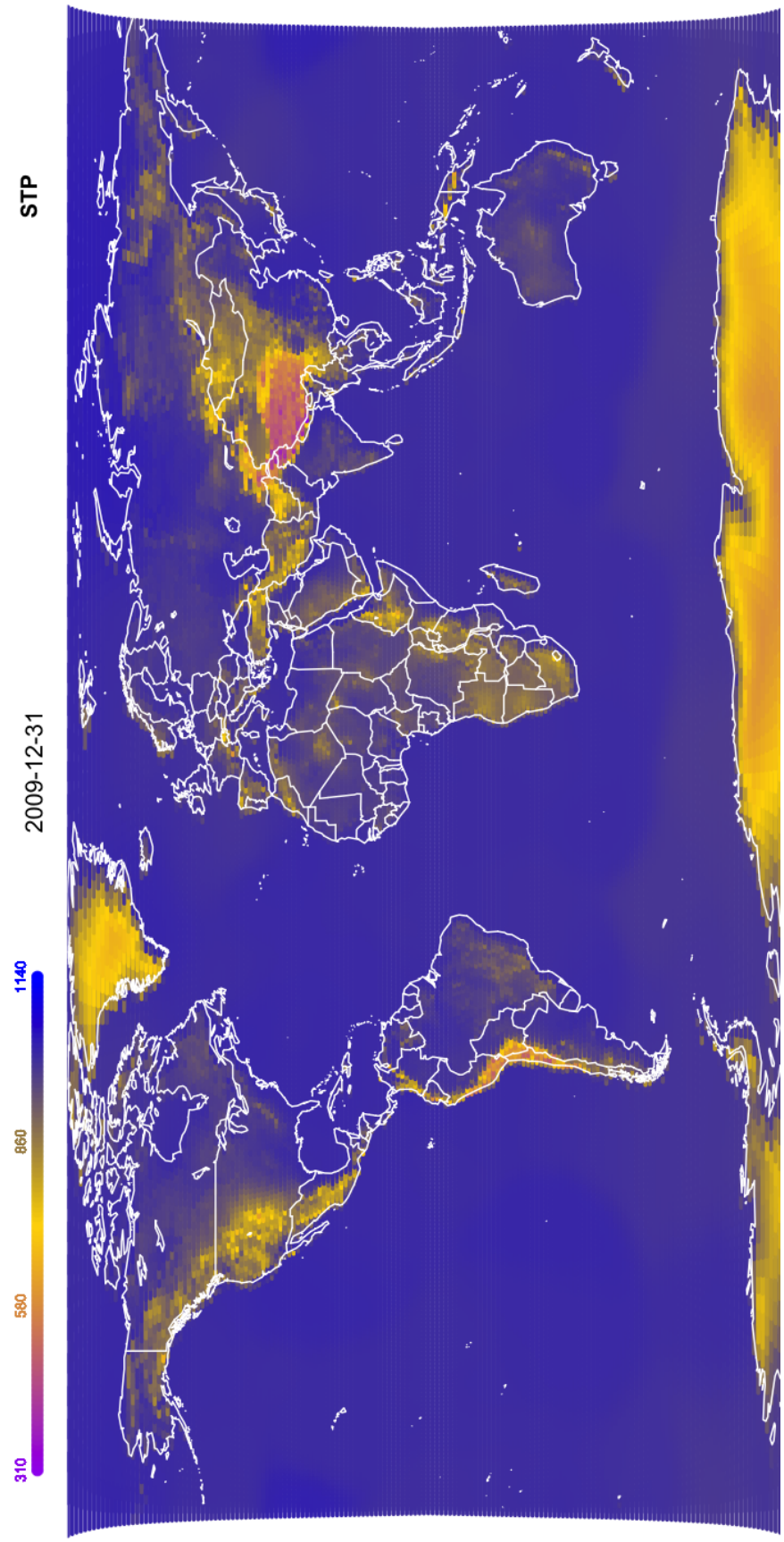

Figure 7.15: Snapshot of globally imputed STP daily Temperature, using WIDALS.

