2017-09

A hierarchical multivariate Bayesian approach to ensemble model output statistics in atmospheric prediction

Wendt, Robert D. T.

Monterey, California: Naval Postgraduate School

http://hdl.handle.net/10945/56188

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

A HIERARCHICAL MULTIVARIATE BAYESIAN APPROACH TO ENSEMBLE MODEL OUTPUT STATISTICS IN ATMOSPHERIC PREDICTION

by

Robert D. T. Wendt

September 2017

Dissertation Supervisor: Wendell A. Nuss

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704–0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2017	**3. REPORT TYPE AND DATES COVERED** Dissertation

4. TITLE AND SUBTITLE
A HIERARCHICAL MULTIVARIATE BAYESIAN APPROACH TO ENSEMBLE MODEL OUTPUT STATISTICS IN ATMOSPHERIC PREDICTION

5. FUNDING NUMBERS

6. AUTHOR(S) Robert D. T. Wendt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

Previous research in statistical post-processing has found systematic deficiencies in deterministic forecast guidance. As a result, ensemble forecasts of sensible weather variables often manifest biased central tendencies and anomalous dispersion. In this way, the numerical weather prediction community has largely focused on upgrades to upstream model components to improve forecast performance—that is, innovations in data assimilation, governing dynamics, numerical techniques, and various parameterizations of subgrid-scale processes. However, this dissertation explores the efficacy of statistical post-processing methods downstream of these dynamical model components with a hierarchical multivariate Bayesian approach to ensemble model output statistics. This technique directly parameterizes meteorological phenomena with probability distributions that describe the intrinsic structure of observable data. Bayesian posterior beliefs in model parameter were conditioned on previous observations and dynamical predictors available outside of the parent ensemble. An adaptive variant of the random-walk Metropolis algorithm was used to complete the inference scheme with block-wise multiparameter updates. This produced calibrated multivariate posterior predictive distributions (PPD) for 24-hour forecasts of diurnal extrema in surface temperature and wind speed. These Bayesian PPDs reliably characterized forecast uncertainty and outperformed the parent ensemble and a classical least-squares approach to multivariate multiple linear regression using both measures-oriented and distributions-oriented scoring rules.

14. SUBJECT TERMS
ensemble model output statistics, statistical post-processing, multivariate multiple linear regression, Bayesian data analysis, Bayesian hierarchical modeling, Markov chain Monte Carlo methods, Metropolis algorithm, machine learning, atmospheric prediction

15. NUMBER OF PAGES
221

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

A HIERARCHICAL MULTIVARIATE BAYESIAN APPROACH TO ENSEMBLE MODEL OUTPUT STATISTICS IN ATMOSPHERIC PREDICTION

Robert D. T. Wendt
Lieutenant Commander, United States Navy
B.S., University of Illinois at Urbana-Champaign, 2003
M.S., University of Illinois at Urbana-Champaign, 2005
M.S., Naval Postgraduate School, 2014

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN METEOROLOGY

from the

NAVAL POSTGRADUATE SCHOOL
September 2017

Approved by:

Dr. Wendell Nuss
Professor of Meteorology
Dissertation Supervisor

Dr. James Hansen
Superintendent, Marine Meteorology
Naval Research Lab

Dr. Patrick Harr
Section Head, AGS
National Science Foundation

Dr. Eric Hendricks
Assoc. Professor of Meteorology
Naval Postgraduate School

Dr. Eva Regnier
Assoc. Professor of Business
Naval Postgraduate School

Dr. Qing Wang
Professor of Meteorology
Naval Postgraduate School

Approved by: Wendell A. Nuss, Chair, Department of Meteorology

Approved by: Douglas Moses, Vice Provost of Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Previous research in statistical post-processing has found systematic deficiencies in deterministic forecast guidance. As a result, ensemble forecasts of sensible weather variables often manifest biased central tendencies and anomalous dispersion. In this way, the numerical weather prediction community has largely focused on upgrades to upstream model components to improve forecast performance—that is, innovations in data assimilation, governing dynamics, numerical techniques, and various parameterizations of subgrid-scale processes. However, this dissertation explores the efficacy of statistical post-processing methods downstream of these dynamical model components with a hierarchical multivariate Bayesian approach to ensemble model output statistics. This technique directly parameterizes meteorological phenomena with probability distributions that describe the intrinsic structure of observable data. Bayesian posterior beliefs in model parameter were conditioned on previous observations and dynamical predictors available outside of the parent ensemble. An adaptive variant of the random-walk Metropolis algorithm was used to complete the inference scheme with block-wise multiparameter updates. This produced calibrated multivariate posterior predictive distributions (PPD) for 24-hour forecasts of diurnal extrema in surface temperature and wind speed. These Bayesian PPDs reliably characterized forecast uncertainty and outperformed the parent ensemble and a classical least-squares approach to multivariate multiple linear regression using both measures-oriented and distributions-oriented scoring rules.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ARW	Advanced Research WRF
BEMOS	Bayesian ensemble model output statistics
BPF	Bayesian processor of forecasts
EMOS	ensemble model output statistics
EPS	ensemble prediction system
IID	independent and identically distributed
JPM	joint probability model
KDE	kernel density estimation
MAE	mean absolute error
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
MOS	model output statistics
MSE	mean squared error
MVN	multivariate normal distribution
MVT	multivariate t-distribution
NMMB	Nonhydrostatic Multiscale Model on B-Grid
NR	non-homogenous regression
NRL	Naval Research Laboratory
NCEP	National Centers for Environmental Prediction
NWP	numerical weather prediction
NWS	National Weather Service
OLS	ordinary least squares
PDF	probability density function
PPD	posterior predictive distribution
SSR	sum of squared residuals
USN	United States Navy
WFO	weather forecast office
WRF	Weather Research and Forecasting model

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

This research is dedicated to the memory of my grandfather, Dr. Dale Edwards, for whom I have tremendous love, respect, and gratitude. His gentle encouragement made so many aspects of my life and career possible. I wish I could have shared this accomplishment with you, Grandpa.

I would like to thank my family for their patience, love, and support during this arduous period of study and research. Emily, Cordelia, Oliver, Eleanor, and Alfred—I love you. To my mother, Cheryl, and my grandmother, Carol: please know how grateful I am for all of your love and support throughout the years. You are simply wonderful.

I would like to express profound and sincere gratitude to my doctoral advisor, Professor Wendell Nuss, for his patience and support. His calm and insightful mentoring made this research possible. I can’t thank you enough for all you have done for me in this endeavor. I would also like to thank the members of my dissertation committee for their guidance and patience. Wendell, Jim, Pat, Eric, Eva, and Qing—thank you.

I would also like to thank my friends, Dr. Jonathan Jarvis and Jonathan Russell, for their unique contributions to my research. I am similarly grateful to Dr. Rob Swanson for helping me to “live fully now.”

Finally, I would like to thank the Naval Postgraduate School community, including faculty, staff, and my fellow students, both past and present, for their fellowship and support.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. RELEVANCE

The *Art of War* emphasizes the value of information in warfare, especially in the context of operational planning, and identifies the “seasons”—or, more broadly, the spatial and temporal variability of the environment—as one of five fundamental factors that affect military engagements (Sun-tzu and Griffith 1964). While intuition might suggest that routine forecasts of sensible weather variables are too prosaic for earnest operational consideration, military history provides an extensive record of events—including Napoleon’s invasion of Russian in 1812, or, more recently, the Normandy landings in 1944—in which the environment, climatology, or weather was crucial to the outcome (e.g., Winters et al. 1998). As a result, forecasts of the ocean, atmosphere, and space domains are routinely consumed by contemporary decision-makers to optimize the distribution of military resources and, indeed, the lethality of weapons systems over short and long timescales. In this way, environmental prediction affects decision superiority at the tactical, operational, and strategic levels of warfare. As described in the *U.S. Navy Information Dominance Roadmap: 2013–2028*, it directly contributes to battlespace awareness—”know the enemy, know the environment” (Leigher 2013).

To this end, numerical weather prediction (NWP) has served as the foundation of operational forecasting since the pioneering work of Rear Admiral Grace Hopper, Jules Charney, and Edward Lorenz in the mid-twentieth century. While the sophistication of various NWP schemes has grown considerably since the field’s inception, especially with the emergence of ensemble prediction systems (EPS), the perennial search for better forecast methodologies is as compelling today as when the first operational forecast was produced by the National Centers for Environmental Prediction (NCEP) in 1950 (Kalnay 2003). To this end, this dissertation explores and develops a specific class of machine-learning techniques in modern data science—that is, Bayesian parameter estimation with Markov chain Monte Carlo (MCMC) sampling methods—as a compelling new approach to statistical post-processing in atmospheric prediction.

B. BACKGROUND

Prediction lies at the heart of the scientific method. Often relying on detailed mathematical descriptions of nature, it provides a conceptual bridge between hypotheses and the empirical evidence that manifests the physical structure of the universe. To this end, determinism has served as an intuitive approach to prediction that has shaped the analytical topography of mathematical physics, and the specialized disciplines that compose its progeny, since Laplace first rigorously engaged the philosophical qualities of probability theory in 1814 (Hawking 1999). In his view, the stochastic margins between observations and predictions are created by a lack of understanding; that is, the natural world may be precisely known, both past and future, with perfect information (Pearl 2000). Ignorance of various components of a dynamical system—including the initial configuration, relevant boundary conditions, or, indeed, the equations that govern its evolution—will obscure its phase space evolution and preclude the accurate identification of future states. Eponymously described as Laplace's demon, this intuitive distillation of natural laws promises arbitrary prognostic skill if sufficient information is obtained.

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. (Laplace 1814)

In this way, scientific determinism ostensibly produces meaningful statements about physical processes for which precise, repeatable predictions are sought. Examples include the periodic motion of astronomical objects orbiting a parent mass or the adiabatic ascent of an idealized parcel of dry air lifted in the atmospheric boundary layer. At time and length scales that are intuitive for human observation, these principles are essentially self-apparent. If one presumes a sufficient understanding of the physical laws that relate observable variables, and indeed their initial values, then counterfactual statements—that is, mathematical descriptions of outcomes or events that have not occurred—are available from first principles (Pearl 2000). Until quantum mechanical notions of uncertainty were formalized in 1926, wherein a precise knowledge of a particle’s position and momentum is reinterpreted within the stochastic framework of the wave function and Schrödinger equation, determinism served as the implicit basis for orthodox scientific thinking—especially in classical mechanics (Hawking 1999). Einstein’s infamous claim “God does not play dice with the universe” is appropriately interpreted in this context; he preferred a traditional view of causality (i.e., deterministic versus stochastic generative processes) in which the perceived stochastic elements of any variable were the products of ignorance—vis-à-vis Laplace’s canonical view of physical systems—and not, as quantum mechanics suggests, an intrinsic facet of nature (Natarajan 2008).

With a similar faith in deterministic principles, and nearly two decades before the disruptive insights of quantum mechanics, Abbe (1901) and Bjerknes (1904) recognized the fundamental utility of scientific determinism and pursued its rigorous application in meteorology (Lynch 2007). They considered fluid dynamics and thermodynamics to be ideal analytical tools by which the real atmosphere, and its complex superposition of physical processes, might be mathematically decomposed and numerically rendered. Bjerknes’ template for weather prediction would become particularly influential; it established a diagnostic step to estimate the initial state of the atmosphere and, perhaps more importantly, a prognostic step to engage the time-dependence of the governing dynamics through the integration of appropriate differential equations (Kalnay 2003). To facilitate the prognostic component of his forecast algorithm, Bjerknes combined the ideal gas law with material conservation statements for density, water vapor, momentum, and internal energy to codify the primitive equations—a set of ubiquitous physical relationships that describe the covariation of fundamental atmospheric scalars.

While theoretically viable, Bjerknes’ diagnostic-prognostic approach was constrained in two primary respects. As described by Kalnay (2003), Charney (1951) correctly recognized that the spatial and temporal distribution of meteorological observations was poorly suited to initial value problems—that is, ordinary differential equations supplied with the appropriate initial conditions. Even if the inadequacy of the contemporary observing network was neglected, however, the analytical difficulties of solving a coupled set of non-linear differential equations were unresolved at that time. Richardson (1922) would address the latter with an application of finite differences to approximate Bjerknes' primitive equations, but his initial trials were nevertheless corrupted by observational deficiencies that produced physically implausible forecasts (Lynch 2007). Meaningful efforts to advance the dynamical approach introduced by Bjerknes and Richardson were eventually successful, perhaps most notably with the numerical integration of the barotropic vorticity equation by Charney et al. (1950) on the Electronic Numerical Integrator and Computer (ENIAC) (Lynch 2007).

Nevertheless, the implicit role of causality in the simulation of physical systems became increasingly relevant to the NWP community. While Charney and his ENIAC collaborators had indeed demonstrated the numerical feasibility of a filtered dynamical predictive scheme, Lorenz (1962) explored the comparative efficacy of contemporary statistical (i.e., linear regression) and dynamical forecast methodologies for sensible weather variables. Lorenz (1963) would go on to reveal a fundamental paradox regarding the sensitivity of unstable dynamical systems, which includes the real atmosphere, to perturbations in their initial state: given enough time, even the most trivial of residuals was capable of producing divergent phase space trajectories that were mathematically indistinguishable from a stochastic process (e.g., Figure 1). Similar to the insights of quantum mechanics, and disrupting intuitive notions of causality consistent with Laplace's demon, Lorenz had established a fundamental basis for chaos theory in meteorology—findings that explicitly invoked randomness as a fundamental component of meteorological prediction.

The implications of the butterfly effect in NWP are difficult to overstate. Kalnay (2003) observes that "even with perfect models and perfect observations the chaotic nature of the atmosphere would impose a finite limit of about two weeks to the predictability of the weather." Lorenz himself would later posit the following definition: "chaos—when the present determines the future, but the approximate present does not approximately determine the future" (Danforth 2013). In this way, chaos theory created a powerful motivation for additional research in computational and statistical methodologies that could better characterize latent stochastic modes of atmospheric dynamics. To this end, Epstein (1969) introduced a statistical approach with stochastic modifications to the governing dynamical equations, which he described as stochastic-dynamic equations, to incorporate the analysis uncertainty via probability distributions (Kalnay 2003). Moreover, he recognized the value of quantitative estimates of forecast uncertainty and noted that if "the probabilistic nature of the initialization is unavoidable, then so also is the probabilistic nature of the prediction" (Epstein 1969).

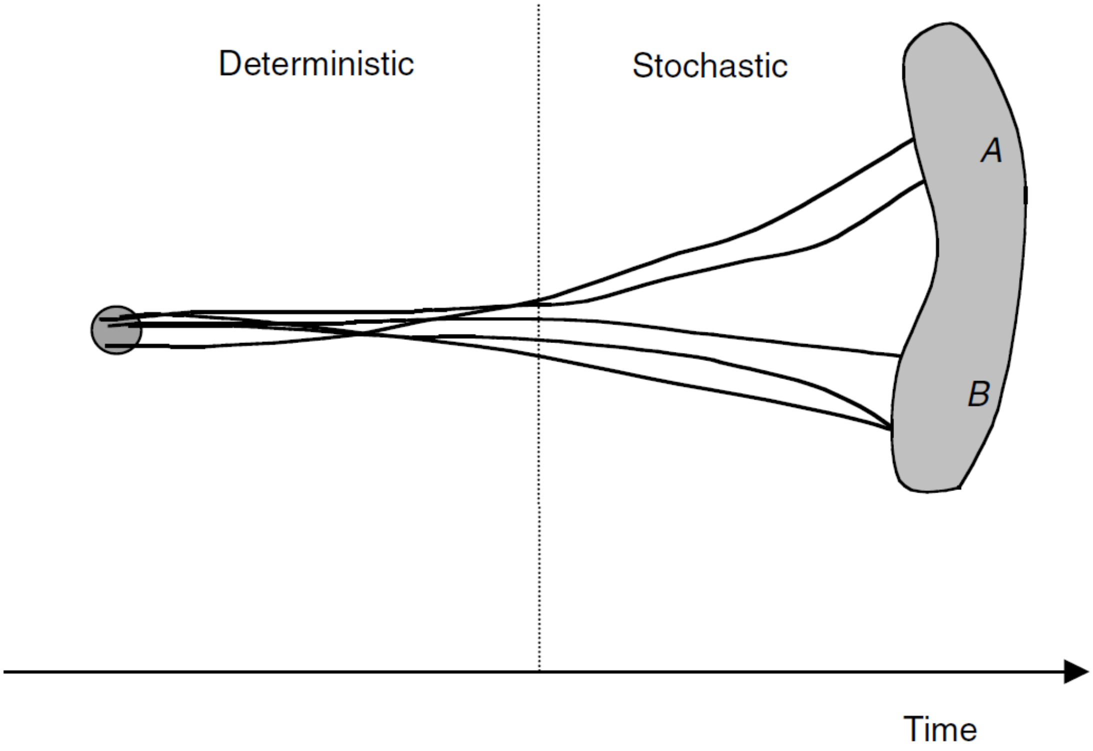


Figure 1. Time evolution of an unstable dynamical system via ensemble modeling. Source: Kalnay (2003). Uncertainty in the estimation of the initial state is represented by small-amplitude perturbations inside the grey circle (left). The sensitivity of the forecast to ensemble model deficiencies is indicated by the divergence of the black forecast curves in the stochastic region (right). The prognostic uncertainty at an arbitrary forecast time is represented by the dispersion of the discrete forecast estimates (black curves).

The computational intractability of Epstein’s complete stochastic-dynamic forecast algorithm was later considered by Leith (1974)—especially for models that contained a large number of degrees of freedom. Consequently, he suggested a more efficient application of available computational methods, wherein a sufficiently large set of deterministic solutions—each initialized from small-amplitude perturbations in the initial state—could be numerically sampled and statistically compared. While only a modest extension of Epstein (1969), his intuitive Monte Carlo forecasting approach would nevertheless become the foundation for modern ensemble forecasting (Kalnay 2003). Leith (1974) was also notable for exploring various optimal regression correction techniques to minimize mean-squared error (MSE) in finite ensemble schemes—an early precursor of modern statistical post-processing. In particular, he observed:

> The averaging together of individual numerical forecasts in a Monte Carlo forecasting procedure has the effect of filtering out small scale structure at forecast times for which there is little accuracy left. A further linear regression step has been shown to provide an optimal estimate of the forecast state of the atmosphere which, of course, tends to preserve the more predictable large-scale anomalies. At late times all skill is lost, forecast anomalies tend to zero, and the forecast consists of the climatological mean field. (Leith 1974)

From this perspective, the noisy details of individual NWP model flows become less relevant than a time series of descriptive statistics summarizing their central tendency and spread—especially at long forecast times. Irrespective of the likelihood of their origins, as indicated by probability distributions evaluated over the perturbed initial conditions, the ensemble members compose a filtered signal of prognostic consensus—the ensemble mean—that demonstrates enhanced time-averaged forecast skill when compared with a control solution (Figure 2a; Kalnay 2003). The dispersion of the ensemble members at a specified time similarly provides a *prima facia* estimate of the forecast uncertainty (Figures 1 and 2) that ostensibly marginalizes out deficiencies in the observations, model dynamics, and finite-difference approximations (Richter 2012). Moreover, the time-evolution of the ensemble’s dispersion may communicate the intrinsic variability of forecast uncertainty. Figure 2 demonstrates this point: the uncertainty (i.e., dispersion) suggested by the ensemble on the left (Figure 2a) increases monotonically with time, while the uncertainty communicated by the ensemble on the right (Figure 2b) is smaller and essentially steady state—despite poorer accuracy from the ensemble mean.

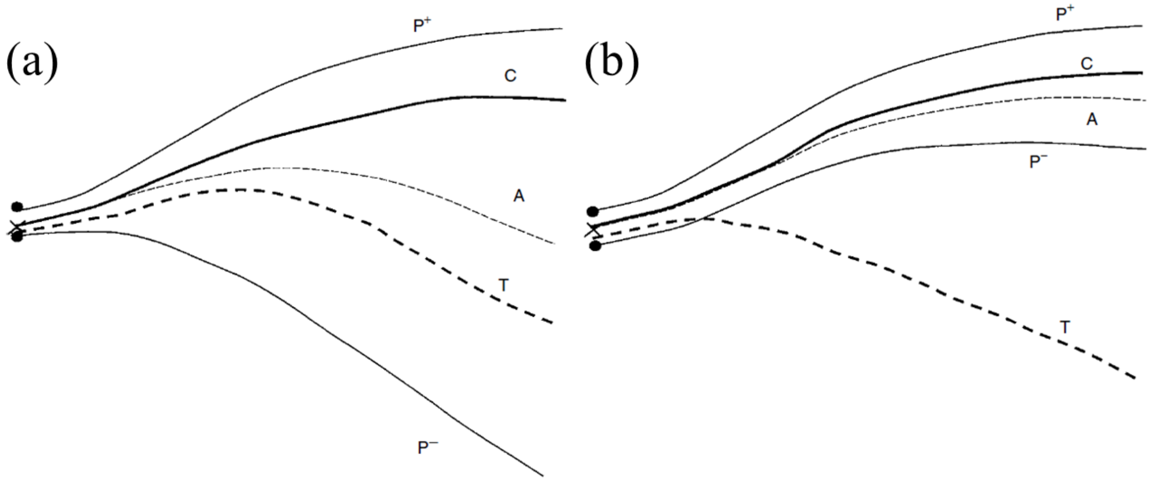


Figure 2. Examples of ensemble forecasts with standard model components. Source: Kalnay (2003). Components include positive perturbation(s) P+, negative perturbation(s) P-, a dynamical control C, an ensemble mean A, and the observed truth T. The distribution on the left (right) is considered desirable (undesirable) because truth was observed close to (far from) the ensemble mean solution and inside (outside) of the range of ensemble members during the forecast period.