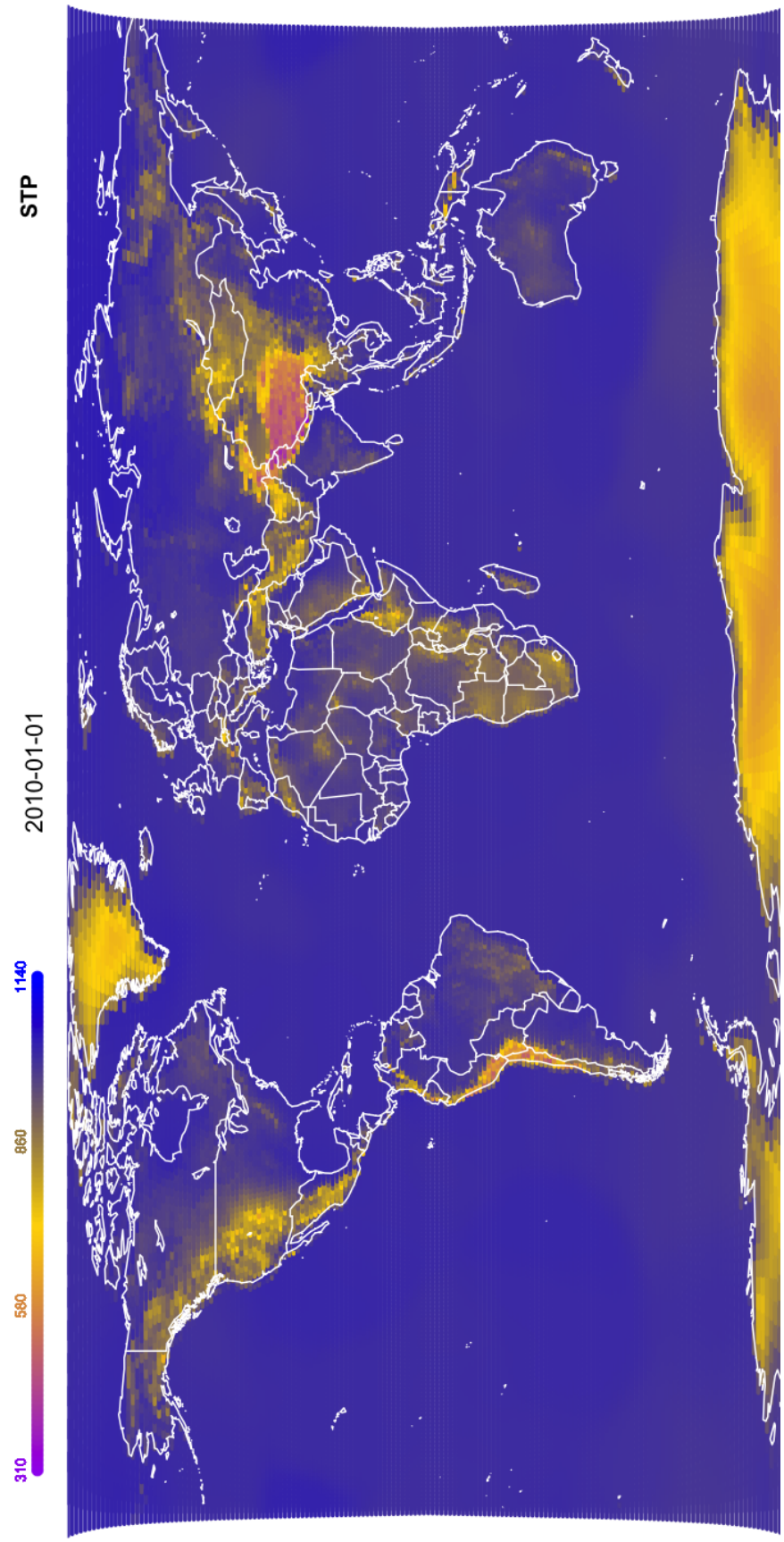

Figure 7.16: Snapshot of globally imputed STP daily Temperature, using WIDALS.

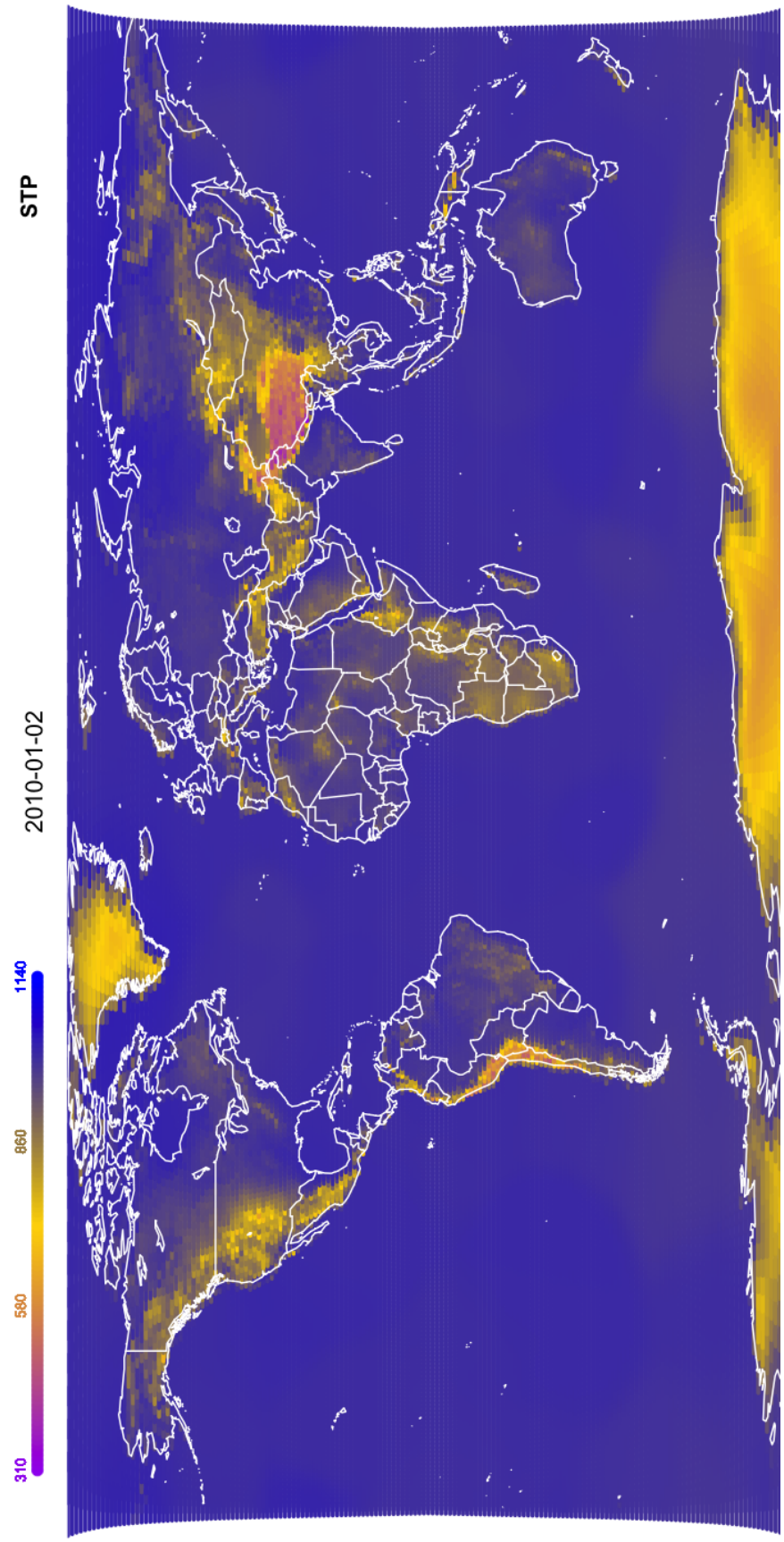

Figure 7.17: Snapshot of globally imputed STP daily Temperature, using WIDALS.

For SLP, Figures 7.18-7.20 show raster maps of our predicted values for the same three days. The downward migration of the polar cold-front, as manifest by SLP, can be clearly seen.