In this way, the conceptual foundation of the ensemble approach is appropriately viewed as a Monte Carlo simulation—a form numerical integration over perturbations in the initial conditions, governing dynamics, and numerics—that heuristically samples probable atmospheric states to extract uncertainty information from deterministic models. The resulting ensemble members are themselves discrete deterministic estimates, but they compose a sparsely-populated predictive distribution—the ensemble—that measures the sensitivity of the forecast to various stochastic modeling elements vis-à-vis chaos theory. As a result, it is frequently assumed that the probabilistic forecasts produced by an EPS are credible estimates of sensible weather variables; at the very least, they provide some quantitative measure of forecast uncertainty—information unavailable from single-valued, deterministic NWP schemes.

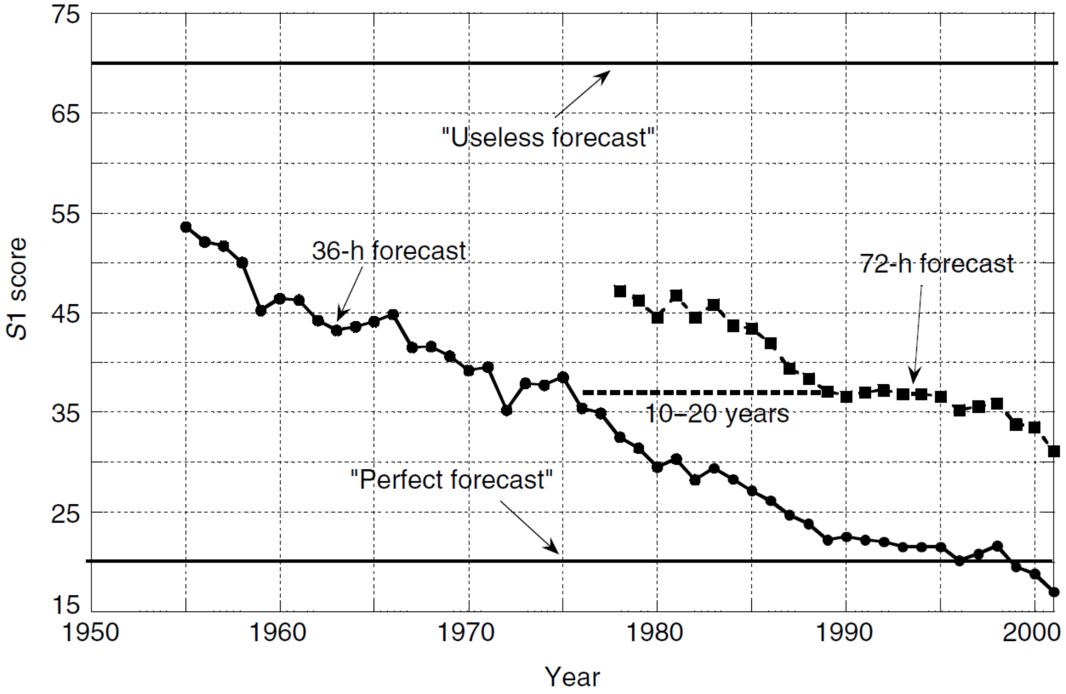


Figure 3. Historical comparison of operational NCEP model performance. Source: Kalnay (2003). S1 scores correspond to 36-hour and 72-hour forecasts of area-averaged 500 hPa horizontal pressure gradient. Advances in NWP methodologies provide predictive performance at 72 hours that is equivalent to 36-hour performance two decades earlier.

Initial investigations into optimal methods of perturbing ensemble model components—especially the initial conditions—eventually produced the first operational EPS in December 1992 at NCEP and the European Centre for Medium-Range Weather Forecasts (ECMWF) (Kalnay 2003). Combined with reliable advances in computational power consistent with Moore’s law (Moore 1965), the computational sophistication and analytical complexity of modern numerical approaches to atmospheric prediction notably increased. As a result, quasi-monotonic improvements in model skill have been observed since the early contributions of Charney et al. in 1950 (e.g., Figure 3; Kalnay 2003).

Nevertheless, Lorenz’s predictability barrier remains a fundamental limitation in meteorology—even with the benefits of contemporary ensemble forecasting. The NWP community has largely focused on upgrades to *upstream* model components—that is, improvements in data assimilation, the governing dynamics, numerical techniques, and various parameterizations of subgrid-scale processes—to extract additional predictive performance from contemporary objective guidance. While these efforts have proven effective, as described by Kalnay (2003) in the 36-hour to 72-hour skill conversion described for NCEP operational S1 scores (Figure 3), the search for latent model skill remains an active area of research. This is especially true at long forecast times, where Leith (1974) observes that dynamical predictions—even within an EPS—compare unfavorably to estimates derived from suitable climatology.

However, comparatively little attention has been given to rigorous statistical analyses of forecast data *downstream* of the aforementioned model components. This last point is ironic considering the state of NWP in mid-twentieth century; Kalnay (2003) describes it as a period of “dark years” defined by a bifurcation in modeling approaches—that is, between statistical and dynamical forecasting methodologies—from which the statistical approach nearly emerged as the victor (Kalnay 2003). While Gneiting et al. (2005) acknowledges the efficacy of dynamical models and the ensemble approach to forecasting, both in the research literature and operational forecasting, it is somewhat curious that the benefits of enhanced computational power, which partially fueled the success of modern dynamical NWP, were not applied with greater enthusiasm to contemporary statistical approaches as well. The fundamental stochastic features of atmospheric prediction articulated by Lorenz’s butterfly effect, and the ensemble approach developed by Epstein and Leith in response, have an intuitive connection with the language and tools of statistical inference.

In particular, Robert and Casella (2011) observe that a MCMC “revolution,” which itself was nurtured by improvements in computational power during the 1980s, led to an important resurgence of complex Bayesian models within the statistical community (e.g., Geman and Geman 1984, Gelfand and Smith 1990). In the decades that preceded this technological revolution, classical notions of statistical inference—that is, the canonical method of least squares, or ordinary least squares (OLS), pioneered by Gauss, Legendre, and Laplace and, later, the technique of maximum likelihood estimation (MLE) introduced by Fisher (1922)—dominated Bayesian approaches to parameter estimation (Feigelson 2015). However, MCMC sampling methods have revitalized Bayesian estimation and made important contributions in machine learning, deep learning, and artificial intelligence—all exceptionally active areas of contemporary research. In this sense, this dissertation reunites the dynamical and statistical approaches to NWP—not to replace the former, but to help reanimate the latter with some of the latest tools in contemporary data science and statistics.

C. MOTIVATION

Gneiting et al. (2005) observes that earlier studies—including Eckel and Walters (1998), Stensrud and Yussouf (2003), and Scherrer et al. (2004)—reveal that "ensemble variance and related measures of ensemble spread are skillful indicators of the accuracy of the ensemble mean forecast." While this validates the basic elements of an intuitive EPS spread-skill relationship, the body of contemporary statistical post-processing research, which includes the seminal work of Gneiting et al. (2005) and Raftery et al. (2005), in addition to Möller et al. (2013), Richter (2012), Veenhuis (2013), Williams et al. (2014), Hodyss et al. (2016), and Baran and Möller (2017), suggests that ensemble central tendency and spread are also frequently biased and overconfident. The underdispersive tendencies of ensemble output are especially noteworthy, as much of the research pursued heretofore has focused on the statistical bias correction of single-valued forecasts—especially the influential regression method of model output statistics (MOS) introduced by Glahn and Lowry (1972) (Thorarinsdottir and Gneiting 2010). Unfortunately, MOS produces a point estimate of truth that cannot directly calibrate anomalous EPS dispersion.

In this way, a fundamental motivation for the ensemble approach can be diminished. The discrete forecast estimates produced by an EPS, which ostensibly provide an accurate characterization of the dynamical system as it evolves in phase space, are frequently unrepresentative of the sensitivity of the forecast to deficient information encoded in the dynamical model. To illustrate this point, let the ensemble model flows in Figure 2 represent surface temperatures at an arbitrary model grid point. If the scenario depicted in Figure 2b is common, with the true state notably displaced from the central tendency of the ensemble members, then the EPS likely contains a warm or cold bias. If the dispersion of the ensemble members is similarly insufficient for the climatological variability of the predicted variable, so that a set of observations could not be considered independent and identically distributed (IID) samples from a set of corresponding ensemble distributions, then the EPS could be classified as uncalibrated. As described by Gneiting (2014), Gneiting et al. (2007) developed the relevant definition from Murphy and Winkler (1987):

> We propose a diagnostic approach to the evaluation of predictive performance that is based on the paradigm of maximizing the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the observed values. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. The more concentrated the predictive distributions are, the sharper the forecasts, and the sharper the better, subject to calibration. (Gneiting et al. 2007)

A well-calibrated EPS will therefore engage both elements of ensemble post-processing—that is, correcting biased central tendency and dispersion—since it is not possible for “the observations [to] be reasonably interpreted as drawn from the predictive distributions” if their mean is biased or their spread is anomalous (Gneiting 2014). To verify this claim, consider the predictive distributions depicted in Figure 4 for maximum diurnal surface wind speed at an arbitrary model grid point. The raw ensemble forecasts, which are assembled according to Gneiting et al. (2005) by fitting a Gaussian distribution to the underlying ensemble mean and variance, are sharper and have smaller means than the Bayesian posterior predictive distribution (PPD) and climatological reference distribution. If one were to consider the dashed red vertical line as truth for this pedagogical example, then it would be difficult to consider this observation as randomly drawn from either of the raw ensemble distributions—the probability density is simply too low. Climatology and the Bayesian PPD assign non-trivial probability densities to truth, but the dispersion of each could prove to be inappropriate for a set of observations if, for example, they were tightly clustered (i.e., small variance) around the current value.

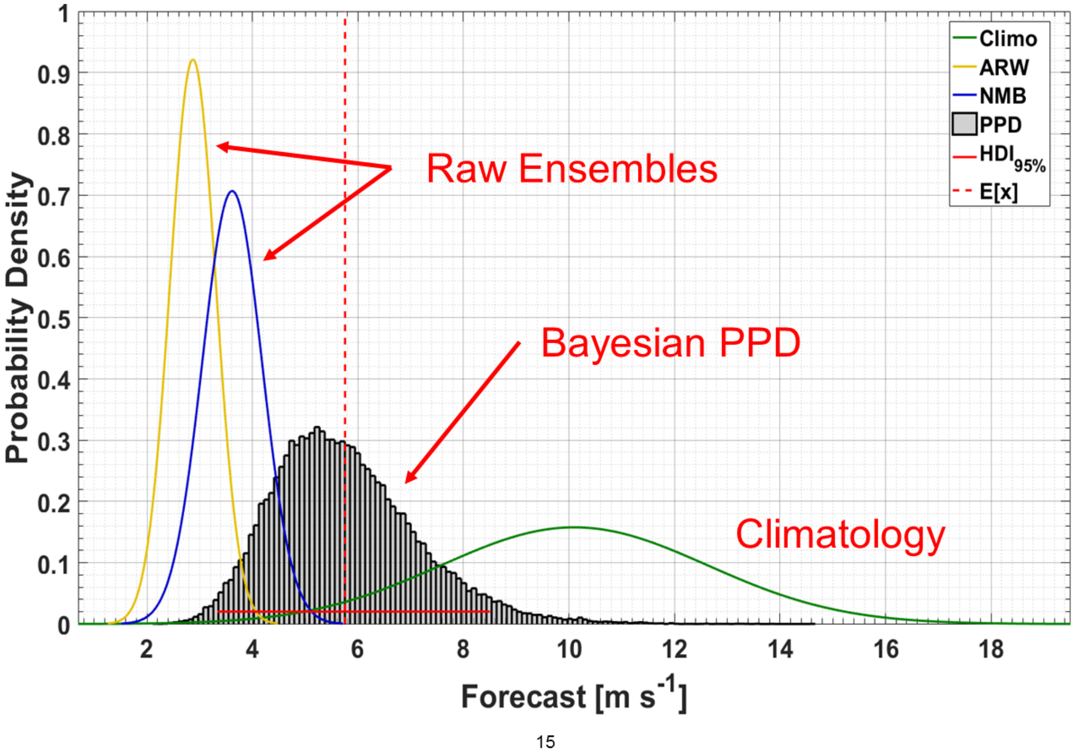


Figure 4. Predictive distributions for 24-hour forecasts of diurnal maximum surface wind speed. The raw ensemble output (yellow and blue) is sharper (i.e., more concentrated) than Bayesian (histogram) and climatological (green) estimates. A highest density interval (HDI; solid red) describes the narrowest portion of the PPD containing a specified probabilistic mass fraction (e.g., 95% as shown).

As this comparison is repeated over many forecast trials, one can properly assess the calibration of the predictive distributions via rank histograms and probability integral transforms (PIT) vis-à-vis Dawid (1984), Diebold et al. (1998), and Hamill (2001) (Gneiting et al. 2007). In particular, it would be possible for a predictive distribution to have sufficient dispersion to describe the variability of future observations but have a central tendency that is systematically high or low. Conversely, the predictive distribution may demonstrate no meaningful bias in the mean yet present a spread that is over- or underdispersive relative to observed surface wind speeds. In this way, a predictive distribution must have a relatively unbiased central tendency and dispersion for one to reasonably classify the observations as IID samples from the corresponding probabilistic forecasts. Statistical post-processing should therefore seek calibrated distributional estimates that are more accurate, vis-à-vis the central tendency, and sufficiently dispersive to represent the fundamental uncertainty of the underlying prediction.

Various statistical post-processing methodologies have been introduced to identify and correct systematic patterns of error in previous forecasts. Hodyss et al. (2016) provides a thorough summary of methods commonly encountered in the research literature today, including forecast analogues (e.g., Hamill et al. 2015) and various forms of kernel density estimation (KDE); the latter includes ensemble dressing (e.g., Wang and Bishop 2005), ensemble regression (Unger et al. 2009), and Bayesian model averaging (BMA; e.g., Raftery et al. 2005 and Vrugt et al. 2008). In this way, KDE forms the basis for a broader class of post-processing techniques and provides an intuitive scheme for transforming a sample of discrete predictions (i.e., kernels) into a continuous forecast distribution (e.g., Figure 5). The latter point is critical for techniques that cannot directly produce probabilistic forecasts for continuous variables; it also provides an explanation for the popularity of KDE—especially for sparsely-populated predictive distributions (e.g., most ensembles). As a non-parametric approach to density estimation, KDE makes no explicit assumptions about the data or the physical processes that influence their generation—save for the shape of the selected kernel function and the size of the bandwidth (i.e., smoothing) parameter. Juban et al. (2007) note the following:

> There are two main categories of density estimation methods: parametric and non-parametric. In the parametric framework, a distribution family is chosen, e.g., the Gaussian distribution. Then, the parameters of the distribution are estimated from the available data. In the non-parametric framework the distribution is directly estimated from the data based on a weaker hypothesis on the underlying distribution. The main drawback of the non-parametric approach is that it requires larger data sets than the parametric one to attain equivalent estimations. The main advantage is that it limits estimation errors due to incorrect hypotheses on the underlying distribution family. (Juban et al. 2007)

Although various weighting techniques may be employed to manage the influence of the smoothed data (e.g., the BMA approach introduced by Raftery et al. 2005), KDE methods nevertheless lack a framework for estimating more complex model structures (e.g., covariance among predicted variables, correlated predictor variables, or multi-level clustering patterns in the data). By design, they also obviate the need for parameter estimation and the distributional hypotheses described by Juban et al. 2007. By comparison, a parametric modeler is able to choose which continuous probability distribution(s) might provide a good fit to the underlying shape of the data (e.g., the log-normal distribution fit to positively-skewed wind speeds in Figure 4). Using the terminology of Krzysztofowicz and Evans (2008), knowledge of the physical relationships between atmospheric variables (e.g., Bjerknes' primitive equations) may also inform the modeler's selection of independent predictor variables, based on their assumed covariation with the desired predicted quantities (i.e., the predictands), and, indeed, the analytical structure between these latent variables (e.g., multiple linear vs. non-linear regression models).

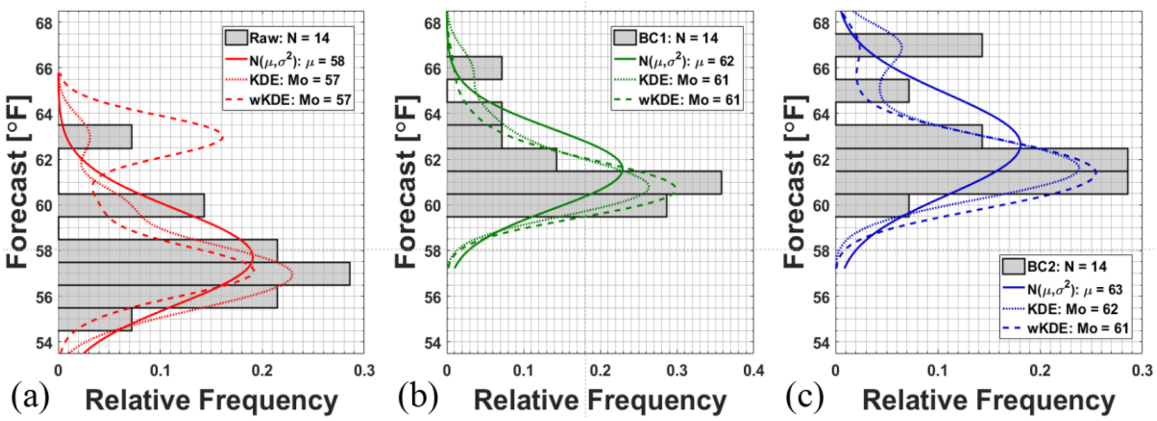


Figure 5. Meteorological application for KDE. Raw forecast estimates (a) are compared with post-processed estimates (b and c) derived from linear regression analyses. Bias corrections derived from mean error statistics are applied as an offset (b) or coefficient (c) to raw forecast guidance. Solid curves represent Gaussian distributions fit to the indicated forecasts based on the sample mean and variance. Dotted (dashed) curves represent unweighted (skill-weighted) KDE schemes.

Hodyss et al. (2016) examines the KDE/BMA technique introduced by Raftery et al. (2005) in greater detail and notes several deficiencies; these include under-dispersive forecasts at long lead times (Wilks 2006), overfitting (Hamill 2007), and extreme forecasts (Bishop and Shanley 2008). Hodyss et al. (2016) also identifies an issue with weighted, regression-corrected forecasts in BMA: if the order of these adjustments is applied incorrectly, an over-weighting of climatology is produced. To this end, a direct application of Bayesian methods is suggested as an optimal alternative that minimizes error variances in post-processed forecasts (Hodyss et al. 2016). The latter point serves as a primary motivation for the research in this dissertation, for which additional theoretical development is reserved for Chapter II. In this way, previous studies have demonstrated the diversity of KDE and BMA techniques and, indeed, some of the disadvantages of the non-parametric approaches to statistical post-processing.

The advanced method of ensemble model output statistics (EMOS), which is also sometimes described as non-homogenous regression (NR), is a parametric approach to statistical post-processing that fits a single continuous distribution to a set of training data (Gneiting 2014). Introduced by Gneiting et al. (2005) to offer an "easy-to-implement post-processing technique that addresses both forecast bias and underdispersion and takes into account the spread-skill relationship," it further parameterizes the location and scale of a continuous distribution—often Gaussian—with a linear combination of ensemble predictors (Gneiting et al. 2005). Once the regression parameters have been estimated by the method of minimum Continuous Ranked Probability Score (CRPS), contemporary predictors are inserted into the multiple linear regression framework to estimate new distributional parameters (e.g., the location and scale of a Gaussian distribution) appropriate for current forecast conditions. In this way, a linear combination of raw ensemble predictions and regression parameters is transformed into a probabilistic forecast over the desired sensible weather variable (e.g., Figure 6). Forecast distributions produced in this manner will have the same parametric form as the original distribution fit to the data. Gneiting et al. (2005) notably found that NR/EMOS predictive distributions were sharp and well-calibrated relative to raw ensemble output (e.g., Figure

7) and bias-corrected peers. The analytical details of this approach will be developed in Chapter II.

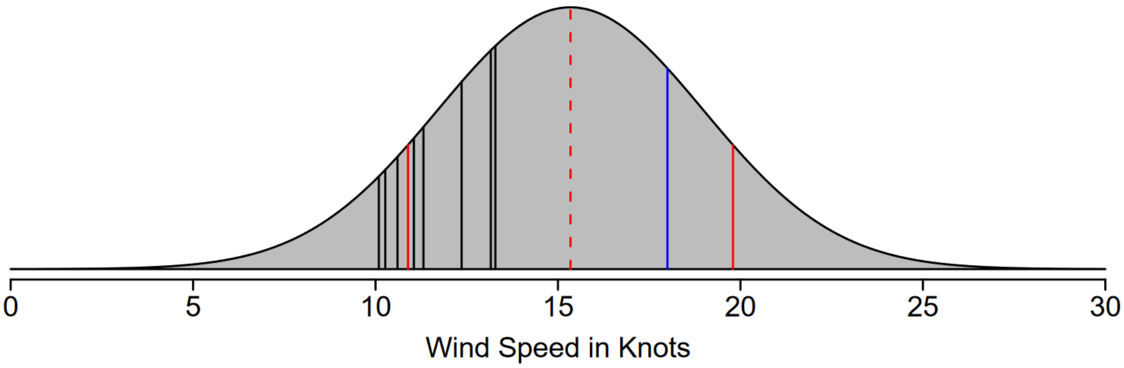


Figure 6. Example of a probabilistic forecast for maximum surface wind speeds produced by the NR/EMOS approach. Source: Thorarinsdottir and Gneiting (2010). Raw ensemble output is indicated by the dashed vertical lines; the median of the post-processed forecast distribution is indicated by the center-most solid red line; the boundaries of the 77.8% central prediction interval are outboard in solid red; truth is indicated in blue. The forecast is for The Dalles, Oregon, valid 14 JUN 2003, with a 48-hour lead time.

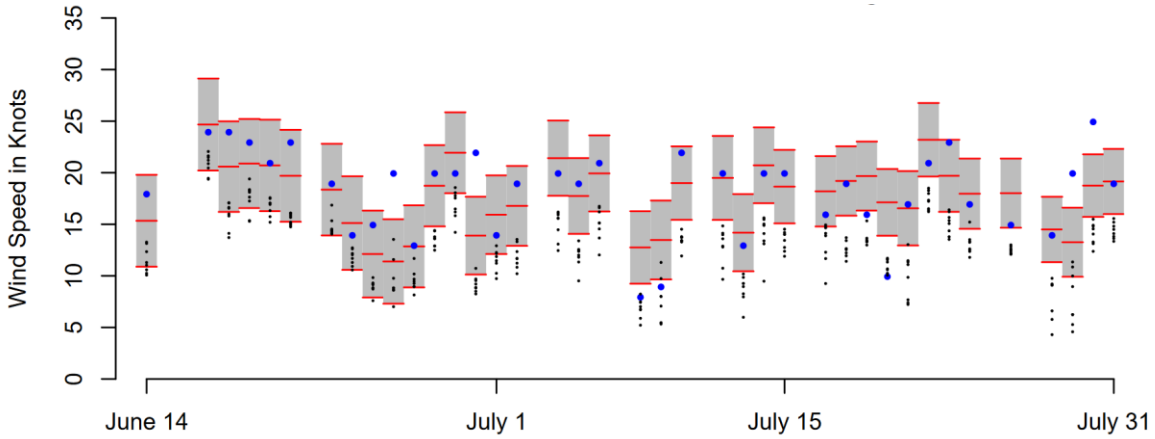


Figure 7. Time series of probabilistic forecasts. Source: Thorarinsdottir and Gneiting (2010). Observed maximum surface wind speeds (blue) are shown relative to 77.8% central prediction intervals derived from the NR/EMOS post-processing method. Black dots indicate the raw output of the parent ensemble. The statistical coverage of the predictive intervals suggests that the generative model is well-calibrated so that the observations can be reasonably considered IID samples from the corresponding forecast distributions vis-à-vis Gneiting et al. (2007). The forecasts are for The Dalles, Oregon, valid 14 JUN 2003 through 31 JUL 2003, with 48-hour lead times.

Schuhen et al. (2012) later extended the NR/EMOS work of Thorarinsdottir and Johnson (2011) to two-dimensional wind vector predictions and found that a bivariate EMOS model produced calibrated predictive distributions with favorable performance margins relative to raw ensemble output and other forms of statistical post-processing. Baran and Möller (2017) extended the approach to predictions of wind speed and temperature and found performance comparable to bivariate BMA; however, their bivariate EMOS model was more computationally efficient. Inspired by the hierarchical approach to BMA suggested by Narzo and Cocchi (2010), Richter (2012) introduced the first Bayesian extension of EMOS—which the author describes as Bayesian ensemble model output statistics (BEMOS)—to consider the benefits of an alternate parameter estimation scheme. Bayesian inference inverts the canonical probability statement to treat the data as fixed and the model parameters as random variables (Casella 2008). This results in a "direct quantification of uncertainty" in parameter estimation that seeks full posterior probability distributions for model parameters conditioned on *a priori* model beliefs and available training data (Gelman et al. 2013).

Benefiting from the MCMC revolution described by Robert and Casella (2011), elements of the broader statistical community observed that Bayesian estimation had several fundamental advantages over comparable classical statistical methods. In his development of multiple linear regression techniques for Bayesian data analysis, which have vital relevance to this dissertation, Kruschke (2014) notes that

> one of the benefits of Bayesian analysis is that correlations of credible parameter values are explicit in the posterior distribution. Traditional statistical methods provide only a single "best" (e.g., MLE) parameter value, without indicating the trade-offs among parameter values. The Bayesian posterior, however, naturally reveals trade-offs and redundancies among parameters. It is up to the user to actually look for and interpret the correlations of parameters, of course. Another benefit of Bayesian analysis is that the model doesn't "explode" when predictors are correlated. If predictors are correlated, the joint uncertainty in the regression coefficients is evident in the posterior, but the model happily generates a posterior regardless of correlations in the predictors. The classical, one-best-solution method is much less robust in the presence of strongly correlated predictors. (Kruschke 2014)

The classical methods of OLS regression and MLE produce single-valued parameter estimates that optimize their associated cost functions—regression residuals and the likelihood function, respectively—without explicit regard for the uncertainty of the underlying inference. To be sure, both methods can be paired with traditional confidence intervals (e.g., the ubiquitous 95% confidence intervals commonly found in hypothesis testing for statistical significance); however, the uncertainty information communicated by Bayesian and non-Bayesian inferences is not equivalent. Casella (2008) posits that classical methods of parameter estimation assume the "data are repeatable random samples [but] the underlying parameters remain constant during this repeatable process;" the Bayesian approach proceeds as if "the data are fixed [and] parameters are unknown and described probabilistically." In this way, classical confidence intervals make uncertainty statements regarding the frequency with which future confidence intervals will include the fixed, true value of the unknown parameter—not the probability that the single, indicated interval contains this true value. By comparison, Bayesian intervals for posterior parameter belief, which are often described as credible intervals, characterize the probability that the true parameter value lies in the indicated interval given the fixed data sample. Gelman et al. (2013) describes the impact of this distinction:

> A primary motivation for Bayesian thinking is that it facilitates a common-sense interpretation of statistical conclusions. For instance, a Bayesian (probability) interval for an unknown quantity of interest can be directly regarding as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice. Recently in applied statistics, increased emphasis has been placed on interval estimation rather than hypothesis testing, and this provides a strong impetus to the Bayesian viewpoint, since it seems likely that most users of standard confidence intervals give them a common-sense Bayesian interpretation. (Gelman et al. 2013)

As a result, Bayesian estimation provides intuitive uncertainty information that is consistent with the practical interpretations of common, lay users (e.g., most meteorologists). Gelman et al. (2013) adds that “the essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.” Casella (2008) provides a relevant pedagogical comparison in Figure 8 for Aminoglycoside-Associated Nephrotoxicity (AAN) incidence rates related to various methods of drug action in a meta-analysis of pharmacological studies by Kim et al. (2004). While similar, the Bayesian and frequentist intervals have non-trivial distinctions in central tendencies and spread; they also contain disparate uncertainty information. Bayesian (blue) intervals indicate a 90% probability of finding the true rate of AAN inside of the indicated interval given current studies; frequentist (black) intervals suggest that 90% of similar intervals constructed from other studies will contain the true AAN rate. It should be noted that the frequentist intervals lack probability statements about true AAN incidence rates—that is, the quantity that motivated the statistical analysis.

In this way, the character of the uncertainty information communicated by an EPS becomes relevant to the Bayesian extension of NR/EMOS by Richter (2012).

Meteorologists interpret the spread of an ensemble as generalized forecast uncertainty; that is, greater (less) spread indicates more (less) model sensitivity to defects in the initial conditions, governing dynamics, physics parameterizations, or numerics. Kalnay (2003) observes that this spread communicates the “reliability of the forecast: if the ensemble forecasts are quite different from each other, it is clear that at least some of them are wrong, whereas if there is good agreement among the forecasts, there is more reason to be confident about the forecast.” Perhaps more importantly, the ensemble method “provide[s] a quantitative basis for probabilistic forecasting” that is consistent with the Monte Carlo methods invoked by Leith (1974) (Kalnay 2003). Finally, Kalnay (2003) notes that “[t]he quantitative relationship between the ensemble spread and the forecast error (or conversely, between the forecast agreement and the forecast skill) has yet to be firmly established, but is now routinely taken into consideration by human forecasters.” In simple terms, meteorologists frequently interpret ensemble dispersion as forecast uncertainty and use the distribution of ensemble members to assemble *ad hoc* probabilistic forecasts estimates.

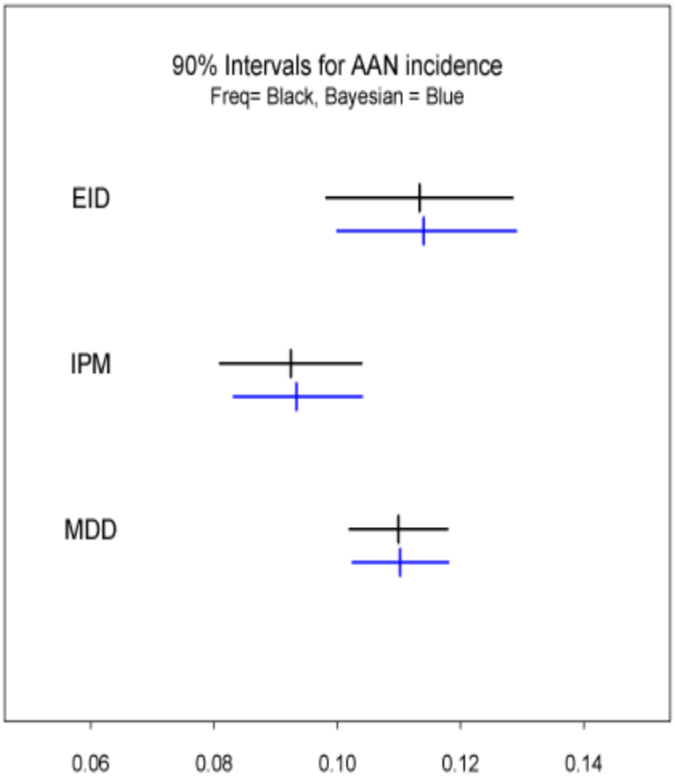
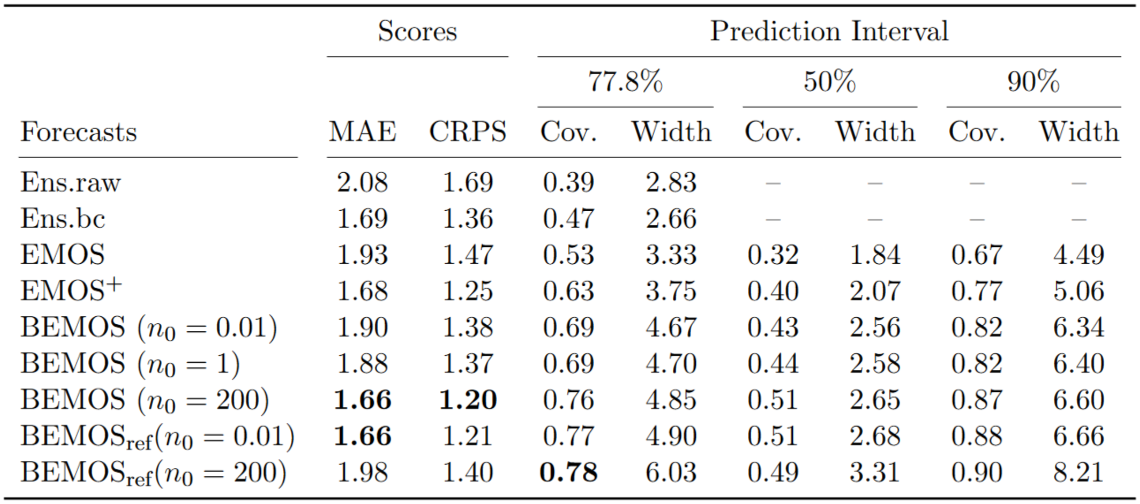


Figure 8. Statistical inferences associated with a meta-analysis of pharmacological studies by Kim et al. (2004). Source: Casella 2008. The relative frequency of Aminoglycoside-Associated Nephrotoxicity (AAN) incidence is examined relative to extended-interval dosing (EID), individualized pharmacokinetic monitoring (IPM), and multiple-daily dosing (MDD). Interval estimates for AAN rates are indicated for classical frequentist (black) and Bayesian (blue) methods (from).

When these insights are combined with the understanding that ensemble forecasts are often biased and underdispersive, it becomes apparent that the probabilistic forecast conclusions meteorologists derive are often unreliable and inconsistent with their assumptions. The NR/EMOS approach introduced by Gneiting et al. (2005) is notable for engaging both aspects of ensemble post-processing and for providing a simple method of sampling from calibrated, parametric forecast distributions. However, it lacks the utility of Bayesian parameter estimation which provides the statistical modeler with tremendous flexibility to choose sophisticated joint probability models that explicitly communicate the uncertainty of the underlying inference (Gelman et al. 2013). Moreover, non-

Bayesian methods of post-processing invoke frequentist uncertainty estimates that do not communicate common-sense probability statements consistent with the Bayesian descriptions of Casella (2008), Gelman et al. (2013), and Kruschke (2014). Bayesian uncertainty estimates are delivered in parallel with the primary parameter estimation; there is no need for additional computation. They also make intuitive probability statements that explicitly propagate coupled regression parameter uncertainty forward to the final predictive distributions vis-à-vis Kruschke (2014).



	Scores		Prediction Interval					
			77.8%		50%		90%	
Forecasts	MAE	CRPS	Cov.	Width	Cov.	Width	Cov.	Width
Ens.raw	2.08	1.69	0.39	2.83	–	–	–	–
Ens.bc	1.69	1.36	0.47	2.66	–	–	–	–
EMOS	1.93	1.47	0.53	3.33	0.32	1.84	0.67	4.49
EMOS$^+$	1.68	1.25	0.63	3.75	0.40	2.07	0.77	5.06
BEMOS ($n_0 = 0.01$)	1.90	1.38	0.69	4.67	0.43	2.56	0.82	6.34
BEMOS ($n_0 = 1$)	1.88	1.37	0.69	4.70	0.44	2.58	0.82	6.40
BEMOS ($n_0 = 200$)	**1.66**	**1.20**	0.76	4.85	0.51	2.65	0.87	6.60
BEMOS$_{\text{ref}}$($n_0 = 0.01$)	**1.66**	1.21	0.77	4.90	0.51	2.68	0.88	6.66
BEMOS$_{\text{ref}}$($n_0 = 200$)	1.98	1.40	**0.78**	6.03	0.49	3.31	0.90	8.21

Table 1. Performance comparison of competing approaches to NR/EMOS. Source: Richter (2012). Lower values of mean absolute error (MAE) and CRPS indicate better forecast performance; both metrics have units that match the predictand. Width describes the size of the indicated central prediction interval that includes the specified probabilistic mass; coverage describes the statistical consistency between the indicated forecast probabilities and the frequency with which truth was observed within these intervals. The tuning parameter controls the dispersion of the Bayesian prior distributions considered by Richter (2012); small values of the tuning parameter indicate a flatter, more uninformative prior while larger values would signal more confidence in specific values of model parameters.

In this way, the BEMOS approach introduced by Richter (2012) is consistent with the recommendations of Hodyss et al. (2016) and forms the fundamental basis for this dissertation. It also found “significantly better calibration” than the original EMOS approach when both were applied to the 8-member University of Washington Mesoscale Ensemble in 48-hour surface temperature forecasts for locations in the Pacific Northwest (Richter 2012; Table 1). However, the present work extends the BEMOS technique of Richter (2012) in five important ways: it permits full multivariate predictions consistent with Schuhen et al. (2012) and Baran and Möller (2015); it uses hierarchical parameter structure, which treats the hyperparameters of Bayesian prior distributions as random variables with their own hyperprior distributions, to permit hyperparameter inference from the data; it uses an adaptive multivariate form of the Metropolis algorithm, with block-wise component updates, to permit more sophisticated, non-conjugate probability models (e.g., the aforementioned hierarchical structure); it can fit the t location-scale distribution (i.e., a non-standard form of the classic Student’s t-distribution) to data; and, finally, it uses logarithmic data transformations to produce asymmetric predictive distributions for sensible weather variables with intrinsic skew (e.g., surface wind speeds). These items are consistent the recommendations of Gneiting (2014) for “cutting edge areas that are in critical need of research and development;” this includes “more sophisticated methods for the statistical post-processing of combined events and, closely related, for the development of theoretically principled techniques for the evaluation of probabilistic forecasts of multivariate quantities and events” (Gneiting 2014).

For comparison, both Gneiting et al. (2005) and Richter (2012) were limited to univariate predictions with Gaussian likelihood functions and symmetric predictive distributions. Moreover, Richter (2012) relied on a more restrictive form of MCMC sampling—that is, Gibbs sampling—that requires analytical knowledge of Bayesian conditional posterior distribution; the random-walk Metropolis sampler adapted for the present work requires no *a priori* knowledge of the analytical structure of the Bayesian posterior distribution. Richter (2012) also introduced a manual tuning parameter (i.e., n_o in Table 1) to control the confidence of Bayesian prior information over model parameters; their conclusions recommended an “automatic way of determining [this tuning] parameter,” and the hierarchical structure of the model developed for this dissertation addresses this point directly. While Gneiting et al. (2005) is also notable for including ensemble variance as a model parameter vis-à-vis the spread-skill relationship, research exploring the intuitive relationship between EPS skill and spread (e.g., Hopson 2014) suggests that limitations may exist. Similar to Richter (2012), the present work does not express model (co)variance as a linear combination of EPS dispersion statistics and, therefore, does not parameterize the ensemble spread-skill relationship.

Finally, it is helpful to consider the motivations for this research in the context of the fundamental challenges facing any modern forecaster—that is, the attractive perspective of Laplace's Demon and the difficulty of estimating sensible weather variables with incomplete and imperfect data vis-à-vis the Butterfly Effect. *Every* forecast contains uncertainty. Since multiple estimates of a sensible weather variable are frequently available, the forecast process is essentially, and quite appropriately, an application of statistical inference. Even when human forecasters leverage their professional experience, which is often gained through long periods of academic study and careful comparisons of skilled objective guidance with the corresponding observations, their contributions represent an *ad hoc* "nudge" toward outcomes they believe to be more likely. From this perspective, virtually all subjective guidance can be viewed as a form of statistical post-processing; in one form or another, data are evaluated (e.g., the ensemble mean), patterns are recognized, and raw objective guidance is adjusted to create the official forecast. No contemporary human forecaster is able to discretize the governing equations and mentally integrate them forward in time to produce discrete, numerical estimates for the maximum temperature in Monterey, CA tomorrow. Moreover, meteorologists are unlikely to consider identical sets of predictors and, even if they do, they often arrive at distinct—and sometimes incompatible—forecast conclusions. In this way, they are intrinsically contaminated with human error and any number of biases (e.g., sampling, confirmation, etc.). As a result, human forecasters are better positioned to monitor model performance, add prognostic detail in the mesoscale domain, and communicate forecast uncertainty. They also embed dynamical knowledge in the models themselves, and this is arguably their most important contribution. However, statistical post-processing techniques enable us to formalize these human, *ad* *hoc* “nudges” on larger sets of data and within a more rigorous analytical framework—that is, to “automate” the subjective qualities of forecasting with more objective methods.

D. OBJECTIVES

As an extension of the NR/EMOS methods that comprise Gneiting et al. (2005), Schuhen et al. (2012), Baran and Möller (2015), and Richter (2012), in addition to more generalized Bayesian approaches considered by Raftery et al. (2005), Rajagopalan et al. (2002), Luo et al. (2007), Krzysztofowicz and Evans (2008), and Veenhuis (2013), this dissertation evaluates a specific class of machine learning algorithms in modern data science—that is, Bayesian parameter estimation with MCMC sampling methods—as a compelling approach to statistical post-processing in operational forecasting. In this way, the present work extends the direct Bayesian schemes advocated by Richter (2012) and Hodyss et al. (2016) and, indeed, the MCMC methods explored by Vrugt et al. (2008) and Richter (2012); Hodyss et al. (2016) itself represents a development of the Bayesian processor of forecasts method (BPF) introduced by Krzysztofowicz and Evans (2008). In particular, this research evaluated a hierarchical multivariate Bayesian approach to multiple linear regression as a desirable technique for producing sharp, calibrated predictive distributions vis-à-vis Gneiting et al. (2007). These probabilistic forecasts should accurately characterize the forecast uncertainty of the desired sensible weather variables. The performance the of the resulting Bayesian PPDs are evaluated by measures-oriented and distributions-oriented scoring rules when the training period, ensemble predictor variables, and parametric form of the joint probability model (JPM) are adjusted. Bayesian PPD performance has been compared with traditional regression estimates produced by OLS and MLE. While the analytical and computational details of this scheme will be reserved for subsequent chapters, this Bayesian estimation/MCMC model framework has been selected to address the issues identified in the previous section. These include

1. Calibrating ensemble guidance to simultaneously correct forecast deficiencies associated with biased central tendencies and anomalous dispersion

2. Mitigating deficiencies observed with non-parametric density estimation (e.g., KDE/BMA)

3. Using Bayesian interval estimation to report forecast uncertainty with explicit probability statements that are consistent with the common-sense interpretations of most consumers

The present work exploits the robust sampling capabilities of MCMC methods, including a special case of the ubiquitous Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953 and Hastings 1970), to diagnose the forecast "habits" of parent models within a Bayesian framework. The Bayesian approach stores posterior beliefs in probability distributions over unobservable, latent model parameters. This results in a "direct quantification of uncertainty" from *a priori* model assumptions conditioned on available training data (Gelman et al. 2013). However, additional manipulations are required to transform posterior distributions over unobservable model parameters into forecast distributions over *unobserved* observable random variables. To this end, Bayesian PPDs extend the inference framework to synthetic samples of observable predictands conditioned on current guidance, past model performance, and subjective model assumptions. Interval estimates of the predictands are similarly delivered by credible intervals (regions) that describe the probability mass contained within a specified interval (region) under the marginal (joint) PPD. Unlike the frequentist confidence intervals that they replace, Bayesian credible intervals make intuitive probability statements about the true value of estimands (Gelman et al. 2013). In this way, Bayesian PPDs constructed from MCMC sampling represent the core of the present work; they compose a quantitative probability statement that has been heuristically calibrated to match the intrinsic uncertainty of the forecast problem.