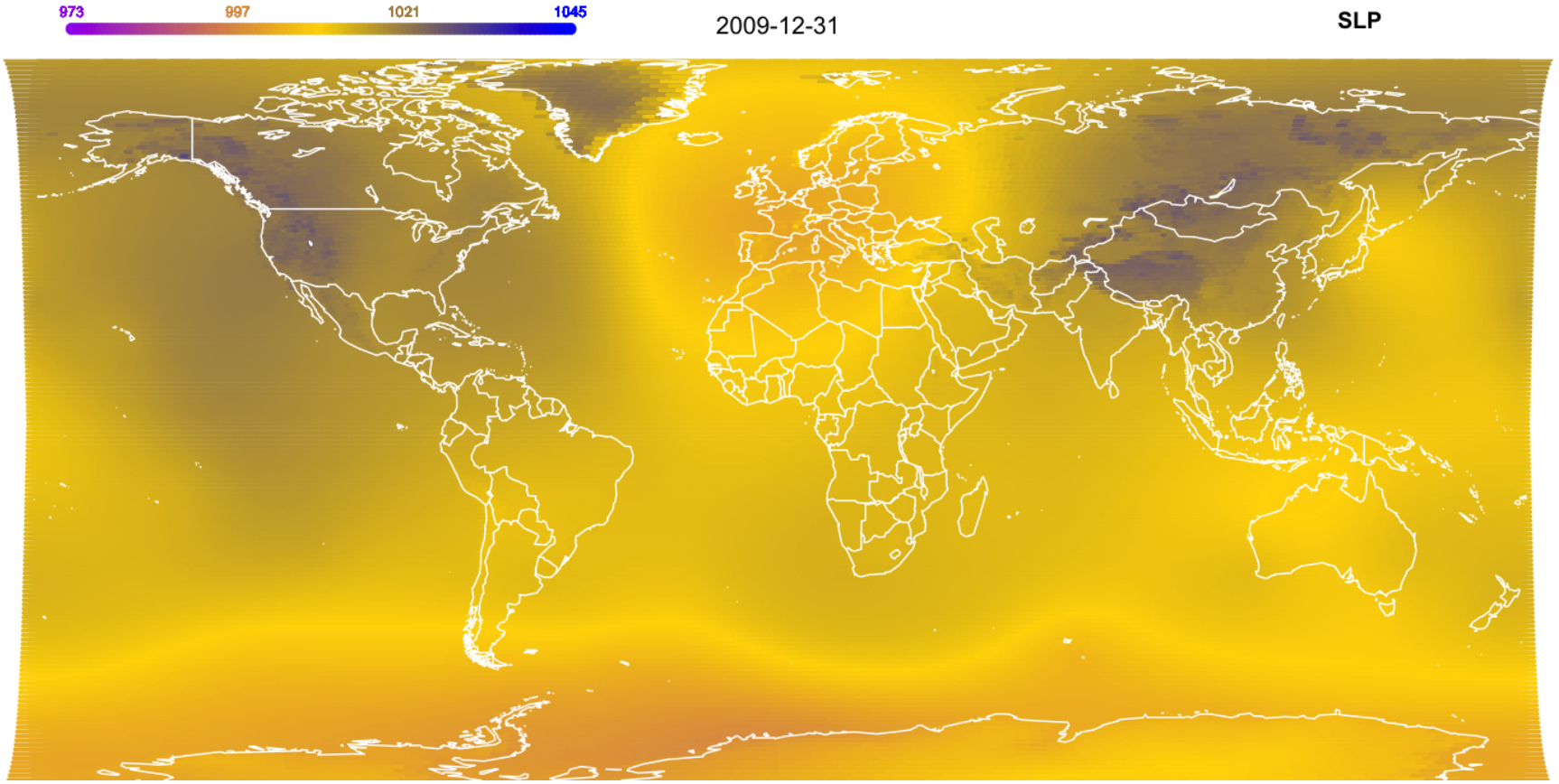

Figure 7.18: Snapshot of globally imputed SLP daily Temperature, using WIDALS.