To properly facilitate the direct Bayesian approach to parameter estimation and ensemble calibration, this research breaks from Einstein's view of causality and, in a figurative sense, *plays dice* with the atmosphere. Similar to the theoretical framework of Gneiting et al. (2007) and the NR/EMOS approach of Gneiting et al. (2005), a stochastic conceptual model will be assumed for relevant meteorological processes in which nature may be thought to draw observations (i.e., realizations of truth vis-à-vis Krzysztofowicz and Evans (2008) from hypothetical populations of credible atmospheric states. In this context, the forecast problem is no longer focused on dynamical relationships between observable weather variables, as it is assumed that the practical limitations of chaos theory prevent their precise specification through purely deterministic means. The emphasis now lies with a Bayesian estimation of unobservable model parameters that define an optimal fit of parametric, distributional beliefs to training data. In this way, the present work uses a combination of statistical, computational, and dynamical methods to calibrate ensemble guidance. Most notably,

1. **Bayesian estimation** inverts the canonical probability statement to seek inferences about model parameters from observed data instead of inferences about unobserved data from latent model parameters (G13). While the latter is recovered through the use of Bayesian PPDs, this research will nevertheless focus on distributional assumptions that parameterize relevant physical processes (e.g., a Gaussian likelihood function for diurnal maximum surface temperatures) and *a priori* beliefs (e.g., hierarchical Gaussian priors for the mean and log-variance of the model parameters) that characterize our assumed knowledge of the model parameters, and the intrinsic shape of the data, before the inference is completed.

2. **Markov chain Monte Carlo** (MCMC) sampling methods are frequently required to complete statistical inferences for real modeling problems (e.g., the atmosphere)—especially when higher-dimensional, multi-level, multiparameter Bayesian models are invoked. Although Bayesian inference provides a flexible template for statistical modeling that scales well with complicated systems, the posterior densities are often analytically intractable (G13). To this end, MCMC methods provide a robust numerical recipe for retrieving samples from so-called “target” distributions that are otherwise unobtainable.

3. **Dynamical meteorology**, which generally describes the deterministic reasoning of the physics of the atmosphere, is essential to inform appropriate modeling choices for the statistical relationships that describe the expected location and intrinsic spread of the training data and structure among the model parameters. While Bayesian statisticians are generally better-equipped to build predictive models with exceptional complexity, a subject matter expert (i.e., a meteorologist) can use their specialized knowledge of the data and the physical processes to shape important elements of the full Bayesian probability model.

THIS PAGE INTENTIONALLY LEFT BLANK

II. THEORY

The following sections describe fundamental concepts relevant to the hierarchical multivariate Bayesian multiple linear regression model developed for this dissertation. In particular, these ideas will permit an appropriate evaluation of the research questions and methods detailed in Chapter III.

A. KERNEL DENSITY ESTIMATION

Inspired by the utility of spectral decomposition in time series analysis, Rosenblatt (1956) and Parzen (1962) introduced the KDE technique to facilitate the non-parametric estimation of probability distributions. A bandwidth parameter and kernel (i.e., weighting) function are analytically combined to produce a linear combination of probability density functions (PDFs)—each centered on a discrete member of the sample (Figure 9a)—that smooth data to form a continuous probability distribution over the random variable of interest (Figure 9b). More specifically, the PDF f_n is analytically described by

$$f_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K(\frac{x-y}{h}) dF_n(y) = \frac{1}{nh} \sum_{j=1}^{n} K(\frac{x-X_j}{h}), \tag{1}$$

where subscript j identifies the independent and identically distributed (IID) random variables X_1, X_2, …, X_n, h denotes a positive bandwidth parameter, K signifies a kernel (or weighting) function, and F_n represents the cumulative distribution function with dummy variable y (Parzen 1962).

While the present work has explicitly pursued a parametric framework for Bayesian inference and probabilistic forecasting, KDE will still be required to evaluate some of the proper scoring rules described in Chapter III. In particular, the logarithmic score of Gneiting and Raftery (2007) requires the *a posteriori* probability density of observations. Since each predictive distribution is a collection of discrete forecast samples, KDE is needed to estimate the value of the corresponding PDF at the indicated observation. Additionally, KDE is used as a visualization aid for joint and marginal probability distributions sampled from MCMC methods in Bayesian posterior parameter beliefs and PPDs.

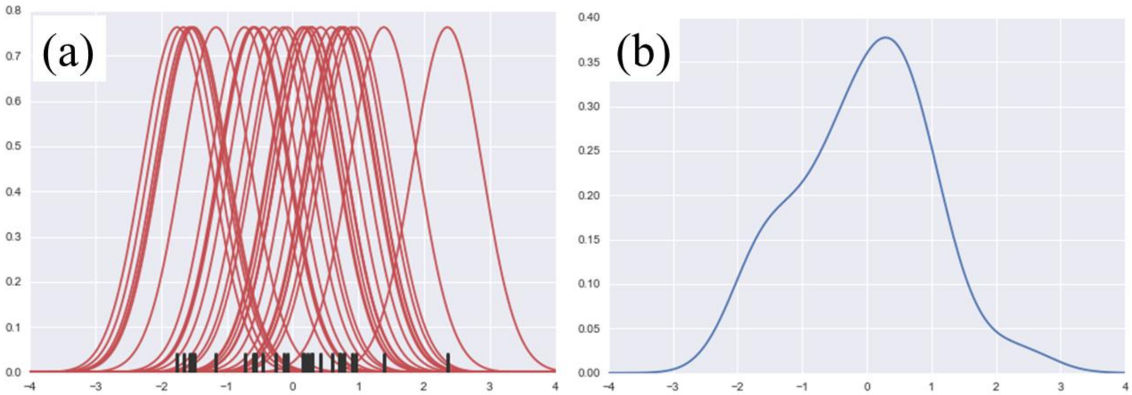


Figure 9. Example plots depicting the KDE technique. Source: Waskom 2015. Graphical representation of Gaussian kernel functions fit to an arbitrary set of discrete data in (a). In combination with the bandwidth parameter, the parameters of the Gaussian kernel function contribute to the structure of the resulting non-parametric continuous probability distribution.

B. ENSEMBLE MODEL OUTPUT STATISTICS

The method of ensemble model output statistics was introduced by Gneiting et al. (2005) to fit a single parametric distribution to training data in a simple statistical model (e.g., Equation 2). If one assumes that sensible weather variables are realized stochastically according to a Gaussian statistical relationship, then the NR/EMOS technique may be described as

$$y \sim N(\mu, \sigma^2), \tag{2}$$

where μ and σ^2 are the location (i.e., mean) and scale (i.e., variance) parameters of a Gaussian (i.e., normal) distribution; predictand y, the sensible weather variable for which a prediction is sought, is an IID random sample drawn from this PDF. EMOS models also use multiple linear regression to characterize the sensitivity of a univariate weather quantity—that is, the predictand—to a linear combination of independent predictor variables gathered from an EPS. In this way, it is similar to the regression framework of MOS; however, the latter assumes independent predictors have a deterministic relationship with the dependent predictand. The linear MOS concept can be described generically for multiple predictors as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K, \tag{3}$$

where $x_1, x_2, ..., x_K$ are a set of ensemble predictors, $\beta_1, \beta_2, ..., \beta_K$ are their associated regression coefficients (i.e., parameters), β_0 is the intercept term, and y is a sensible weather variable for which a prediction is sought. The scatter plot of forecasts and their corresponding observations in Figure 10 illustrates the basic elements of this predictor-response relationship when only a single predictor (i.e., $y = \beta_0 + \beta_1 x_1$) is considered. Each order pair represents a two-dimensional evaluation of Equation 3 for maximum diurnal surface wind speeds over 228 forecast trials. The vertical deviation between the dashed red and solid black lines represents the state-dependent forecast bias.

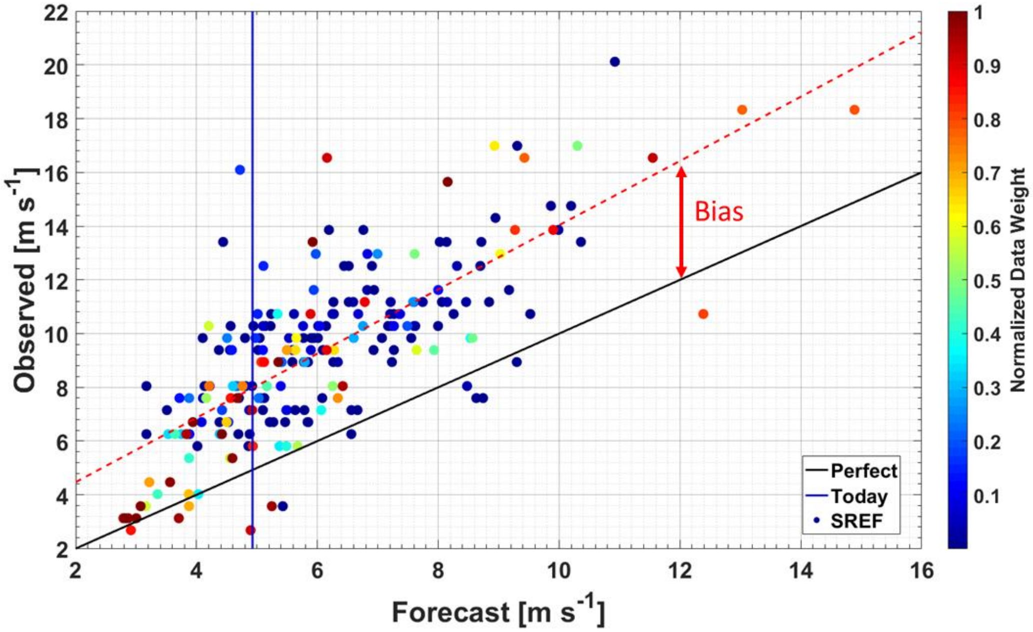


Figure 10. Forecast verification plot showing bias in ensemble mean forecasts of maximum diurnal surface wind speeds. The solid black line represents a perfect forecast-observation relationship; the dashed red line indicates a linear OLS regression fit to the data. The colorbar at right indicates the normalized age and weighting of the training data.

In this way, the MOS technique characterized by Equation 3 seeks a linear transformation that maps raw model output $x_1, x_2, ..., x_K$ to forecast estimates $f_n(x_1, x_2, ..., x_K)$ while optimizing an associated cost function. In simple terms, we seek a set of mathematical instructions that, on average, *would* have nudged raw model output, which is indicated by the abscissa of the independent predictor variable, toward the ordinate of the corresponding observation. To this end, the dashed red line in Figure 10 identifies the OLS solution to this regression, and it also represents the desired bias correction for the aforementioned two-dimensional MOS framework (i.e., $y = \beta_0 + \beta_1 x_1$).

In regression analysis, the intercept and regression parameters—that is, the "betas" in Equation 3—must be estimated from the data. We seek estimates for these betas that best fit the assumed mathematical relationship against the noisy superposition of random effects and errors contained in the set of independent and dependent variables. This is demonstrated by the variability of observed data around the OLS regression line in Figure 10. The randomness encoded in the data are expressed over repeated forecast instances, each with its own complicated interactions between measurement errors and EPS defects, where multiple combinations of predictors and predictands produce an overdetermined linear system of equations (Ron 2010). This linear system may be visualized as

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots + \beta_K x_{1,K} \\ y_2 &= \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots + \beta_K x_{2,K} \\ &\ldots \\ y_N &= \beta_0 + \beta_1 x_{N,1} + \beta_2 x_{N,2} + \ldots + \beta_K x_{N,K}, \end{aligned} \tag{4}$$

where each forecast trial adds an additional constraint to the linear model. This system can be described in matrix notation as $y = \beta_0 + Xb^T$, where

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix} \tag{5}$$

and $b = [\beta_1 \ldots \beta_K]$. The NxK matrix of predictor variables in Equation 5 is known as the design matrix. An alternative formulation of Equation 5 includes the intercept parameter in the beta vector and adds a column of ones (i.e., an identity column) to matrix X to permit a concise description of the linear system with $K-1$ predictors as

$$y = Xb^T, \tag{6}$$

where $b = [\beta_0 \ldots \beta_{K-1}]$ and the NxK design matrix is now written as

$$X = \begin{bmatrix} 1 & \cdots & x_{1,K-1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{N,K-1} \end{bmatrix}. \tag{7}$$

The classic OLS methods described in Chapter I engage this problem directly by minimizing the sum of squared residuals (SSR) between the true predictands and estimates derived from a linear combination of regression parameters and predictor variables (i.e., Equation 3). A heuristic representation of the fitting process is depicted in

Figure 11a with multiple combinations of betas producing distinct regression lines. The combination of parameter values that minimizes the SSR is considered the best point estimate for the unknown parameters (Feigelson 2015). In particular, this OLS cost function may be written as

$$SSR = \sum_{i=1}^{N} \left(y_i - x_i b^T \right)^2 , \tag{8}$$

where $x_i = \left[1 \ldots x_{i,K-1} \right]$ and $b = \left[\beta_0 \cdots \beta_{K-1} \right]$. It should be noted that linear transformations of this nature are frequently selected for analytical convenience and computational efficiency (e.g., Lorenz 1962 and Leith 1974); unless the modeler has a compelling motivation to consider nonlinear regression, the former is generally sufficient for most applications (Neter et al. 1996). Moreover, an arbitrary number of predictor variables may be considered from desperate sources (e.g., multi-model ensembles) according to the multiple linear regression framework in Kruschke (2014). While any number of physical processes may be involved in the dynamical realization of an arbitrary set of predictands, statistical regression provides a basis for the estimation of model parameters that appropriately characterize their stochastic (e.g., Equation 2) or deterministic (e.g., Equation 3) realization.

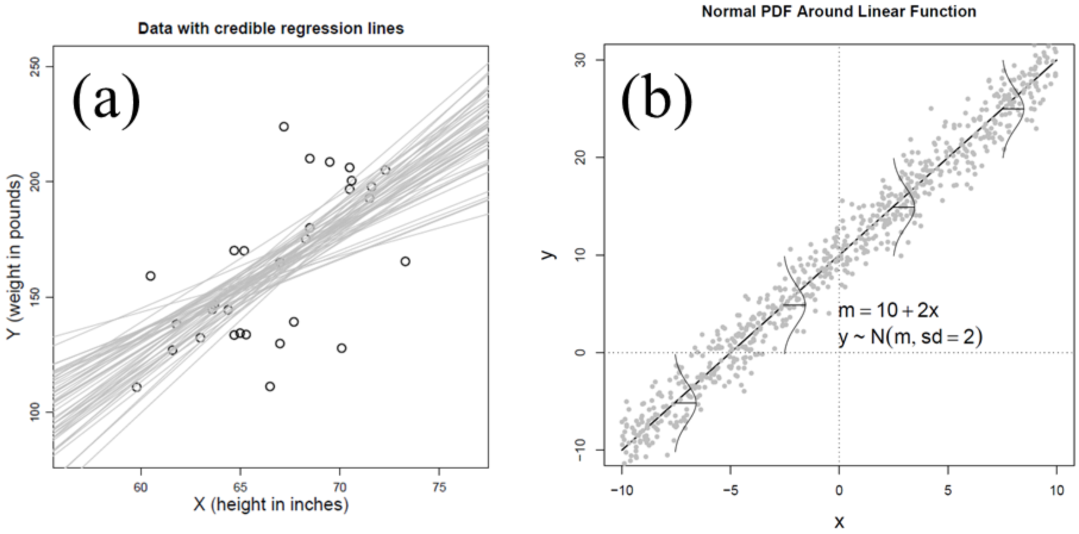


Figure 11. Examples of two-dimensional parameter estimation in regression analysis. Source: Kruschke (2014). Panel (a) depicts various regression lines when different values of the regression parameters are heuristically sampled; panel (b) shows a Gaussian PDF fit to a scatter of training data to illustrate the conceptual foundation of the NR/EMOS technique.

These OLS methods provide a simple template for adjusting contemporary ensemble guidance toward an outcome with greater statistical likelihood—a bias correction for meteorologists. It is also adaptive in the sense that the details of the mapping depend on the state of the independent predictor variables. In this way, Figure 10 demonstrates that a bias correction that is appropriate for a raw wind speed prediction less than 4 $[m\ s^{-1}]$ may not be ideal for a raw prediction exceeding 10 $[m\ s^{-1}]$; while the slope of the linear model is fixed by assumption, it is the vertical separation between the perfect forecast line (solid black) and regression solution (dashed red) that must be induced. This feature is commonly described as a state-dependent bias correction in the statistical post-processing literature (e.g., Hodyss et al. 2016), and it should be noted that the model used in this dissertation follows the standard convention of mapping ensemble output to observations (i.e., observed predictands)—not the inverse scheme explored by Hodyss et al. (2016).

The key innovation of NR/EMOS over traditional MOS, which produces single-valued forecast estimates consistent with Equation 3 and Figure 10, is the additional parameterization of this multiple linear regression framework inside of the parametric distribution in Equation 2. In this way, the statistical modeler is playing dice with the atmosphere; they are making an assumption about the manner by which data are generated—that is, meteorological truth realized stochastically by nature. More specifically, this creates a basis for mapping ensemble components to distributional parameters as

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{K-1} x_{K-1} \tag{9}$$

$$\sigma^2 = \gamma_0 + \gamma_1 s^2, \tag{10}$$

where μ and σ^2 are the location (i.e., mean) and scale (i.e., variance) parameters of the Gaussian PDF, as before, and γ_1 and γ_2 are regression parameters on the variance (i.e., s^2) of the ensemble predictors $x_1, x_2, ..., x_{K-1}$. A pedagogical visualization of this relationship is found in Figure 11b for a two-dimensional regression scenario (i.e., $\mu = \beta_0 + \beta_1 x_1$). For NR/EMOS models, the cost function is modified to estimate the parameters according to the method of minimum CRPS; the details of this analysis are beyond the scope of the dissertation and available to eager readers in Gneiting et al. (2005) and Richter (2012).

Unlike traditional parameter estimation techniques, which make direct inferences about the parameters of the selected continuous distribution (e.g., μ and σ^2), EMOS requires estimates for the parameters of the multiple linear regression specified by Equation 9 and Equation 10. Contemporary predictors (i.e., current values for $x_1, x_2, ..., x_{K-1}$) are then inserted into the multiple linear regression structure of the model to produce updated estimates for the distributional parameters (e.g., μ and σ^2). Once these parameters have been specified, synthetic data (i.e., unobserved observable random variables) can be randomly sampled to produce a calibrated probabilistic forecast over the desired sensible weather variables (e.g., Figure 6 and Figure 7).

C. BAYESIAN ENSEMBLE MODEL OUTPUT STATISTICS

A direct Bayesian extension of Gneiting et al. (2005) was introduced by Richter (2012) to provide an alternate means of parameter estimation for the NR/EMOS technique. A rigorous treatment of the foundational benefits of the Bayesian approach to parameter estimation is beyond the scope of this dissertation and is, in fact, part of larger contemporary debate within the statistical community (e.g., the validity of hypothesis testing, statistical significance, and p-values). The present work does not seek participation in this conflict. Nevertheless, the advantages of intuitive, common-sense Bayesian statistical conclusions detailed by Casella (2008), Gelman et al. (2013), and Kruschke (2014) provide a fundamental motivation for the BEMOS technique. Moreover, Gelman et al. (2013) and Kruschke (2014) note the utility of the Bayesian approach to data analysis in formulating sophisticated probability models. Gelman et al. (2013) defines this process as

> 1. Setting up a *full probability model*—a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.
>
> 2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution*—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
>
> 3. Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps.

Richter (2012) found that a Bayesian approach produced calibrated forecast distributions with improved mean absolute error (MAE) and CRPS performance (Table 1) compared with competing OLS and EMOS methods. Hodyss et al. (2016) similarly advocated for “a direct application of Bayes’ rule” in their examination of BMA in significant wave height forecasts. To this end, this dissertation relies on a Bayesian framework for prediction that produces counterfactual statements about *unobserved* observable forecast quantities conditioned on *a posteriori* beliefs about multiple linear regression parameters (e.g., Equation 9).

"With inferential statistics," Kruschke (2014) observes, "we don't just introspect to find the truth. Instead, we rely on data from observations. Based on the data, what should we believe in?" As described by this dissertation's research motivations, it is advantageous to explicitly quantify the underlying uncertainty associated with an inference. To properly develop the methods (Chapter III) of this dissertation, consider again the pedagogical regression analysis explored in the previous section. If the observed predictands are assigned to random variable y in a two-dimensional stochastic framework as

$$y(x \mid \beta_0, \beta_1, \sigma^2) \sim N(\beta_0 + \beta_1 x, \sigma^2), \tag{11}$$

where y is drawn from a normal distribution with mean $\mu = \beta_0 + \beta_1 x$, σ^2 is the variance associated with Gaussian "noise" in the data, β_0 and β_1 are unknown regression coefficients (i.e., the intercept and slope of the desired post-processing transformation), and x is the independent predictor (i.e., model output), then the regression problem has been suitably reformulated as indicated in Figure 11b. That is, a statistical inference is sought for the latent model parameters (i.e., β_0, β_1, and σ^2) that will provide an ideal fit to the data.