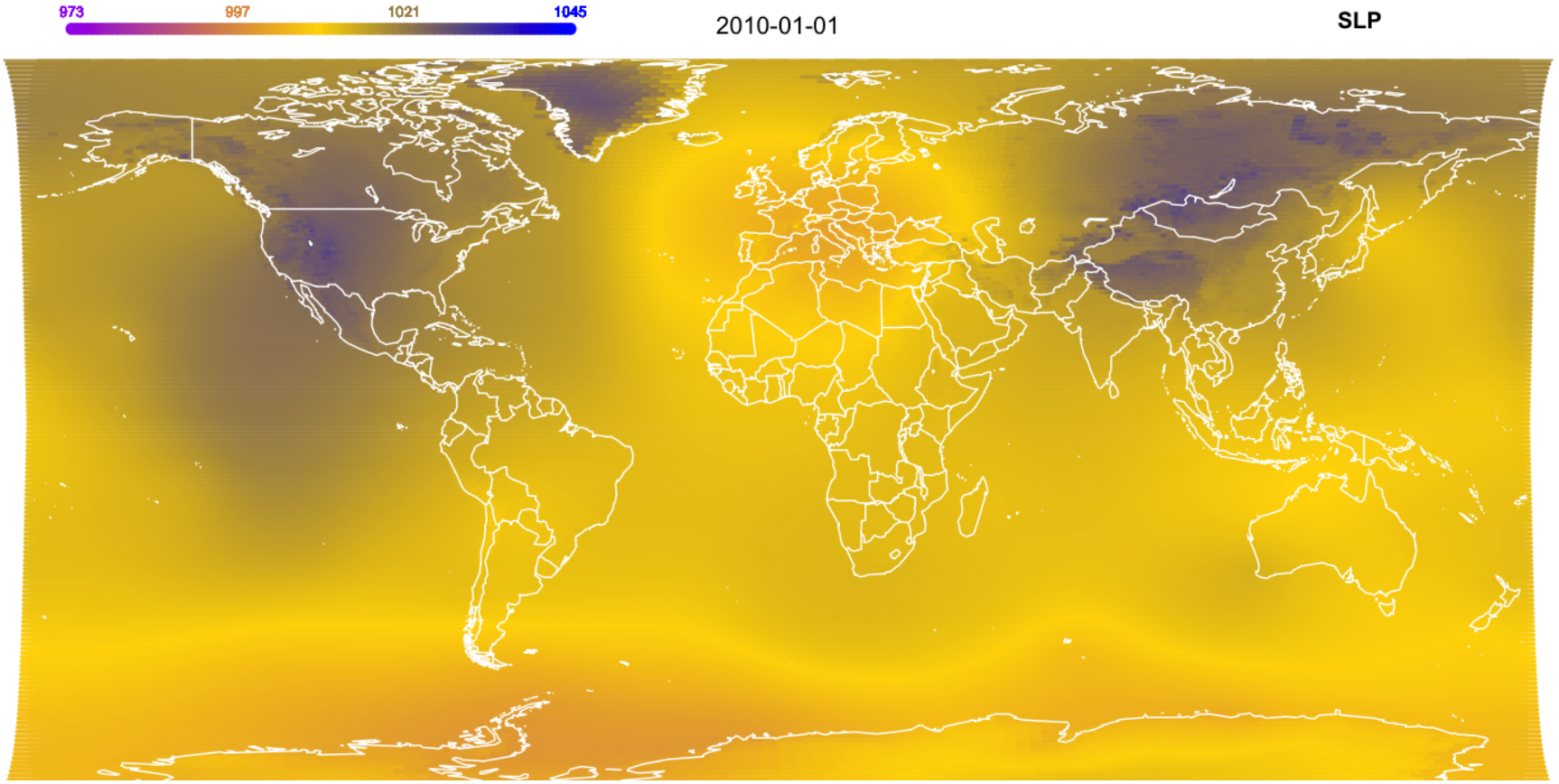

Figure 7.19: Snapshot of globally imputed SLP daily Temperature, using WIDALS.

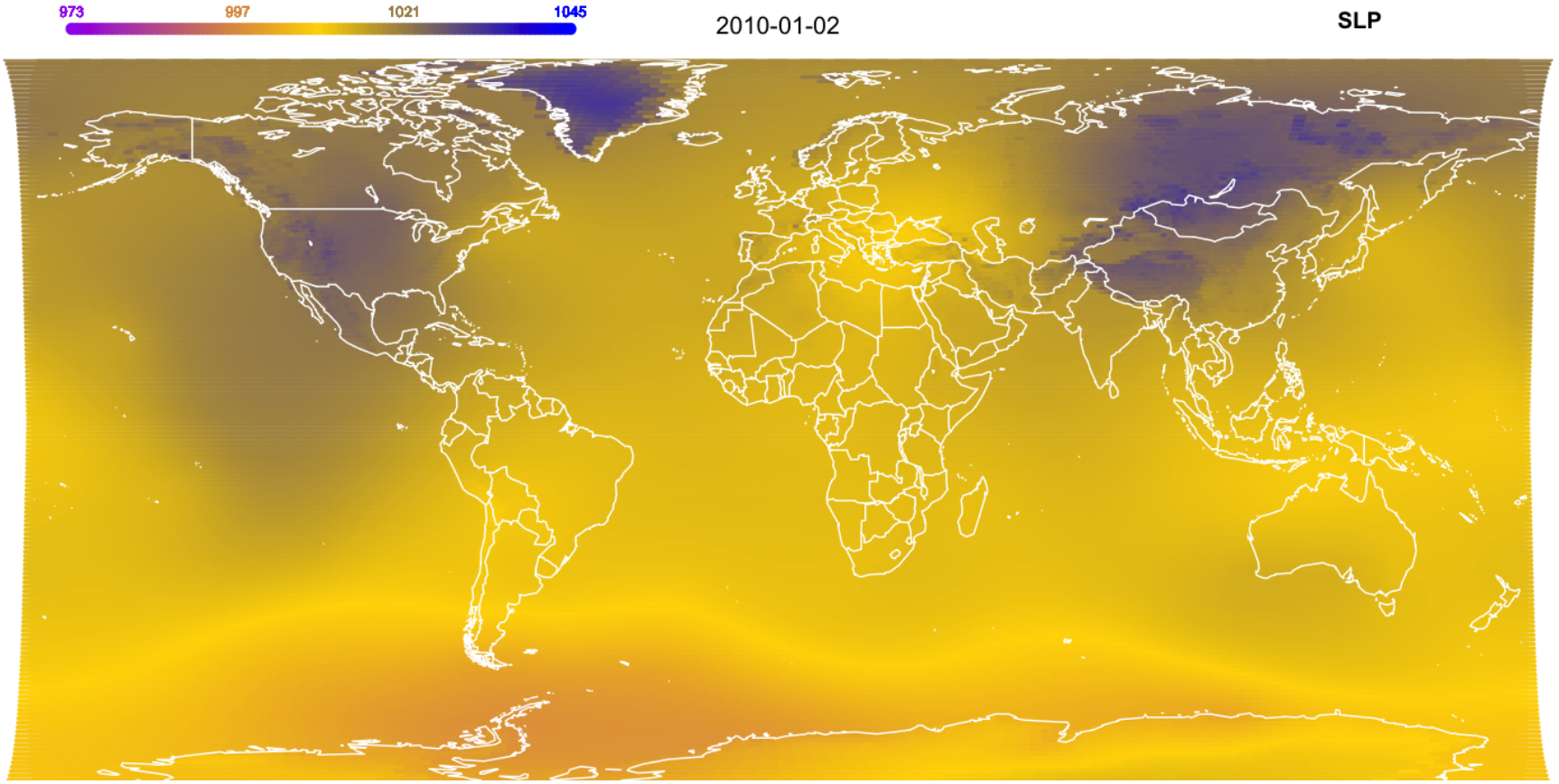

Figure 7.20: Snapshot of globally imputed SLP daily Temperature, using WIDALS.

For VISIB, Figures 7.21-7.23 show raster maps of our predicted values for the same three days.