Kruschke (2014) notes that this parameter estimation can be completed by classical methodologies and, specifically, by the well-known maximum likelihood estimation (MLE) procedure introduced by Fisher (1922). While a full treatment of MLE concepts is also beyond the scope of the present work, the MLE approach does share an important feature with Bayesian parameter estimation. It minimizes a cost function that quantifies the likelihood of the data as the model parameters are varied over the indicated support of the latter. This cost function is known, quite appropriately, as the likelihood function in statistical inference, and it is a function of the latent parameters—not the observed data. If multiple predictors are considered (i.e., Equation 9), MLE seeks an extreme value of

$$l(\underline{\theta}) = \sum_{i=1}^{N} \ln f\left(y_i \mid \underline{\theta}\right) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i b^T)^2, \tag{12}$$

where l is the log-likelihood for the model parameters in vector $\underline{\theta}$ over N forecast trials given the fixed observations in random variable $y = \left[y_1 \ldots y_N\right]$, the i$^{\text{th}}$ predictors in $x_i = \left[1 \ldots x_{i,K-1}\right]$, regression parameters $b = \left[\beta_0 \ldots \beta_{K-1}\right]$, and the unknown variance in σ^2.

As is common in the statistical literature (e.g., Fisher 1922), the log-likelihood has been reported in Equation 12 to simplify the details of the analysis; finding the extremum of the log-likelihood is identical, and often more computationally more convenient, than finding the extremum of the likelihood. It should be noted that a generic likelihood function may be constructed for any parametric distribution by fixing the data (vis-à-vis the Bayesian perspective) and evaluating the product of sequences over N forecast trials of each observation as a function of the parameters of the specified PDF. In this way, the log-likelihood converts the logarithm of a product of sequences into a sum of logarithmic sequences through the canonical properties of logarithms. The Gaussian log-likelihood in Equation 12 is also notable for the additional interaction between the logarithm and the exponential term in the probability density for the Gaussian distribution; as inverse functions, the right-hand side of Equation 12 becomes a convenient sum over the Mahalanobis distance, which is simply proportional to the Euclidean distance in the one-dimensional case considered by Equation 12 (Gu 2008). In matrix notation, Gu (2008) writes the Mahalanobis distance for the multivariate Gaussian distribution (or multivariate normal distribution; MVN) as

$$\Delta^2 = \left(\underline{y} - \underline{\mu}\right)^T \Sigma^{-1}\left(\underline{y} - \underline{\mu}\right), \tag{13}$$

where $\underline{y}$ is vector of predictands, $\underline{\mu}$ is a vector form of Equation 9 that is formed by a matrix product of the predictors and regression coefficients, and Σ defines the associated covariance matrix. This quantity will be critical for the methods detailed in Chapter III where both MVN and multivariate t location-scale (MVT) distributions will be used as likelihood functions in the BEMOS model developed by the present work.

The MLE solution to the procedures described by Equation 11 and Equation 12 will produce single-valued, point estimates of the parameters in the inference; values that are most probable for the sample of data considered given the model structure. The analytical development of similarities between traditional OLS and MLE approaches—especially when residuals are assumed to have a Gaussian distribution—is similarly beyond the scope of this dissertation; however, the eager reader can find examples in Feigelson (2015) and most foundational texts covering regression analyses (e.g., Neter et al. 1996). While this may be sufficient for some applications, there is no explicit uncertainty information for the underlying parameter inference in these approaches. MLE also lacks a mechanism for combining empirical evidence with *a priori* beliefs (e.g., based on other similar data, the variance σ^2 is *probably* less than one). While additional development of classical frequentists methods can mitigate some of these limitations, Bayesian data analysis provides a different approach to the same fundamental optimization problem articulated by Fisher (1922) and, indeed, the model specified by Equation 11 and Equation 12. Casella (2008) even notes that "likelihoodists [are] Bayesians, but they don't know it." More specifically, we seek an inference scheme that permits model parameters as random variables and parameter estimation that quantifies uncertainty explicitly via Bayesian posterior distributions. As a result, the BEMOS technique has been selected by the present work based on the findings of Richter (2012) and Hodyss et al. (2016) and the aforementioned arguments.

A concise description of Bayesian reasoning, which was notably advanced by Laplace and other contemporary mathematicians (Stigler 1990), is provided by Gelman et al. (2013) as a procedure to "invert the probability statement and obtain probability statements about [parameters] given observed [data]." Adopting the notation of Gelman et al. (2013) for unobservable model parameters $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_K)$ and observable data y , Bayes' rule is frequently derived by expanding a generic joint probability distribution, $p(\underline{\theta}, y)$, as

$$p(\underline{\theta}, y) = p(\underline{\theta} \mid y)\, p(y) = p(y \mid \underline{\theta})\, p(\underline{\theta}), \tag{14}$$

where $p(y\,|\,\underline{\theta})$ is the conditional probability of observed data given the model parameters, $p(\underline{\theta}\,|\,y)$ is the conditional probability of model parameters given the observed data, and $p(\underline{\theta})$ and $p(y)$ are the marginal probabilities of the model parameters and the data. An example for observable data y would be observations for an arbitrary sensible weather variable (e.g., surface temperatures). Unobservable, latent model parameters are quantities in the model structure for which direct observations are unavailable (e.g., the mean and variance of a Gaussian PDF); they must be computed with descriptive statistics or otherwise inferred from observable data. Bayes' theorem itself is then revealed by a trivial manipulation of Equation 14 to produce

$$p(\underline{\theta}\,|\,y)=\frac{p(y\,|\,\underline{\theta})\,p(\underline{\theta})}{p(y)}, \tag{15}$$

where $p(y\,|\,\underline{\theta})$ now identifies the well-known likelihood function or, sometimes, the sampling distribution, $p(\underline{\theta})$ is the subjective prior, and $p(y)$ is the prior predictive distribution; the latter is sometimes described, more simply, as the evidence for the observed data with this probability model (Kruschke 2014).

As with Lorenz and chaos theory, the implications of Equation 15 are difficult to overstate. In simple terms, there is value in "inverting the probability statement" to update our beliefs in model parameters, regardless of their subjective origins, and, therefore, refine the inferences they inform with data (Gelman et al. 2013). Unlike the non-parametric schemes and simple linear regression methods previously considered, which work directly with objective evidence in a more naïve fashion, this expression produces a more rigorous basis—and cost function—for the parameter optimization in Equation 11 and Equation 12. This Bayesian framework can similarly exploit the chain rule of probability indicated by Equation 14, as described by Hodyss et al. (2016), to construct Bayesian JPMs with an arbitrarily large number of parameters. In this way, Bayesian reasoning can be easily scaled to generate tractable probability models of exceptional sophistication (Gelman et al. 2013). To this end, the multiple linear regression framework of Kruschke (2014) extends the description provided by Equation 11 and Figure 11b to include multiple predictors that are *potential* covariates of the predictand. That is, a predictor may ultimately be found to have little influence on a predictand and, therefore, be appropriately associated with a trivial beta-slope (i.e., small values of the normalized regression coefficients). In simple terms, this means that OLS, MLE, and Bayesian parameter estimation schemes can diagnose the sensitivity of a predictand to variance in the independent predictor and, if necessary, return a null result. For these reasons, the model uncertainty analysis of Richter (2012) is considered superfluous by the present work.

With the foundation of a Bayesian inference framework established, the full multiple linear regression model may now be expressed for a Gaussian likelihood function as

$$y(x\,|\,\underline{\theta}) \sim N(\sum_{i=0}^{K-1}\beta_i x_i, \sigma^2), \tag{16}$$

where the location parameter μ in Equation 2 has been reparametrized with $K-1$ predictors in an expanded linear model with homoscedasticity (i.e., constant, state-independent variance); using this notation, $x_0 = 1$. In this way, the Bayesian probability model is now structured to capture the sensitivity of the observable predictand to multiple predictor variables through the magnitudes of the various regression coefficients. Predictions extracted from this generative model will now include the relevant variability of multiple covariates, which may themselves be correlated in the linear model. This multiplicative interaction can be similarly modeled in Equation 16 with products of predictors (e.g., $\beta_{12}x_1x_2$) that are nonlinear functions of the predictors but linear functions of the regression coefficients (Kruschke 2014).

If the inference structure is to be further extended to capture covariance among multiple response variables, in such a way that y becomes a vector $\underline{y}$ of statistically coupled predictions for several predictands (e.g., tropical cyclone position *and* intensity), then a *multivariate* multiple linear regression framework is required. More specifically, Equation 12 and Equation 16 can be modified according to Sinay and Hsu (2014) and Gu (2008) for a linear system to construct a MVN log-likelihood as

$$l(B,\Sigma) \propto -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(y_i^T - x_i^T B)\Sigma^{-1}(y_i^T - x_i^T B)^T, \qquad (17)$$

where l is the log-likelihood of the observations, y_i^T is a $1xM$ vector of observations, x_i^T is an $1xK$ vector of independent predictors, B is a KxM parameter matrix of regression coefficients (including the intercepts), and Σ is an MxM covariance matrix that replaces the univariate variance σ^2. For a case with three predictands and three predictor variables, which describes a model with $M=3$ and $K-1=3$, the coefficient and covariance matrices expand as

$$B = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \\ \beta_3 & \beta_4 & \beta_5 \\ \beta_6 & \beta_7 & \beta_8 \\ \beta_9 & \beta_{10} & \beta_{11} \end{bmatrix} \qquad (18)$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}, \qquad (19)$$

where σ_i describes the standard deviation of the corresponding predictand y_i and ρ_{ij} identifies the Pearson correlation coefficient (PCC) between predictands y_i and y_j. The first column of the coefficient matrix describes the sensitivity of y_1 to variance along each predictive dimension of x_i^T using multiple linear regression; that is, β_0 is the intercept associated with y_1, β_3 measures the sensitivity of y_1 to predictor x_1, β_6 measures the sensitivity of y_1 to predictor x_2, and β_9 measures the sensitivity of y_1 to predictor x_3. Similarly, β_1 is the intercept associated with y_2, β_4 measures the sensitivity of y_2 to predictor x_1, β_7 measures the sensitivity of y_2 to predictor x_2, and β_{10} measures the sensitivity of y_2 to predictor x_3. Finally, β_2 is the intercept associated with y_3, β_5 measures the sensitivity of y_3 to predictor x_1, β_8 measures the sensitivity of y_3 to predictor x_2, and β_{11} measures the sensitivity of y_3 to predictor x_3. In this way, a stochastic formulation of the multivariate multiple linear regression framework with $M = 3$ and $K - 1 = 3$ is appropriately visualized as a linear system with

$$\begin{aligned} \mu_{n,1} &= \beta_0 + \beta_3 x_{n,1} + \beta_6 x_{n,2} + \beta_9 x_{n,3} \\ \mu_{n,2} &= \beta_1 + \beta_4 x_{n,1} + \beta_7 x_{n,2} + \beta_{10} x_{n,3} \\ \mu_{n,3} &= \beta_2 + \beta_5 x_{n,1} + \beta_8 x_{n,2} + \beta_{11} x_{n,3}, \end{aligned} \tag{20}$$

where μ describes the i^{th} location parameter of $MVN(\underline{\mu}, \Sigma)$ with covariance matrix Σ; subscript n describes the n^{th} forecast trial in a training period with N total forecast-observation data comparisons. The precise number of predictands, M, predictor variables, $K - 1$, and training points, N, is selected by the statistical modeler and the constraints of the problem according to the “underlying scientific problem and the data collection process” (Gelman et al. 2013). Moreover, this multivariate multiple linear regression framework—which represents a combination of modeling approaches assembled from Sinay and Hsu (2014) and K14, but is also commonly found in the statistical literature for linear models (e.g., Neter et al. 1996)—can be inserted into any parametric distribution appropriate for the forecast application. In this way, it is consistent with primary motivations of NR/EMOS, which sought a convenient parametric representation of the forecast uncertainty communicated by a linear combination of EPS output.

Bayesian estimation permits a wide variety of likelihood functions—that is, $p(y \,|\, \underline{\theta})$—when constructing the JPM. In the context of the BEMOS approach to statistical post-processing, the selection of the Bayesian likelihood function specifies the stochastic character of the data generation process—the manner by which meteorological truth is drawn from our hypothetical parametric forecast distributions. While many candidate parametric distributions exist, including the $Beta(\alpha, \beta)$, $Gamma(k, \theta)$, and $\chi^2(k)$, or chi-squared, distributions, the ubiquitous Gaussian—that is, $\mathrm{N}(\mu, \sigma^2)$—finds wide application throughout statistics and Bayesian data analysis. The multivariate form of $\mathrm{N}(\mu, \sigma^2)$ (e.g., Figure 12) is similarly analytically tractable and computationally efficient (e.g., the convenient form of Equation 17). Moreover, $\mathrm{N}(\mu, \sigma^2)$ is statistically conjugate with itself when used as a Bayesian likelihood and prior. Gelman et al. (2013) defines such conjugacy as “the property that the posterior distribution follows the same parametric form as the prior distribution.” The latter was important in the original BEMOS analysis of Richter (2012); it permits direct analytical solutions to Bayesian parameter estimation and, in some cases, the availability of more efficient MCMC techniques (e.g., Gibbs sampling).

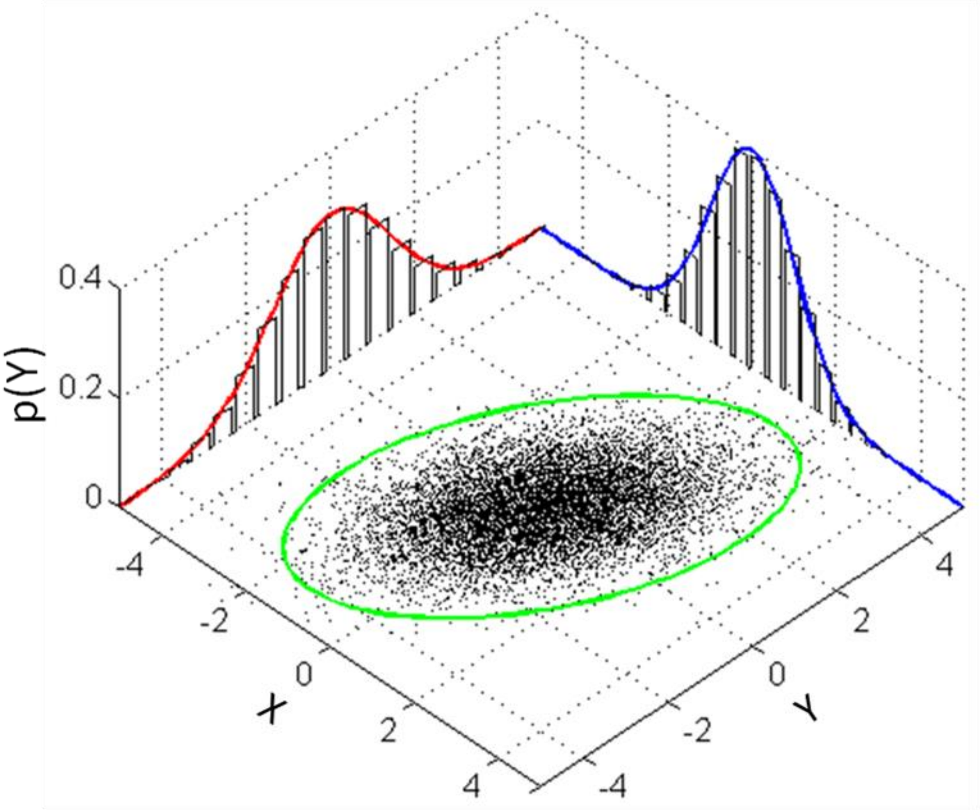


Figure 12. A pedagogical representation of a two-dimensional multivariate Gaussian distribution (MVN) with elliptical presentation (i.e., non-zero covariance elements). Source: Wikipedia 2017a. The scatter of points in the horizontal plane represents a discrete sample from the joint probability distribution. Each dimension of the joint PDF can be decomposed into one-dimensional Gaussian marginal distributions.

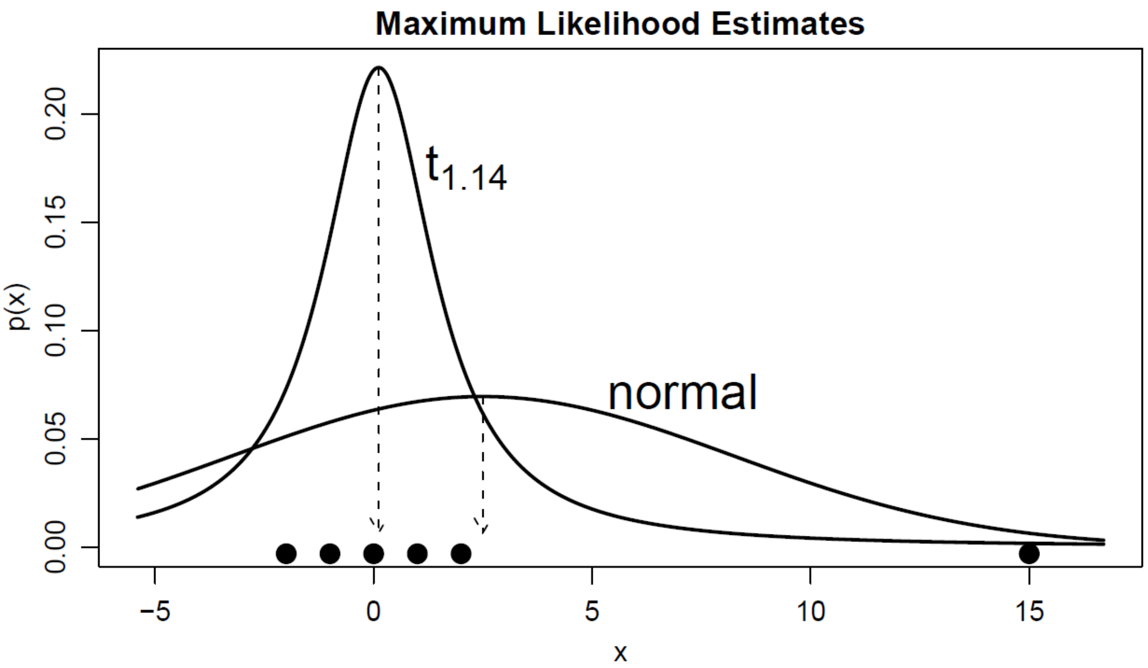


Figure 13. Pedagogical comparison of univariate parameter estimation performed with a Gaussian (i.e., normal) distribution and the classic Student's t-distribution. Source: Kruschke (2014). The location parameter of each distribution is estimated based on the statistical properties of the underlying data. The presence of an outlier (extreme right) causes the location of the Gaussian distribution to shift further towards the outlier than the t-distribution.

While the present work has sought the widest application of the MVN as a likelihood function, there are reasons to consider other distributional forms. Kruschke (2014) notes that outliers in data can sometimes produce undesirable consequences in parameter inference schemes. In particular, Figure 13 demonstrates the influence of a single outlier on the MLE estimates produced by fitting a univariate Gaussian (i.e., normal) and a Student's t-distribution to a "toy" data sample (Kruschke 2014). Both PDFs are symmetric with the bulk of the probabilistic mass concentrated near the central tendency of the distribution; however, the t-distribution shifts some of its probabilistic mass further into the tails. The latter property of the t-distribution becomes advantageous in so-called robust regression, which is defined in this context as a procedure to mitigate the impact of outliers in parameter estimation. In this way, Kruschke (2014) notes:

> Outliers are simply data values that fall unusually far from the model's expected value. Real data often contain outliers, presumably because some extraneous influences have sporadically perturbed the data values. Sometimes these extraneous influences can be identified, and the affected data values can be corrected. But usually we have no way of knowing whether a suspected outlying value was caused by an extraneous influence, or is a genuine reflection of the target being measured. Instead of deleting suspected outliers from the data according to some arbitrary criterion, we retain all the data but use a likelihood function that is less affected by outliers than is the normal distribution. (Kruschke 2014)

The central tendency of the t-distribution in Figure 13 is less sensitive to the outlier at $x = 15$ precisely because it permits more probabilistic mass in the tails, or wings, of its PDF. In other words, the t-distribution is allowed to assign higher probabilistic density to locations further away from its mean and, as a result, can keep its location parameter closer to the central tendency of the rest of the data (i.e., the points assumed not to be outliers). In this way, the t-distribution provides a concise method of robust regression in Bayesian estimation and an attractive alternative to the Gaussian likelihood function examined in Equation 12. Since no separate model specifications are required to formulate this approach (e.g., removing outliers from the training set), it is computationally convenient vis-à-vis the MCMC sampling techniques described in the following section. Kruschke (2014) adds that

> the estimated parameter values for a normal likelihood function can be greatly distorted by outliers in the data. This sensitivity also occurs when the normal distribution is used in linear regression. The normal likelihood function demands that the regression line is vertically close to all the data points, because the likelihood value is tiny for points more than about three standard deviations from the line. Consequently, outlying data points can have disproportionate leverage in the estimate of the regression coefficients. (Kruschke 2014)

Wiecki (2013) considers the practical impact of outliers on Bayesian parameter estimation using the PyMC3 Python package for Bayesian statistical modeling and Probabilistic Machine Learning. Using a collection of normally distributed errors around simulated data (blue), a true regression line can be compared with Bayesian estimation solutions with univariate Gaussian (panel a) and t-distribution (panel b) likelihood functions. The regression performance of each likelihood function is represented by the family of regression solutions (i.e., black lines) in each panel. The outliers found in the top left of each panel have a notable influence on the regression solutions for the Gaussian likelihood function (i.e., panel a); by comparison, the t-distribution likelihood function produces regression estimates that remain closer to the true value (i.e., panel b). As a result, the regression residuals produced by the Gaussian likelihood function are larger near the boundaries of the data interval considered and, perhaps more importantly, a poorer fit to the underlying data.