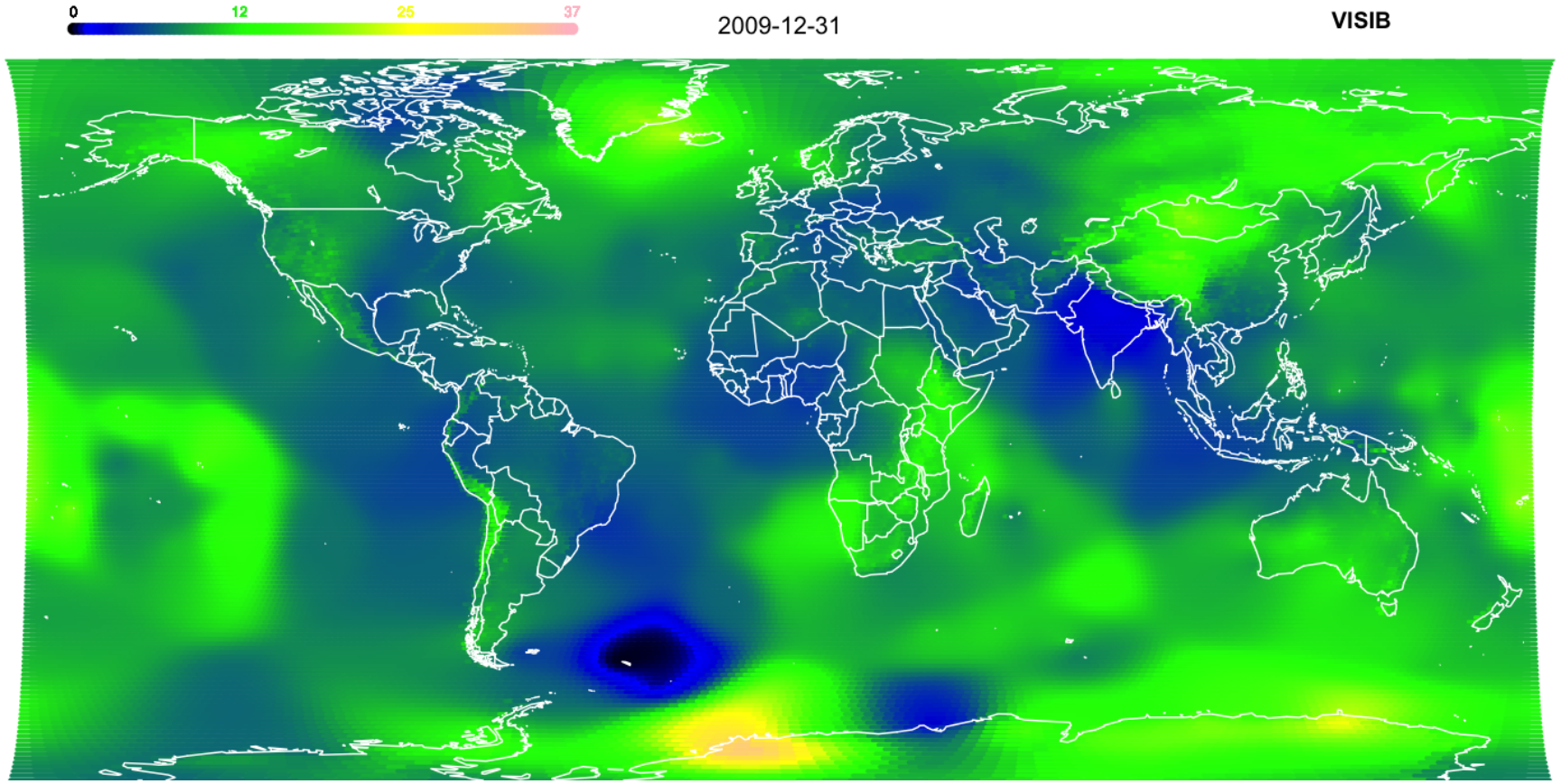

Figure 7.21: Snapshot of globally imputed VISIB daily Temperature, using WIDALS.

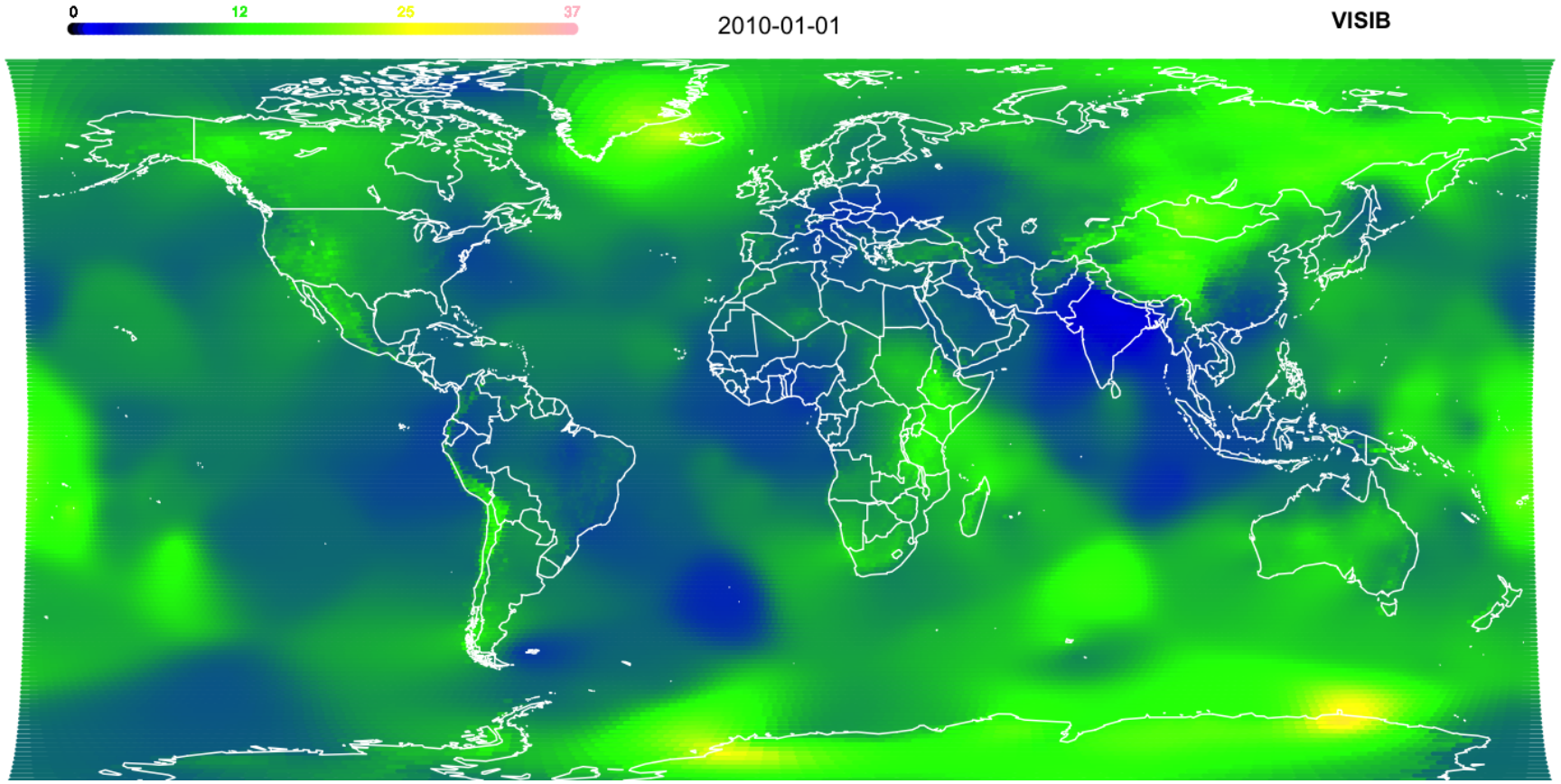

Figure 7.22: Snapshot of globally imputed VISIB daily Temperature, using WIDALS.

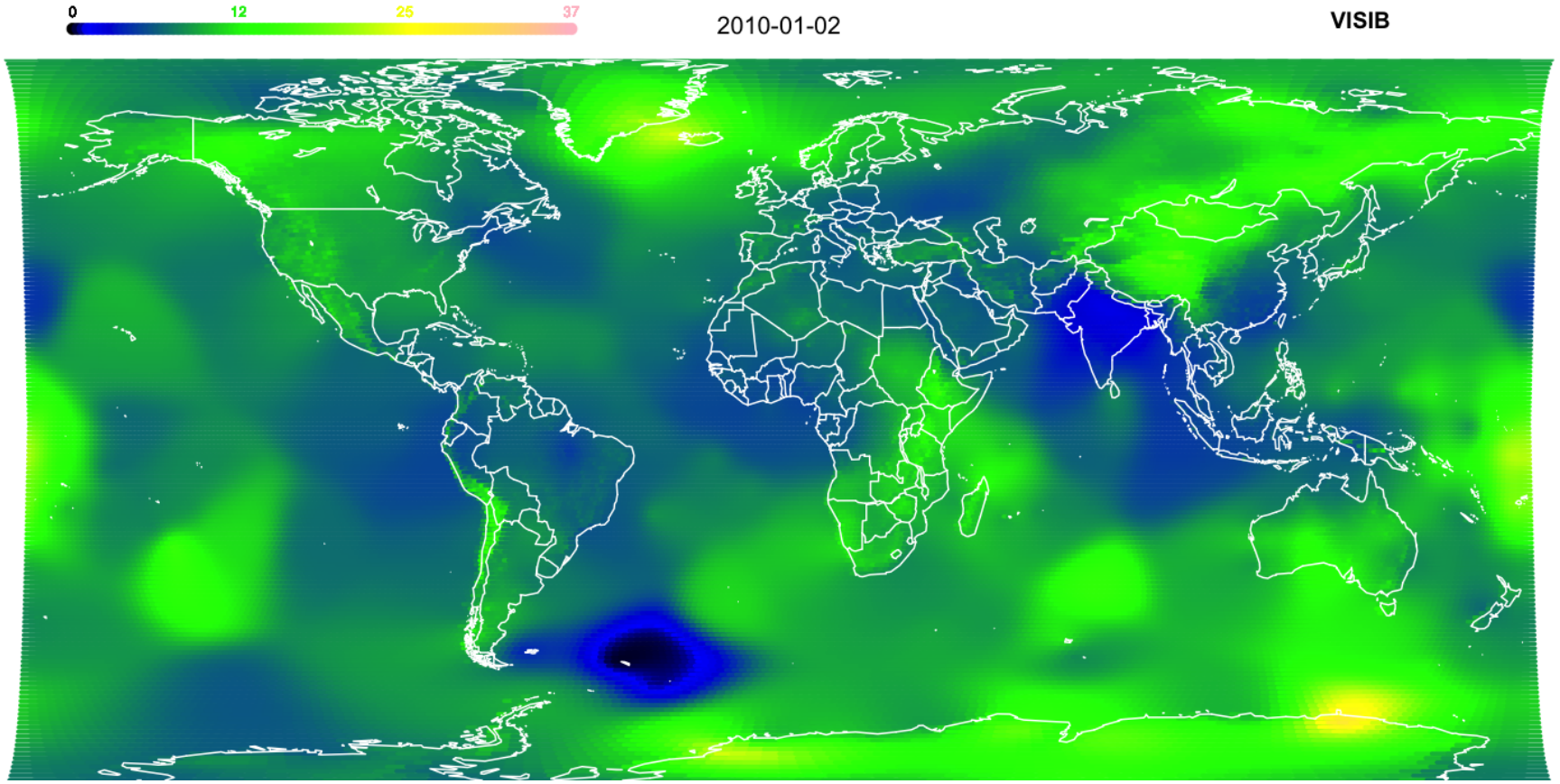

Figure 7.23: Snapshot of globally imputed VISIB daily Temperature, using WIDALS.