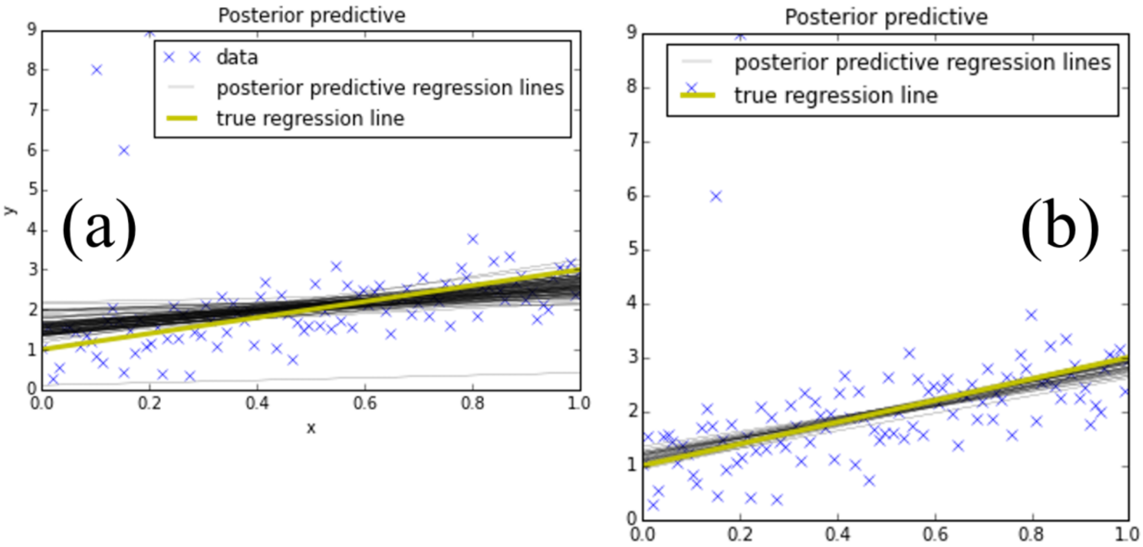


Figure 14. A practical comparison of robust regression in Bayesian parameter estimation using univariate Gaussian (left) and t-distribution (right) likelihood functions. Source: Wiecki (2013). The influence of outliers on each probability model is represented by the fit of the regression solutions (black lines) compared with the true solution (yellow line). The simulated data (blue) was generated assuming normally distributed errors with outliers (top left) added manually.

To this end, a multivariate form of the t location-scale distribution (MVT), which merely describes a non-standardized form of the t-distribution, has been adapted for the present work. The PDF for the MVT is frequently represented as a combination of the MVN and χ^2 (i.e., chi-squared) distributions as

$$t_\nu(\underline{x};\underline{\mu},\Sigma,\nu)=\frac{\Gamma\left[(\nu+M)/2\right]}{\Gamma(\nu/2)\nu^{M/2}\pi^{M/2}\left|\Sigma\right|^{1/2}}\left[1+\frac{1}{\nu}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})\right]^{-(\nu+M)/2}, \qquad (21)$$

where ν is the degrees-of-freedom parameter, $\underline{\mu}$ and $\underline{x}$ are as before, and M specifies the number of predictands (i.e., the dimensionality of $\underline{\mu}$, $\underline{x}$, and $\underline{y}$). The stochastic samples $\underline{x}$ produced by the multivariate PDF t_ν are formed according to

$$\frac{\underline{y}}{\sqrt{u/\nu}}=\underline{x}-\underline{\mu} \qquad (22)$$

where $\underline{y}$ and u are IID samples from the $MVN(\underline{\mu},\Sigma)$ and $\chi^2(\nu)$ (i.e., chi-squared) distributions, respectively, and $\underline{x}$ is "said to have a (nonsingular) multivariate t-distribution" (Lin 1972). Finally, an appropriate modification to the MVN log-likelihood in Equation 17 must be specified. Unfortunately, the lack of an exponential function in Equation 21 renders a convenient form of Equation 17 unattainable; the MVT log-likelihood must therefore be written as

$$l(B,\Sigma)\propto N\ln\left[\frac{\Gamma\left[(\nu+M)/2\right]}{\Gamma(\nu/2)\nu^{M/2}\pi^{M/2}\left|\Sigma\right|^{1/2}}\right]-\frac{\nu+M}{2}\sum_{i=1}^{N}\ln\left[1+\frac{1}{\nu}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right], \quad (23)$$

where $\underline{y}$ now describes a vector of observations from training data, $\underline{\mu}$ is parameterized according to Equation 20, and all other variables are as before. The logarithm of the ratio on the left hand side of Equation 23 has been retained for parsimonious specification in this document. The expression on the right-hand side of Equation 23 also retains a logarithm and removes the benefits of conjugacy in Bayesian probability models with Gaussian prior distributions. This is notable for the MCMC methods described in the next section, as closed-form solutions are not available when JPM conjugacy is broken.

Bayesian inference requires the specification of prior information. In this way, the modeler must choose the distributional forms of each term in the numerator of Equation 15 and what numerical information is stored in each. While the details of specific parametric selections used in this dissertation will be reserved for Chapter III, the interaction of each component in the JPM can be diagrammed to articulate the analytical structure of each aspect of the probability model. Figure 15a shows a JPM diagram for a pedagogical Bayesian model consistent with Equation 17; that is, a univariate stochastic framework for predictand y_i with a Gaussian generative process (i.e., a normal likelihood function with mean μ_i and standard deviation τ_y) wrapped around the multiple linear regression framework for μ_i specified by Equation 9. The distributions above this level represent Bayesian prior information selected for each of the model parameters. For this example, the lone intercept parameter β_0 is given a Gaussian prior with fixed location hyperparameter M_0 and fixed standard deviation hyperparameter T_0; this communicates the modeler's *a priori* belief that model parameter β_0 has a true value inside the support of the indicated distribution.

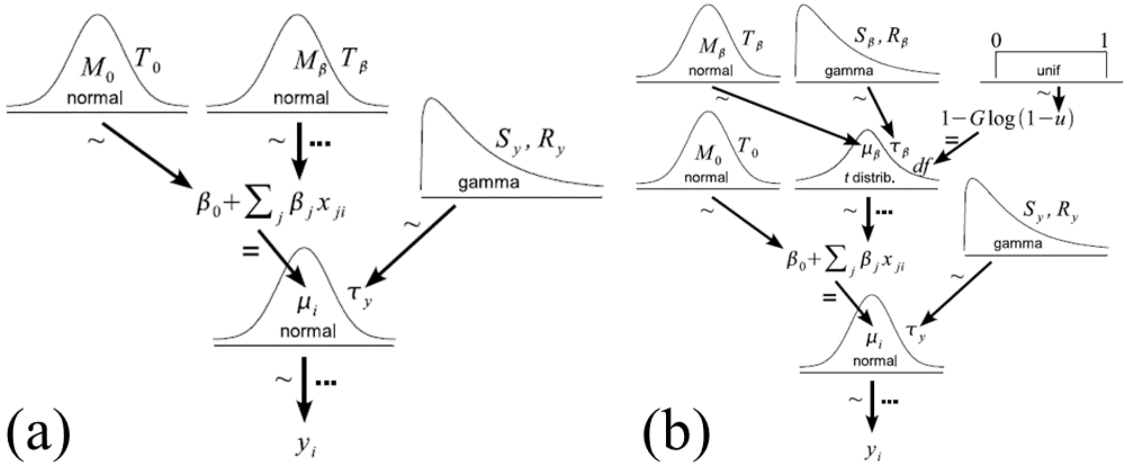


Figure 15. JPM diagrams for two multiple linear regression models. Source: Kruschke (2014). The panel on the left (a) depicts a non-hierarchical model in which the hyperparameters are fixed. The panel on the right (b) depicts a hierarchical model in which the hyperparameters of the Bayesian priors are treated as random variables and influenced by the data via hyperpriors. Standard, non-hierarchical priors with fixed hyperparameters (i.e., hyperpriors) are placed on the hierarchical hyperparameters.

Similarly, the regression parameters β_j are assigned identical Gaussian priors with fixed hyperparameters M_β and T_β—the location and scale parameters of the prior. Finally, the variance of the Gaussian likelihood fit to the data, τ_y, is assigned a Gamma prior distribution with fixed hyperparameters S_y and R_y - the shape and scale parameters of the indicated Gamma distribution. The latter communicates the modeler's *a priori* belief that model parameter τ_y has a true value inside the support of the positively-skewed Gamma distribution with these fixed hyperparameter values. In this way, the Bayesian prior modifies the cost functions of Equation 12 and Equation 17 to weight likely parameter values with additional, external information—a modification to the MLE optimization process introduced by Fisher (1922). It allows a subject matter expert to imbed scientific reasoning into the JPM—not simply from the distribution chosen for the likelihood function, which describes the manner in which observed and *unobserved* data are stochastically realized, but also from beliefs *about* model parameters we include in the Bayesian prior.

However, it should be noted that the absence of Bayesian prior information, which is sometimes described as "letting the data speak for itself," is also considered a part of Bayesian data analysis (Gelman et al. 2013). This is often implemented with early versions of the Bayesian data analysis process, when confidence in *a priori* beliefs is either low or incomplete, through so-called noninformative or weakly informative priors. Gelman et al. (2013) describes a weakly informative distribution as a Bayesian prior "which contains some information—enough to 'regularize' the posterior distribution, that is, to keep it roughly within reasonable bounds—but without attempting to fully capture one's scientific knowledge about the underlying parameter." Using the JPM in Figure 15a as an example, a noninformative approach would simply remove the prior distributions above the likelihood function; this is equivalent to setting $p(\underline{\theta}) \propto 1$ in Equation 15. Unfortunately, the latter method can also produce improper posterior distributions that do not integrate to unity over the full support of the distribution (Gelman et al. 2013). In this way, care must be taken with the selection of Bayesian prior information to ensure posterior beliefs in model parameters are valid. In the present work, JPMs with noninformative priors are labeled as MLE; while not truly equivalent to the classic method introduced by Fisher (1922), they are also not fully Bayesian in character either. However, a primary advantage of Bayesian data analysis with noninformative priors is that it permits an equivalent MLE analysis while simultaneously delivering the associated Bayesian uncertainty estimates—with no separate or additional calculations required. In this regard, Bayesian parameter estimation on the MLE cost function—that is, the likelihood function—might be considered superior to the classical frequentist estimates they replace.

For many Bayesian probability models, including the multiple linear regression framework described by Sinay and Hsu (2014) and Kruschke (2014), the prior distributions can be augmented with hyperparameters that are also influenced by the data. In simple terms, this means the Bayesian priors have hyperparameters that are not fixed; they are treated as random variables that must also be estimated from data. To this end, additional level(s) of stochastic expressions, each with their own hyperparameters and priors can be added above the final level indicated in Figure 15a. This is known as a hierarchical Bayesian model and Figure 15b shows how the pedagogical JPM in Figure 15a can be modified to permit parameter inference on additional levels. While the prior information for the regression intercept β_0 and likelihood τ_y remain unchanged, Figure 15b shows an example of regression coefficient shrinkage; that is, the regression coefficients β_j are now grouped together with a common t-distribution prior and hyperparameters μ_β, τ_β, and df. These hyperparameters correspond to the location, scale, and shape parameters of the t-distribution and are treated as model parameters—that is, they are estimated from the data in the same manner as other model parameters. Since these non-fixed hyperparameters are treated as random variables, they also require prior information. The latter is indicated by a new level of distributions at the top of Figure 15b; each hyperparameter in the t-distribution now has its own prior distribution—a hyperprior—that performs the same function as any prior distribution. More specifically, these hyperpriors describe the modeler's *a priori* belief in the hyperparameters—only now on this new level of information. Bayesian data analysis permits other forms of hierarchical modeling, where clustering in the underlying data can be modeled with parameters organized at different model levels (e.g., global and regional regression coefficients).

Hierarchical models capture additional structure in the parameters that can improve the fit of a JPM to the data. Kruschke (2014) describes the benefit of regression coefficient shrinkage in this context:

> A desirable side-effect of incorporating this prior structure is that the estimates of the regression coefficients experience shrinkage. If many regression coefficients are near zero, then the t distribution will have a high precision (τ parameter), which in turn will shrink the estimates of the regression coefficients. The regression coefficients are mutually informing each other, via the prior knowledge that they should be distributed according to a t-distribution. The shrinkage is desirable not only because it expresses our prior knowledge, but also because it rationally helps control for “false alarms” in declaring that a predictor has a non-zero regression coefficient. When there are many candidate predictors, some of them may spuriously appear to have credibly non-zero regression coefficients, even when the true coefficient is zero. This sort of false alarm is unavoidable because data are randomly sampled, and there will be occasional coincidences of data that are unrepresentative. (Kruschke 2014)

While the present work does not consider the selection of t-distribution priors to be a necessary condition for regression coefficient shrinkage, since the Gaussian may be considered an extension of the t-distribution when the degrees of freedom have become sufficiently large (e.g., rules of thumb indicating $\nu \geq 30$), it does consider the presence of the hierarchical hyperparameters to be sufficient. This assumption will be revisited in Chapter III. To this end, both the MVN and MVT likelihood functions selected for this dissertation use hierarchical Gaussian priors. The analytical details of the resulting hierarchical multivariate Bayesian models, which include the relevant JPM diagrams, will be similarly reserved for Chapter III.

The inferences produced by the methods described heretofore focus exclusively on the parameters of the selected JPM. In this way, Bayesian posteriors are useful for diagnostic analyses of model parameters, because they describe optimal updated statistical beliefs in the elements of the generative model. However, they are not explicitly prognostic. If predictive statements about *unobserved* data are desired, as required by the nature of the statistical post-processing methodologies engaged by the present work, then additional analysis must be pursued. In particular, the posterior belief in model parameters acquired though Bayesian estimation, which is contained in the joint probability distribution of $p(\underline{\theta} \mid y)$, must be weighted by the conditional probability of observing the *unobserved* random variable and summed over all possible values of the model parameters (Gelman et al. 2013). To this end, let z be an *unobserved* observable random variable (i.e., synthetic data), y be an *observed* observable random variable (i.e., training data), and $p(z \mid y)$ describe the posterior predictive distribution as

$$p(z \mid y) = \int p(z, \underline{\theta} \mid y) d\underline{\theta} = \int p(z \mid \underline{\theta}) p(\underline{\theta} \mid y) d\underline{\theta}, \tag{24}$$

where the quantity $p(z \mid \underline{\theta})$ is functionally equivalent to the likelihood selected for the JPM (Gelman et al. 2013). While this expression is analytically intractable for some models, including many implementations of the MVT likelihood function and, indeed, many forms of hierarchical parameter inference, the same MCMC techniques that helped popularize Bayesian estimation can produce joint posterior samples from $p(\underline{\theta} \mid y)$ that are easily repurposed for a discrete analogue of Equation 24. In simple terms, the MCMC samples produced by the methods described in the next section produce solutions for $p(z \mid y)$ with little additional computational expense or modeling effort.

D. MARKOV CHAIN MONTE CARLO METHODS

In some applications, the Bayesian estimation described in the previous section can be completed without the aid of computational techniques. These circumstances typically result from the selection of noninformative or conjugate priors (e.g., panels a, b, and c in Figure 16 from Wiecki 2015), which guarantee that, for a given likelihood function, the Bayesian posterior and prior will belong to the same distributional family (Kruschke 2014). While computationally tractable and conceptually convenient, conjugate priors can limit the scope and utility of Bayesian models—perhaps most notably in the complexity of the JPM permitted by the inference. In this way, MCMC computational methods are frequently invoked to *complete* the inference in Bayesian "models with many parameters and complicated multilayered probability specifications"

(Gelman et al. 2013). While a full description of MCMC theory is beyond the scope of the present work, the eager reader is encouraged to consult Gelman et al. (2013) or Gilks et al. (1996)—among many others. In this way, the essential elements of one popular approach—the Metropolis algorithm (Metropolis et al. 1953)—will be developed to demonstrate their application to the hierarchical multivariate Bayesian probability models identified in Chapter III.

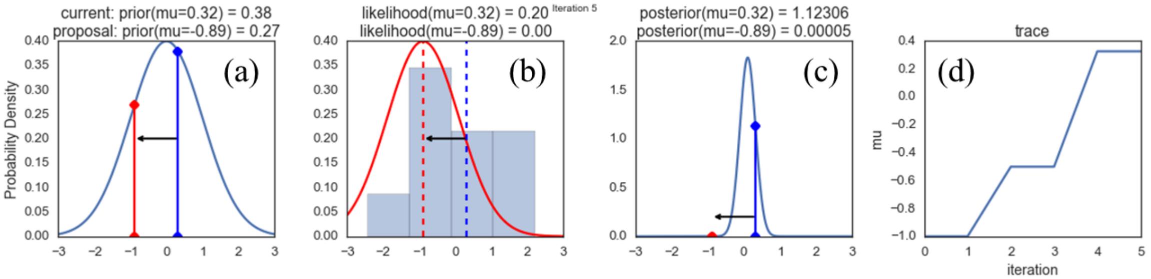


Figure 16. Pedagogical visualization of a Bayesian posterior distribution formed during a single iteration of the Metropolis algorithm. Source: Wiecki (2015). The individual panels depict, from left to right, a comparison of the (a) prior, (b) likelihood, and (c) posterior distributions at the current (blue) and proposed (red) phase space locations. Panel (d) is a trace diagram that records the sequence of chain states visited by the algorithm.

The primary obstacle to arbitrarily complex models in Bayesian data analysis is the specification of the normalizing constant in Equation 15—that is, the denominator in Bayes’ theorem (Gilks et al. 1996). To this end, Gelman et al. (2013) expands $p(y)$ in Equation 15 as

$$p(y) = \int p(y, \underline{\theta}) d\underline{\theta} = \int p(y \mid \underline{\theta}) p(\underline{\theta}) d\underline{\theta}, \tag{25}$$

where the interpretations of all quantities remain unchanged from their original descriptions in Equation 15 and Equation 24. Unlike the other elements of Equation 15, which are functions of the model parameters $\underline{\theta}$, the quantity defined in Equation 25 has marginalized this dependence out of the integrand. Although a constant with respect to the parameters, this expression—which ensures the Bayesian posterior integrates to unity (i.e., produces a PDF)—is frequently unavailable in closed-form (Gelman et al. 2013). Without the benefits of Markov chain simulation, this analytical obstacle prevents an otherwise-viable Bayesian probability model from completing the necessary parameter inferences.

To this end, a computational scheme introduced by Metropolis et al. (1953), and later generalized by Hastings (1970), provides a method of drawing samples from arbitrarily complicated target distributions that are intractable by classical methods (Robert and Casella 2011). Compared to other schemes available for parameter estimation (e.g., the expectation-maximization algorithm in Raftery et al. 2005), MCMC methods offer many advantages (Vrugt et al. 2008). In particular, these MCMC methodologies leverage a special property of Markov chains—that is, the Markov property—for which the development of any state, θ_t, in a sequence of random variables is dependent solely on the previous state, θ_{t-1}, and a transition kernel, T, that stochastically maps every possible state of θ to any other (Gilks et al. 1996). As a result, this sequence of states is considered memoryless; each additional step is independent of all previous states visited and, perhaps more importantly, the sequence of states produced in this manner—that is, the Markov chain—represents a random walk through the phase space of the random variables and the transition kernel (Kruschke 2014).

The efficacy of Markov chain simulation in Bayesian estimation depends on three additional properties: if the chains are observed to be irreducible (i.e., transitions to and from any phase space location have non-zero probability and are possible in finite time), aperiodic (i.e., no diagnosable periodicity in the frequency with which phase space locations are visited), and non-transient, then the Markov chain is guaranteed to "forget" its initial state (Gilks et al. 1996) and converge to a unique stationary distribution (Gelman et al. 2013). MCMC methods exploit these properties by reverse-engineering a Markov chain with a stationary distribution that matches the Bayesian target—that is, $p(\underline{\theta} \mid y) \propto p(y \mid \underline{\theta}) p(\underline{\theta})$, the unnormalized Bayesian posterior. In particular, we seek a Markov chain that visits phase space locations with a relative frequency that is statistically consistent with the probabilistic mass contained within the desired joint probability distribution. The analytical development of this condition, which guarantees the equivalence of the stationary and target distributions, is similarly beyond the scope of this dissertation; however, Gilks et al. (1996), Gelman et al. (2013), and Kruschke (2014) provide the necessary details. Nevertheless, a parsimonious description of the Metropolis algorithm reproduces the most relevant portions of the associated proof.

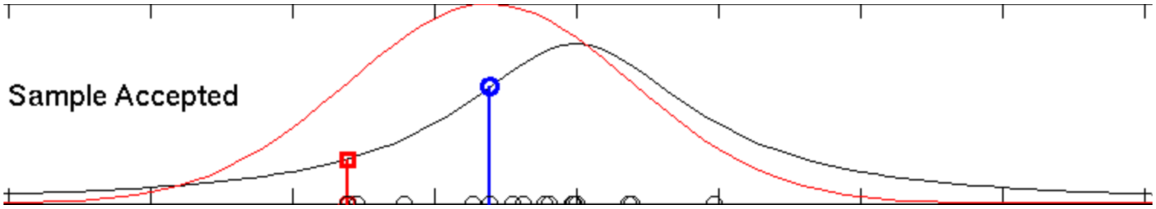


Figure 17. Visualization of a univariate Gaussian proposal distribution (red curve) superimposed on a target distribution (black curve). Source: Stansbury (2012a). Samples from the target are generated according to the Metropolis algorithm, which compares the target density at two locations: the current position (blue vertical line) and at the proposed location (red vertical line).