CHAPTER 8

Conclusions

It can be said that model hermeneutics spawned WIDALS. Multilevel state-space models are powerful means by which we may describe action in space-time. Application to large data presents two related problems. First, the researcher must specify a model form, then attempt to identify a potentially massive number of model hyperparameters from the data. Second, the researcher must use the specified model with the fitted hyperparameters to solve for predictions. Canonical models, alas, are computationally expensive; with sufficiently large data, they are impossible to execute in any practical way.

8.0.6 Wending the Logic

So let us move through the sequence that originally advocated WIDALS. The *latent state* level is virtually always in much lower dimension than the observation space. Moreover, with the regression paradigm in mind, the measurement matrix can be the familiar design matrix. ALS possesses the ability to estimate/predict the latent states utilizing only two scalar hyperparameters. In a sense, ALS is nonparametric; one can feed it any sort of variable, ordinal, dummy, and retrospectively acquire some sense of the quality of the latent state estimates using empiric techniques, such as resampling, cross-validation. Generic ALS, however, omits utilizing observation correlation structure in its estimation of the ALS states. Such a generalization of ALS (call it GALS for the sake of this discussion, so that LS:GLS :: ALS:GALS) may require the dreaded inversion of an $n \times n$ matrix.

However, the off-diagonal observation space correlations have only a trivial effect on the slope estimates (but, of course, can have a dramatic effect observation space prediction). If the diagonal elements vary (the system is heteroskedastic over space), then one can easily account for this without taxing the computational expense (analogous to weighted least squares). Finally, we had noticed time and time again when working with large spacial data sets, especially those with irregularly located sensors, that the Kriging solution (or the LS solution), produced weights whose relationship with distance — call this $w(d^\star)$ — looked strikingly exponential. In many cases, the departures were infinitesimal. Sometimes $w(d^\star)$ would oscillate, dipping below the abscissa. But when even a tiny nugget effect is added (e.g., *extrinsic* measurement error), or the number of supporting sites increased, $w(d^\star)$ flattens out to look again exponential.

If one lets one's mind wander far enough, one might postulate that nonparametric approaches are the statistician's *Critical Theory*. The grand positivist tradition in applied research is to acquire data, *specify* a generative model, estimate parameters, then expectorate confidence intervals and p-values. Model specification, though, is a convenience, and theoretical estimator properties become mere could-have-beens — or, to an obstinate researcher, should-have-beens. This convenience is demagogical.

WIDALS is extremely fast, easy to code, and generates output that is easily digested by any studious researcher. At the worst, WIDALS can be one of many benchmarks available to the researcher *en route* to visualizing and solving large space-time data. At the very best, WIDALS can expeditiously find a near *true* minimum for some sought objective function.

8.0.7 The Future

The reader will have noticed that the WIDALS spacial smoothing weights, $\mathbf{W}$, are, in their essence, application, and everything in between, nothing more than a spacial *transformation*. So too is the LS solution we called $w(d^\star)$ above. Both are merely a result of *folding* space. Such investigations belong to an enterprise so vast that mathematics has dedicated the entire branch of *topology* to it. For the present work, we made no formal excursions into topology for an obvious reason: It wasn't necessary. However, the at-times excellent performance of the constrained-sum exponential function we have employed here certainly serves as a tantalizing cue that other, computationally easy transformations lurk.

In a very early draft of the present work, we had included as applications a number of real-life videos, where WIDALS was called upon for forecasting *missing* frames, or otherwise restoring frames possessing various types of noise. My own personal attraction to this applications, was, admittedly, to at least in some small way follow in the footsteps of my advisor, Professor Ying-Nian Wu, and also Professor Zhowen Tu, a committee member, who have developed far more elegant, powerful and flexible means of image and video analyses. Indeed, human vision recognition makes much more intricate use of local spacial structure, as well as repetitive spacial structure, such as textures, for which WIDALS, at least in its present implementation, cannot meaningfully account. For example, regions of high variance, which can signal a physical edge (the border of a teapot), or an artistic edge (the border of a flower painted on the teapot), can be effectively modeled using particular classes of local transformations, notably, the Gabor function, or Canny edge detection. Of course, the current formulation of WIDALS permits inclusion of spacial transformation or convolution at either the mean functions level, or the smoothing level. Consider that the smoothing level acts as a de-noiser, inviting the potential inclusion of a third level where more detail image recognition filters may live — that is, WIDALS might serve as a fast

image preprocessor.

References

[ABS00] Felix Abramovich, Trevor C. Bailey, and Theofanis Sapatinas. "Wavelet Analysis and its Statistical Applications." *The Statistician*, **49**:1–29, 2000. Part 1.

[AM79] Brian D.O. Anderson and John B. Moore. *Optimal Filtering*. Dover, Mineola, NY, 1979. republished in 2005.

[Bar08] Christopher David Barr. *Applications of Voronoi tessellations in point pattern analysis*. PhD thesis, UCLA, Rick Paik Schoenberg (Committee Chair), 2008.

[BBN02] Stephen D. Billings, Rick K. Beatson, and Garry N. Newsam. "Interpolation of geophysical data using continuous global surfaces." *Geophysics*, **67**(6):1810–1822, 2002.

[BBR09] Mostafa Bendahmane, Raimund Bürger, and Ricardo Ruiz-Baier. "A Multiresolution Space-Time Adaptive Scheme for the Bidomain Model in Electrocardiology." *Wiley Online Library*, October 2009.

[Ber04] Richard Berk. *Regression Analysis: A Constructive Critique*. Sage, Thousand Oaks, CA, 2004.

[CJ08] Noel Cressie and Gardar Johannesson. "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 2008.

[CR05] Joaquin Quiñonero Candela and Carl Edward Rasmussen. "A Unifying View of Sparse Approximate Gaussian Process Regression." *Journal of Machine Learning Research*, **6**:1939–1959, 2005.

[Cre93] Noel A.C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., 1993.

[CSK10] Noel Cressie, Tao Shi, and Emily L. Kang. "Fixed Rank Filtering for Spatio-Temporal Data." *Journal of Computational and Graphical Statistics*, **19**(3):724–745, 2010.

[CW02] Noel Cressie and Christopher K. Wikle. "Space-time Kalman filter." *Encyclopedia of Environmetrics*, **4**:2045–2049, 2002.

[Efr99] Sam Efromovich. *Nonparametric Curve Estimation (Springer Series in Statistics)*, volume 1862. Springer U.S., New York, 1999.

[FGN06] Reinhard Furrer, Marc G. Genton, and Douglas Nychka. "Covariance Tapering for Interpolation of Large Spatial Datasets." *Journal of Computational and Graphical Statistics*, **15**:502–523, 2006.

[FS09] Reinhard Furrer and Stephan R. Sain. "Spatial Model Fitting for Large Datasets with Applications to Climate and Microarray Problems." *Statistics and Computing*, **19**(2):113–128, 2009.

[Gar99] Michael Garland. "Multiresolution Modeling: Survey and Future Opportunities." *Eurographics*, 1999.

[GDG10] Alan E. Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of Spatial Statistics*. CRC Press, Boca Raton, 2010.

[Goo11] Google. "Google Elevation API." https://developers.google.com/maps/documentation/elevation/, 2011. [Online; last accessed 2012-07-04].

[Gun94] S. Gunnarsson. "On Covariance Modification and Regularization in Recursive Least Squares Identification." *In 10th IFAC Symposium on System Identification — SYSID'94*, **2**:661–666, July 1994. Copenhagen, Denmark.

[Gun96] S. Gunnarsson. "Combining Tracking and Regularization in Recursive Least Squares Identification." *Proceedings of the 35th Conference on Decision and Control*, December 1996. Kobe, Japan.

[Hay02] Simon Haykin. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, forth edition, 2002.

[HC96] Hsin-Cheng Huang and Noel Cressie. "Spatio-temporal prediction of snow water equivalent using the Kalman filter." *Computational Statistics & Data Analysis*, **22**:159–175, 1996.

[HCG02] Hsin-Cheng Huang, Noel Cressie, and John Gabrosek. "Fast, Resolution-Consistent Spatial Prediction of Global Processes from Satellite Data." *Journal of Computational and Graphical Statistics*, **11**(1):63–88, March 2002.

[Hoc09] Ronald R. Hocking. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley, New York, 2009.

[JC02] G. Johannesson and N. Cressie. "Variance-Covariance Modeling and Estimation for Multi-Resolution Spatial Models." *Technical Report*, (700), December 2002.

[JC04] G. Johannesson and N. Cressie. "Finding large-scale spatial trends in massive, global, environmental datasets." *Environmetrics*, **15**:1–44, 2004.

[JCH07] Gardar Johannesson, Noel Cressie, and Hsin-Cheng Huang. "Dynamic multi-resolution spatial models." *Environmental and Ecological Statistics*, **14**:5–25, 2007.

[Jea99] Pierre Delfiner Jean-Paul Chilès. *Geostatistics: modeling spatial uncertainty.* John Wiley & Sons, New York, NY, 1999.

[KJ99] Phaedon C. Kyriakidis and André G. Journel. "Geostatistical Space-Time Models: A Review." *Mathematical Geology*, **31**, 1999.

[Lju98] Lennart Ljung. *System Identification: Theory for the User (2nd Edition).* Prentice Hall PTR, 1998.

[LSM07] Giovanna Jona Lasino, Sujit K. Sahu, and Kanti V. Mardia. *Bayesian Statistics and its Applications*, chapter Modeling rainfall data using a Bayesian Kriged-Kalman model, pp. 301–318. Anamaya Publishers, New Delhi, India, 2007.

[LYS99] Chi Sing Leung, G.H. Young, J. Sum, and W.K. Kan. "On the Regularization of Forgetting Recursive Least Square." *IEEE Transactions on Neural Networks*, **10**(6), November 1999.

[McC05] J. Huston McCulloch. "The Kalman Foundations of Adaptive Least Squares, With Application to U.S. Inflation." *Unpublished*, 2005.

[MW01] R. Van der Merwe and E.A. Wan. "The Square-Root Unscented Kalman Filter for State and Parameter-Estimation." In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 6, pp. 3461 –3464, 2001.

[NHS12] Douglas Nychka, Dorit Hammerling, Stephen Sain, and Tia Lerud. *LatticeKrig: Multiresolution Kriging based on Markov random fields*, 2012. R package version 2.2.1.

[RDS96] Gabriel Rodriguez-Yam, Richard A. Davis, and Louis L. Scharf. "Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression." 1996. Technical report, Colorado State University, Fort Collins.

[Say03] Ali H. Sayed. *Fundamentals of Adaptive Filetering.* John Wiley & Sons, Hoboken, N.J., 2003.

[Sch11] Frederic Paik Schoenberg. "Tessellations." *UCLA Department of Statistics Papers*, 8 2011.

[SH89] Michael L. Stein and Mark S. Handcock. "Some Asymptotic Properties of Kriging When the Covariance Function Is Misspecified." *Mathematical Geology*, **21**(2), 1989.

[She68] Donald Shepard. "A two-dimensional interpolation function for irregularly-spaced data." *Proceedings – 1968 ACM National Conference*, pp. 517–524, 1968.

[SM05a] Sujit K. Sahu and Kanti V. Mardia. "A Bayesian kriged Kalman model for short-term forecasting of air pollution levels." *Applied Statistics*, **54**(1):223–244, 2005.

[SM05b] Sujit K. Sahu and Kanti V. Mardia. "Recent trends in modeling spatio-temporal data." *Proceedings of the special meeting on Statistics and Environment, Messina*, 2005.

[SS06] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications, With R Examples*. Springer, N.Y., second edition, 2006.

[Ste02] Michael L. Stein. "The Screening Effect in Kriging." *The Annals of Statistics*, **30**(1):298–323, 2002.

[Was06] Larry Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[WH97] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, second edition, 1997.

[Zes12a] Dave Zes. *SSsimple: Solve, Fit, Simulate State Space Systems*, 2012. R package version 0.6.0.

[Zes12b] Dave Zes. *widals: Fit, forecast, predict massive spacio-temporal data*, 2012. R package version 0.5.