To generate samples from an arbitrary target distribution over $\underline{\theta}$, we seek a Markov chain that produces a random walk through the multi-dimensional phase space of the parameters (Kruschke 2014). To accomplish this, Metropolis et al. (1953) introduced a simple modification to Markov chain mechanics that ensures the chain "finds" the desired stationary distribution (Gelman et al. 2013). If the unnormalized Bayesian posterior is the target (e.g., the black distribution in Figure 17), the Metropolis algorithm generates each step in the resulting sequence by evaluating the quantity $p(\underline{\theta} \mid y)$ at two locations in the phase space of the model: the current location, $\underline{\theta}_t$, and a proposal location, $\underline{\theta}_{t+1}$ (red vertical line), which is stochastically drawn from a symmetric proposal distribution (red distribution) centered on the current value of $\underline{\theta}_t$ (blue vertical line). Recall that the *observed* observable data y is fixed by our Bayesian assumptions; the ratio of the target densities at the proposed and current locations (i.e., the ratio of the ordinates of the red box and blue circle in Figure 17) may now be written as

$$r = \frac{p(\underline{\theta}_{t+1} \mid y)}{p(\underline{\theta}_t \mid y)} = \frac{\dfrac{p(y \mid \underline{\theta}_{t+1})\, p(\underline{\theta}_{t+1})}{p(y)}}{\dfrac{p(y \mid \underline{\theta}_t)\, p(\underline{\theta}_t)}{p(y)}} \tag{26}$$

$$r = \frac{p(\underline{\theta}_{t+1} \mid y)}{p(\underline{\theta}_t \mid y)} = \frac{p(y \mid \underline{\theta}_{t+1})}{p(y \mid \underline{\theta}_t)} \frac{p(\underline{\theta}_{t+1})}{p(\underline{\theta}_t)}, \tag{27}$$

where the expression in Equation 27 is, quite conveniently, the ratio of the unnormalized Bayesian posterior (i.e., the product of the likelihood function and the prior) at the proposed and current phase-space locations.

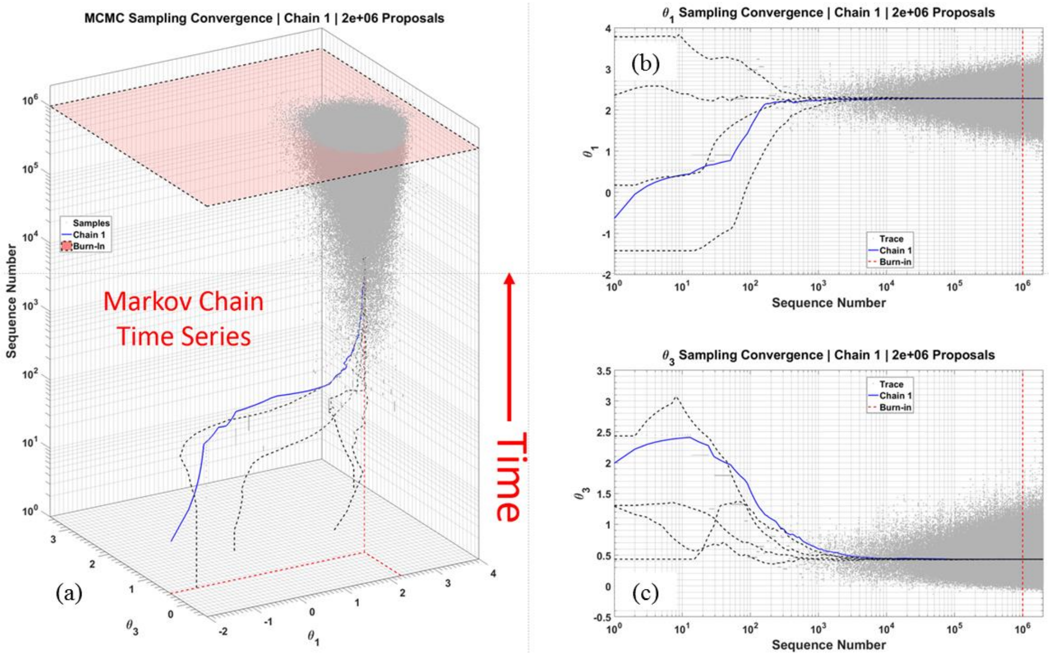


Figure 18. Multiparameter trace diagram for five parallel Markov chains in a semi-log time series of discrete phase space locations (grey dots). The moving average of each chain (solid blue and dashed black curves) is superimposed over the discrete MCMC samples (grey filled circles). Panel (a) shows a joint trace for two model parameters; panels (b) and (c) depict the associated marginal traces (i.e., cross-sectional views) for these parameters. The discrete time sequence of the Markov chain increases vertically (up) in panel (a) and horizontally (right) in panels (b) and (c).

Using this expression, transitions are accepted with probability $\min(r,1)$; that is, the location $\underline{\theta}_{t+1}$ is "accepted" and appended to the sequence if a random number drawn from the open interval $(0,1)$ is less than r (Gelman et al. 2013). If this condition is not satisfied, then $\underline{\theta}_{t+1}$ is "rejected" and the current value of $\underline{\theta}_t$ is appended as the next element of the chain. This process is repeated in the above manner until the moving average of multi-dimensional states reaches its stationary distribution. Monte Carlo samples of the desired target distribution may be taken from any portion of the chain that has reached its stationary distribution (i.e., converged) in this manner. It should be noted that proposal distribution is also sometimes known as the transition distribution, transitional kernel, or jumping distribution. If this proposal distribution is not symmetric, then the generalized Metropolis-Hastings algorithm is required; analytical details for the latter are available in Gilks et al. (1996) and Gelman et al. (2013).

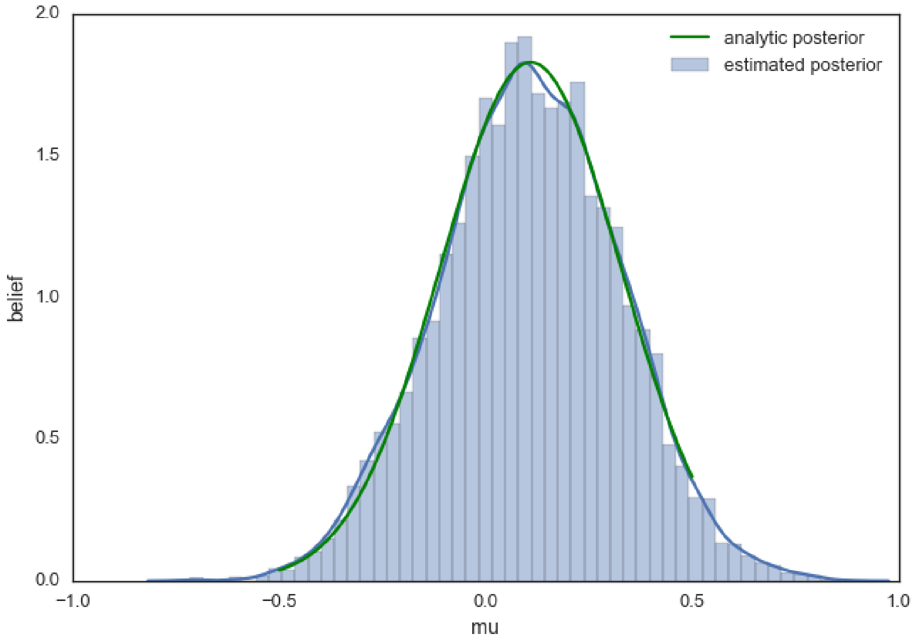


Figure 19. Comparison of analytical (green plot) and MCMC (blue histogram) solutions for a simple Bayesian inference scheme with a conjugate normal prior distribution. Source: Wiecki (2015). The normal posterior distribution represented in the histogram was generated by extending the Metropolis algorithm represented in Figure 16 through 15,000 additional MCMC iterations.

The Markov chain created by this procedure does not discriminate between multi-dimensional states that have been accepted or rejected; the Metropolis algorithm merely records the sequence of phase space locations visited during this random walk (e.g., Figure 18). In this way, transitions to phase-space locations with lower target density are permitted (e.g., the red square has a lower density than the blue circle in Figure 17); however, we always accept transitions to locations with a greater density in the joint probability distribution and, as a result, these locations will be recorded more frequently in the multi-dimensional sequence of the evolving chain (the black circles at the bottom of Figure 17). Equation 26 and Equation 27 also reveal the mechanism by which the normalizing constant is managed in MCMC algorithms. While crucial to the completion of the Bayesian inference by purely analytical means, the normalization constant $p(y)$ is absent from the ratio in Equation 27 because it does not depend on the variability of the multi-dimensional parameter $\underline{\theta}$ and has been identically cancelled on the right-hand side of Equation 26; the data are considered fixed and all dependence on model parameters has been marginalized out of the integrand in Equation 25. The resulting expression, which requires only the likelihood function and the Bayesian prior, can be computationally evaluated for arbitrarily complex Bayesian models. This gives the Bayesian approach to parameter estimation tremendous flexibility in JPM selection and, ideally, the fit of the inference scheme to data conditioned on model assumption and *a priori* beliefs.

Finally, the posterior samples produced by MCMC methods require additional manipulation to make probability statements about *unobserved* observable data (i.e., predictions). To this end, the likelihood function found in MLE and Bayesian estimation can be similarly tasked to produce samples from Bayesian PPDs (Gelman et al. 2013). This reflects the goals of the NR/EMOS technique, which sought a simple parametric representation of the probabilistic forecasts meteorologist extract of ensembles. MCMC methods make this process conceptually and computational efficient. As suggested by Equation 2, a Gaussian likelihood function similarly assumes that an *unobserved* observable random variable z is stochastically drawn as

$$z \sim N(\mu, \sigma^2), \tag{28}$$

where μ and σ^2 are location and shape parameters that may be arbitrarily reparametrized in a linear model (e.g., Equation 9). Multi-dimensional samples from the joint posterior distribution of unobservable, latent parameters $\underline{\theta}$ (e.g., Figure 18) can be combined with updated predictor variables (e.g., current values for the ensemble predictors in Equation 3) to iteratively populate the parametric elements of the corresponding predictive distribution (e.g., Equation 9). In this way, PPD samples are generated by a discrete formulation of Equation 24 as

$$p(z \mid y) = \sum p(z \mid \underline{\theta}) p(\underline{\theta} \mid y), \tag{29}$$

where each sample of z is stochastically drawn from the likelihood specified by $p(z \mid \underline{\theta})$ according to the multi-dimensional posterior parameter samples in $p(\underline{\theta} \mid y)$. While no prohibitions on the number of z samples drawn from the "warmed," final $\underline{\theta}$ samples are known to exist, so long as the same number of z samples drawn from each cross-sectional element of the warmed MCMC chain are equal, the present work assumes that a 1:1 ratio between z and $\underline{\theta}$ samples—that is, one z sample for each $\underline{\theta}_i$—is sufficient to accurately specify the desired PPDs. The details of PPD construction will be further developed in Chapter III. In this way, Bayesian estimation and MCMC methods provide a complete inference structure, for both unobservable model parameters and unobserved observable data, in a wide variety of practical predictive modeling applications.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODS

The present work adapts a number of techniques from statistical post processing, statistics, and data science to produce a functional predictive model. At its core, the hierarchical multivariate Bayesian approach to atmospheric prediction developed for this dissertation represents an extension of the multivariate multiple linear regression detailed by Sinay and Hsu (2014) and Kruschke (2014), the stochastic predictive framework of Gneiting et al. (2005), and perhaps most importantly, the Bayesian methods explored by Richter (2012). However, the BEMOS model in the present work is distinct from Gneiting et al. (2005) and Richter (2012) in the following ways:

1. It permits full multivariate predictions, with coupled covariance between vector predictands, with MVN and MVT likelihood functions.

2. It uses hierarchical parameter structure, which treats the hyperparameters of Bayesian prior distributions as random variables with their own hyperprior distributions, to permit hyperparameter inference from the data.

3. It uses a customized adaptive multivariate Metropolis sampler, with block-wise component updates, to permit more sophisticated, non-conjugate probability models (e.g., the aforementioned hierarchical structure).

4. It can engage robust regression with multivariate t location-scale distributions (i.e., MVT likelihood functions).

5. It uses logarithmic data transformations to produce asymmetric predictive distributions for sensible weather variables with intrinsic skew.

As described by the research motivations in the introduction, these model innovations are consistent with the recommendations of Gneiting (2014)—especially the desire for coupled statistical predictions of sensible weather variables. They also attempt to make a unique contribution to statistical post-processing research in meteorology. In this way, the Naval Postgraduate School (NPS) BEMOS model represents a proof of concept that may ultimately be applied with more rigor to nearly any predictive application where training data are available (e.g., electromagnetic propagation, tropical cyclone, and surface spectral reflection prediction).

A. RESEARCH QUESTIONS

The first conjecture of this dissertation is that a MVT likelihood function will provide better predictive performance, with forecast distributions that are better calibrated, than a MVN likelihood function in the NPS BEMOS JPM. In this way, the likelihood function contains the modeler's belief in the stochastic generative process that best describes the distribution of the observable data. It also represents a core component of the model's cost function. With the data fixed, the likelihood function describes how the model parameters may be heuristically perturbed over the support of the prior to make inferences regarding the underlying model parameters. As described in Chapter II, the Gaussian PDF is convenient in statistical modeling; the analytical form of the PDF is easily adapted in log-likelihood expressions (e.g., Equation 12), and $N(\mu,\sigma^2)$ is conjugate with itself in Bayesian posterior inference. Many of the relevant investigations cited in the literature review (e.g., Gneiting et al. 2005, Krzysztofowicz and Evans 2008, Richter 2012, and Hodyss et al. 2016) fit Gaussian likelihood functions. However, the advantages of robust regression are well known in applied statistics (e.g., Wiecki 2013 and Kruschke 2014). To this end, the present work will explore the comparative efficacy of the MVN and MVT distributions as likelihood functions in the application considered by Chapter IV. The t-distribution produces location parameter estimates that are less sensitive to outliers vis-à-vis robust regression; this property is assumed to be desirable in atmospheric prediction.

The second conjecture of this dissertation is that hierarchical Bayesian priors will provide better predictive performance than noninformative priors in the NPS BEMOS JPM. This research question therefore explores the sensitivity of predictive performance to Bayesian prior information. Chapter II introduced some of the basic distinctions between the MLE and Bayesian approaches to parameter estimation. While the former introduced the analytical form of the likelihood as a cost function, the latter effectively extends this approach by combining it with prior information. It also produces full posterior distributions over the model parameters of interest with explicit uncertainty information that is delivered in parallel with the primary inference—especially when paired with MCMC sampling techniques. To investigate the utility of Bayesian priors, especially with the hierarchical methods detailed by Gelman et al. (2013) and Kruschke (2014) for regression coefficient shrinkage (e.g., Figure 15b), the present work will compare hierarchical Bayesian model instances against similar models with noninformative priors—that is, $p(\underline{\theta}) \propto 1$; the former are label as "NPS" models and the latter as "MLE" models. These hierarchical JPMs contain informative and, ostensibly, meaningful prior information that should contribute to enhanced predictive performance. It should be noted that the "MLE" model instances created by the present work are not truly representative of the original MLE approach introduced by Fisher (1922), because they benefit from a Bayesian/MCMC formulation of the likelihood function. In this way, Bayesian predictive models with identical likelihood functions, for both the MVN and MVT distributions, will also be compared with OLS methods to see if hierarchical prior information improves predictive performance relative to classical techniques.

The third conjecture of this dissertation is that larger amounts training data will produce better predictive performance than smaller sets of training data. Moreover, this conjecture further assumes that training data that is more meteorologically similar (e.g., synoptic pattern) to the test period will produce an NPS BEMOS model that performs better than when trained with data that is identical in size yet more meteorologically dissimilar to the test period. This research question therefore examines the impact of training period length and character on predictive performance. The former describes the amount of data given to the model (e.g., two months vs. 12 months); the latter describes the similarity of the training data to the test data used to evaluate predictive performance (e.g., training data from April and May 2016 to evaluate predictive performance during the same months in 2017). The majority of the statistical post-processing literature considers this question at some length (e.g., Raftery et al. 2005, Gneiting et al. 2005, Berrocal et al. 2008, and Hodyss et al. 2016), as it has relevance to the obtainable forecast skill, its associated computational cost, and the need for countermeasures against overfitting the data. This analysis may be broadly described as statistical cross-validation;

contemporary inferences are evaluated against one another as training samples of varying length and/or composition are considered. In simple terms, this dissertation will assume more training data are better and, when data sparsity is a constraint, similar data are preferable. This conjecture has particular relevance in operational applications, where large sets of quality-controlled training data may be unavailable or irrelevant (e.g., forecast personnel deployed internationally to data-sparse regions). Moreover, operational schedules may preclude model training with large data; in these circumstances, computational efficiency is paramount.

The fourth conjecture of this dissertation is that a NPS BEMOS model trained on ensemble control elements (i.e., information available outside of the parent EPS) will be able to outperform the raw EPS and, indeed, comparable statistical reference methods (i.e., OLS) dressed with frequentist variance estimates. This conjecture further assumes that predictors based on the central tendency (i.e., mean) of the parent ensemble will provide better predictive performance than the aforementioned control predictors. While multiple linear regression permits an arbitrarily large number of predictor variables, the computational efficiency of such an approach can rapidly diminish—especially for Bayesian inferences completed with MCMC methods. The practical limitations associated operational applications therefore provide an additional incentive for parsimonious model specifications. However, all of the statistical methods previously described are more computationally efficient than the EPS schemes they seek to improve. In this way, the present work seeks a foundational comparison with the ensemble approach itself. More specifically, it seeks a performance comparison with EPS output that assesses the importance of the information provided by the ensemble distribution. If a Bayesian generative model can offer predictive performance that meets or exceeds raw ensemble output, for both measures-oriented and distributions-oriented scoring rules, then the computational cost EPS forecasts will need to be carefully considered against the merits of comparatively more efficient statistical approaches. To this end, model instances trained with ensemble control predictors—information available outside of an EPS—will be compared with similar variants trained with predictors derived from the ensemble mean. The latter models will be identified with “NPSR” and “OLSR” model instances.

B. DATA TRANSFORMATIONS

The parameter estimation schemes examined in this dissertation operate on large systems of linear equations (e.g., Equation 4 and Equation 20). While the cost functions in Equation 8 and Equation 11 can process raw training data (i.e., no systematic manipulations with analytical functions), it is common in the statistical literature to apply mathematical transformations (e.g., Manikandan 2010, Gelman et al. 2013, Kruschke 2014 and Hodyss et al. 2016) to the data before the primary analysis begins. The goal of such transformations is to shape the numerical character of the data (e.g., it’s central tendency and spread) to make them more suitable for the assumptions and structure of the model.

1. Z-Score Standardization

Standardization in statistical analysis describes a transformation that shifts and scales each member of the data so that the resulting set has a convenient central tendency and spread (e.g., $\mu = 0$ and $\sigma^2 = 1$). It is common practice in statistical modeling (e.g., Neter et al. 1996 and Kruschke 2014) to compute z-scores for each dimension of the training data—that is both, predictors and predictands. The transformed data may then be written as

$$z(x) = \frac{x - \mu_s}{s_x}, \tag{30}$$

where x describes an arbitrary data point, μ_s is the mean of the raw data sample, and s_x is the standard deviation of the raw sample; the transformed element, $z(x)$, now belongs to a sample that has been centered (i.e., $\mu_z = 0$) and normalized (i.e., $s_z^2 = 1$). This form of standardization permits comparisons between beta regression coefficients, which indicate the impact of each variable in the data set.

Standardized regression coefficients reflect the normalized sensitivity of the associated predictand to normalized perturbations in a single predictor while all other predictors are held constant; if $\beta_3 = 0.5$ in Equation 18, for example, then this implies y_1 will experience $0.5s_y$ change for every s_x change in x_1. Moreover, standardized data have no units and permit the mixing of predictor variables of different types (e.g., temperatures and wind speeds) in the same linear model. Neter et al. (1996) advise caution when interpreting beta regression coefficients, because correlations between predictors can mask their “comparative importance.” However, Kruschke (2014) notes the importance of standardization in MCMC sampling as illustrated in Figure 20:

> In principle, we could run the [model] on the raw x, y data. In practice, however, the attempt often fails. There’s nothing wrong with the mathematics or logic, the problem is that believable values of the slope and intercept parameters tend to be tightly correlated, and this narrow diagonal zone of believability is difficult for sampling algorithms to explore…MCMC sampling can be made much more efficient if the data are standardized. Standardizing each variable is straight forward. (Kruschke 2014)

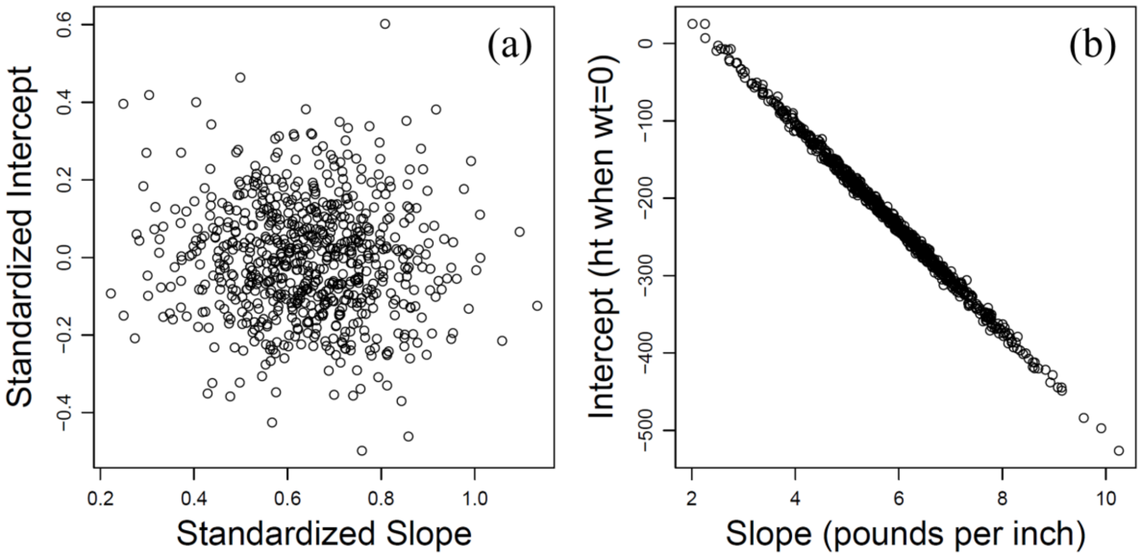


Figure 20. Comparison of standardized (left; a) and raw (right; b) regression parameter estimation procedures. Source: Kruschke (2014). The intercept and slope parameters of a linear regression model with a single predictor. MCMC samples from each approach are indicated with black circles. The "intercept" corresponds to "height" data when the "weight" variable is zero with pedagogical "toy" data.

To this end, the present work has adopted standardized regression coefficients for three primary reasons:

1. The MCMC-convergence of some Bayesian JPMs was found to rely strongly on z-score standardization; Figure 20b depicts the issues for MCMC methods: "sampling from such a tightly correlated distribution is typically very difficult to do directly. It is difficult to discover a point in the narrow zone in the first place. Then, having discovered a viable point, the chain does not move efficiently" (Kruschke 2014).

2. The relative strength in predictive influence can generally be inferred by the magnitude of the associated regression parameter (e.g., $\beta_3 \gg \beta_6$ in Equation 18 would suggest that x_1 is a better predictor for y_1 than x_2); conversely, small beta values indicate relatively poor predictive influence—a null result that can be equally useful in prioritizing or eliminating superfluous features in the training data.

3. Standardized data permits mixed predictors for the same predictand; for example, temperature, wind speed, and pressure could be used in a

multiple linear regression model to predict any other sensible weather variable—regardless of units.

2. Log Transformations

Hodyss et al. (2016) considered an additional step to their direct approach to Bayesian predictive modeling; they applied a log transformation to the raw data to make it “more Gaussian before analytic evaluation through Bayes’ rule.” Like z-score standardization, this transformation is common in statistical analysis (e.g., Gelman et al. 2013 and Kruschke 2014). The goal of logarithmic transformations is to shape the data to fit the parametric assumptions of the statistical model. In parameter estimation schemes like MLE and Bayesian data analysis, this is employed to improve the fit of the likelihood function to training data—especially when the natural variability of the data is expressed with positive skew. Figure 21 illustrates the impact of various log transformations on so-called “toy” (i.e., pedagogical) data. The IID samples in random variable X (top panel; blue) have a positive (right) skew and a non-Gaussian presentation; however, a log transformation of X produces a more symmetric distribution (top panel; green) with no skew. The IID samples in random variable Y (bottom panel; blue) are normally distributed with no perceptible skew and a Gaussian presentation; however, a log transformation applied to this data produces a distribution (bottom panel; green) with a comparatively smaller negative (left) skew and a quasi-Gaussian presentation.

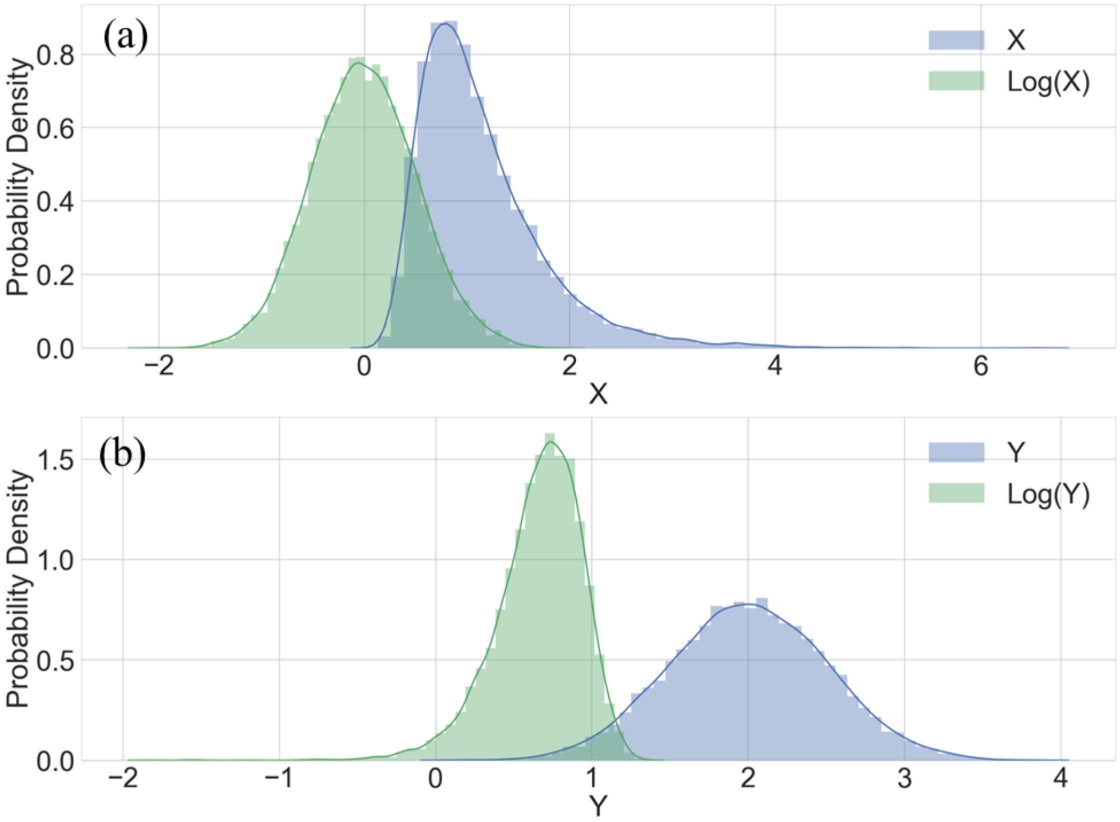


Figure 21. Visualization of log transformations for pedagogical "toy" data with positive (blue; panel a) and zero (blue; panel b) skew. In both cases, the resulting distributions (green; both panels) have negligible skew and Gaussian presentations.

In this way, log transformations can give raw data with undesirable positive skew and non-Gaussian presentations a more suitable Gaussian appearance in the transformed coordinate space of the statistical model. Similarly, log transformations applied to raw data that is already Gaussian in appearance have little effect; while the spread of the resulting distribution has changed, the skew of the transformed data is comparatively small and does not notably degrade its Gaussian appearance. In both cases, changes to the central tendency and spread of the transformed data will be shifted and scaled by the z-score standardization techniques in the previous section so that the final data are relatively consistent with a standard normal distribution—that is, $N(0,1)$. In this way, the present work uses a combination of contemporary data transformation techniques to ensure raw training data provides a good fit to MVN and MVT likelihood functions described in Chapter II and the following section.

C. MODEL STRUCTURE

1. Likelihood Functions

The NPS BEMOS model uses MVN and MVT likelihood functions. The analytical details of the latter were introduced in Chapter II with Equation 21, Equation 22, and Equation 23. The log-likelihood of the former was introduced in Equation 17; however, an important computational modification has been introduced into the present work. The so-called "trace trick" in Gu (2008) is frequently invoked in linear algebra to simplify the triple matrix product in Equation 17, which is a multivariate form of the Mahalanobis distance from Equation 13, to generate an updated expression for the log-likelihood. The simplified MVN log-likelihood may therefore be written as

$$l(\mathrm{B},\Sigma) = -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}tr\left[\Sigma^{-1}\sum_{i=1}^{N}(y_i^T - x_i^T B)^T(y_i^T - x_i^T B)\right], \tag{31}$$

where l is the log-likelihood of data with N forecast instances, y_i^T is a $1xM$ vector of observations, x_i^T is an $1xK$ vector of independent predictors, B is a KxM parameter matrix of regression coefficients (including the intercepts), Σ is an MxM covariance matrix that replaces the univariate variance σ^2, and tr describes the matrix trace of the quantity in brackets.

It should be noted that Σ^{-1} has been conveniently removed from the Mahalanobis distance term and, perhaps most importantly, brought outside of the summation of sequences where it can be computed a single time. Moreover, the quantity inside of the summation is recognizable as a Euclidean outer product for each forecast instance of the training data—that is, $(y_i^T - x_i^T B) \otimes (y_i^T - x_i^T B)$. The summation is therefore expanded as

$$\sum_{i=1}^{N}(y_i^T - x_i^T B)^T(y_i^T - x_i^T B) = [Y - XB]^T[Y - XB], \tag{32}$$

where the capital letters now indicate full matrix represents of each variable over the N forecast instances of the training period; Y is a NxM matrix of observations, X is a NxK matrix of predictors, and B is a KxM matrix of regression parameters. Finally, the matrix trace of the quantity inside of the brackets in Equation 31 can now be solved as a linear system (i.e., $x = A^{-1}b \Rightarrow Ax = b$) according to

$$tr\left[\Sigma^{-1}\sum_{i=1}^{N}(y_i^T - x_i^T B)^T (y_i^T - x_i^T B)\right] = tr\left[\Sigma^{-1}[Y - XB]^T [Y - XB]\right] \tag{33}$$

where Σ^{-1} and $[Y - XB]^T [Y - XB]$ assume the role of A and b, respectively, in the linear system $Ax = b$. When simplified to Equation 33, the MVN log-likelihood expression in Equation 31 is considerably more efficient than the MVT form in Equation 23; it also notably improves the computational speed of MCMC sampling. This performance impact should be considered when evaluating the relative merits of these likelihood functions—especially when priors are added to the JPM.

2. Covariance Parameter Decomposition

The MVN and MVT likelihood functions in the previous section share an identical covariance structure for the predictands that will be described in the forecast application of Chapter IV. This is self-evident from the original description of the MVN log-likelihood in Equation 17, with univariate variance σ^2 replaced by a full covariance matrix as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}, \tag{34}$$

where σ_i describes the standard deviation of the corresponding predictand y_i and ρ_{ij} identifies the Pearson correlation coefficient (PCC) between predictands y_i and y_j. Equation 34 is appropriate for a model with three predictands—that is, $M = 3$. Equation 34 is similarly applicable to the MVT distribution because of the analytical relationships described by Equations 21 and 22; that is, samples from the MVT were generated by combining samples from the MVN with covariance Σ and samples taken from the univariate chi-squared distribution. In this way, Equation 21 explicitly invokes the same covariance matrix for a random variable that is distributed according to a multivariate extension of the classic t-distribution.

Nevertheless, the information contained within the covariance matrix has important constraints. To have physical significance, the standard deviations in Equation 34 must have nonnegative values. Moreover, the matrix inversion required by Equation 23 and Equation 31 must be permissible for all parameter samples generated by the random-walk Metropolis algorithm adapted by this dissertation. This is more precisely expressed by the requirement that covariance matrix Σ be positive semidefinite (or nonnegative definite) (Liu and Daniels 2006). More specifically, a positive semidefinite covariance matrix will satisfy

$$x^T \Sigma x \geq 0, \tag{35}$$

where x describes an arbitrary real-valued column vector. For the Gaussian proposal distribution considered by the MCMC methods in Chapter II, the Metropolis algorithm will generate real-valued proposals according to the location and scale hyperparameters of that distribution. However, these proposals are not guaranteed to satisfy Equation 35. As a result, we seek a collection of raw model parameters that are compatible with these constraints.

Cholesky (LDL) decomposition provides a solution. Krishnamoorthy and Menon (2013) note that the LDL decomposition of an arbitrary real-valued positive-definite matrix M may be written as

$$M = LDL^T, \tag{36}$$

where L is a lower triangular matrix and D is a diagonal matrix with positive elements. This decomposition becomes convenient when generating MCMC samples for covariance matrix parameters, because the lower triangular matrix L is only constrained to assume real values. A Gaussian proposal distribution in the Metropolis algorithm will produce both positive and negative samples for model parameters according to the details of the memoryless random walk structure of the evolving Markov chain. However, these samples can be repurposed according to Equation 36 to populate the elements of L and D directly with

$$\Sigma = LDL^T = \begin{bmatrix} 1 & 0 & 0 \\ \theta_{15} & 1 & 0 \\ \theta_{16} & \theta_{17} & 1 \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\theta_{12}} & 0 & 0 \\ 0 & \mathrm{e}^{\theta_{13}} & 0 \\ 0 & 0 & \mathrm{e}^{\theta_{14}} \end{bmatrix} \begin{bmatrix} 1 & \theta_{15} & \theta_{16} \\ 0 & 1 & \theta_{17} \\ 0 & 0 & 1 \end{bmatrix}, \tag{37}$$

where θ_i describes the associated raw model parameters. In simple terms, the MCMC sampler populates L and D directly; it does not generate proposals for the parameters in Equation 34. This matrix triple product is calculated for each step in the Markov chain so that Σ can be included in the cost functions of Equation 23 and Equation 31 as

$$\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \theta_{15} & 1 & 0 \\ \theta_{16} & \theta_{17} & 1 \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\theta_{12}} & 0 & 0 \\ 0 & \mathrm{e}^{\theta_{13}} & 0 \\ 0 & 0 & \mathrm{e}^{\theta_{14}} \end{bmatrix} \begin{bmatrix} 1 & \theta_{15} & \theta_{16} \\ 0 & 1 & \theta_{17} \\ 0 & 0 & 1 \end{bmatrix}. \tag{38}$$

After some linear algebra, these raw samples can be transformed into the common, intuitive elements of the covariance matrix—this includes the off-diagonal PCC terms.

3. Prior Distributions

The Bayesian prior distribution in Equation 14 arises from the decomposition of an arbitrary joint probability distribution, $p(\underline{\theta}, y)$, according to the chain rule of probability. In particular, we observe that the unnormalized Bayesian Posterior becomes

$$p(\underline{\theta} \mid y) \propto p(y \mid \underline{\theta}) p(\underline{\theta}), \tag{39}$$

where $p(y \mid \underline{\theta})$ is the conditional probability of observed data given the model parameters (i.e., the likelihood function), $p(\underline{\theta} \mid y)$ is the conditional probability of model parameters given the observed data (i.e., the posterior), and $p(\underline{\theta})$ is the marginal probability of the model parameters (i.e., the prior). We recognize the latter as the full vector representation of the Bayesian prior information, which is more appropriately written as joint probability distribution for all model parameters according to

$$p(\underline{\theta}) = p(\theta_0, \theta_1, \ldots, \theta_{P-1}), \tag{40}$$

where all quantities are as before for an NPS BEMOS model with P raw model parameters. Glickman and van Dyk (2007) note that it is common to assume that Equation 40 can be decomposed into a product of sequences by assuming that each model parameter has mutually independent prior information so that

$$p(\underline{\theta}) = p(\theta_0, \theta_1, \ldots, \theta_{P-1}) = \prod_{i=0}^{P-1} p(\theta_i). \tag{41}$$

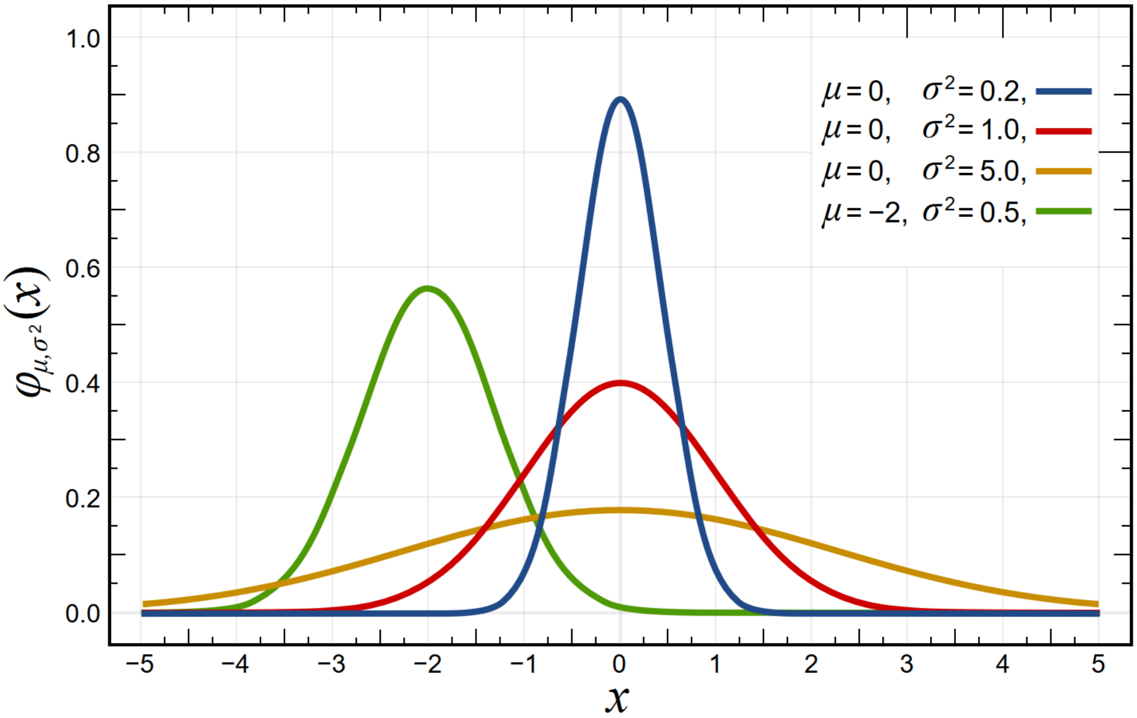


Figure 22. Gaussian prior distributions for random variable x according to various values of hyperparameters μ and σ^2. Source: Wikipedia 2017b. The ordinate describes the probability density assigned to the associated abscissa of parameter x.

In this way, each model parameter can have a distinct distribution assigned to it that describes the *a priori* belief in reasonable values for that parameter. This PDF is a function of the model parameter, but it also depends on the hyperparameters of the prior distribution. For a Gaussian prior (e.g., Figure 22), the appropriate distribution is

$$p(\theta) = f(\theta \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}, \tag{42}$$

where θ is the random variable associated with an arbitrary model parameter; μ and σ^2 are known as the hyperparameters of the prior and they control the location and scale of the modeler's *a priori* belief in this parameter. In simple terms, this PDF is a function of the model parameter because it is plotted over the random variable θ.

The hierarchical framework for beta regression shrinkage described by Kruschke (2014) treats these hyperparameters as model parameters themselves; they behave as ordinary parameters that must be inferred by the data. Moreover, some of the parameters at the primary model level (i.e., parameters on the bottom tier of Figure 15), which are designated as Level 1 parameters by the present work, are assigned identical hyperparameters consistent with *a priori* knowledge about the structural relationships between these parameters; in particular, we generally expect to find the beta regression coefficients in Equation 20 clustered around small values of β, which is consistent with the belief that the predictand(s) will have modest sensitivity with most model predictors (Kruschke 2014). However, this assumption and prior structure permits a few regression coefficients to have values displaced further away from the concentration of the aforementioned parameters, consistent with the distribution of probabilistic mass in Figure 22; that is, we would like to assume the betas are themselves normally distributed.

By giving the model freedom to infer the hyperparameters of the Gaussian prior, we allow the data to influence the analytical details of this hierarchical relationship. The present work has adopted the beta regression shrinkage framework of Kruschke (2014) and, in particular, assigned Gaussian prior distributions to all model parameters. For the forecast application described in Chapter IV, the NPS BEMOS model has $M = 3$ and $K - 1 = 3$ so that the regression coefficient and covariance matrices expand as before with

$$B = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \\ \beta_3 & \beta_4 & \beta_5 \\ \beta_6 & \beta_7 & \beta_8 \\ \beta_9 & \beta_{10} & \beta_{11} \end{bmatrix} \tag{43}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}. \tag{44}$$

To this end, the regression coefficients in Equation 43 are structured for the present work with common Gaussian hyperparameters—that is a single location and scale applicable to all beta coefficients (e.g., Figure 15b). This reflects an assumption about the distribution of these parameters in the model—one that seeks beta shrinkage to minimize “false alarms” (vis-à-vis Kruschke 2014) in the posterior inferences they inform.

Unfortunately, the covariance elements in Equation 44 are not strictly appropriate for Gaussian prior distributions; the latter are symmetric and permit both positive and negative values. In this way, we intuitively expect the diagonal elements of Equation 44—the variance parameters—to be restricted to nonnegative values and present a positive (right) skew. This reflects an assumption that variance elements should be positive and, importantly, the probability of finding larger σ_i^2 values should decrease monotonically. The Gamma distributions in Figure 15 are consistent with this description and are frequently selected for Bayesian priors on variance parameters (e.g., Kruschke 2014). However, Gelman et al. (2013) suggests a log transformation of these parameters to give them a more Gaussian appearance. Figure 21 described such benefits with a log transformation applied to raw data with positive (right) skew—as one might expect for the variance elements of Equation 44. To this end, the present work has adopted the log-sigma (i.e., the logarithm of standard deviation parameters) convention of Gelman et al. (2013) so that Gaussian priors on the raw model parameters are consistent with the Gamma priors suggested by Kruschke (2014) in Figure 15. These considerations provided an additional motivation for the LDL decomposition described in the previous section, and they are reflected by the exponential transformations of the diagonal elements of matrix D in Equation 37 and Equation 38. More specifically, we allow a Gaussian proposal distribution to generate real-valued samples for these raw variance parameters; the exponential transformation ensures they become positive in the subsequent LDL matrix triple product. Moreover, Gaussian priors on these raw variance parameters are entirely consistent with both our *a priori* assumptions and the practical constraints of the Metropolis sampler.

The parameters of the covariance matrix can also be modeled with hierarchical structure. As indicated for the beta regression coefficients, which were assumed to be identically distributed (i.e., Gaussian) with common location and scale hyperparameters, the diagonal and off-diagonal elements of the LDL framework can be given a similar treatment. In particular, the covariance elements of Σ associated with lower diagonal matrix L in Equation 37 are assumed to be identically distributed (i.e., Gaussian) with common location and scale hyperparameters. Moreover, the variance elements of Σ associated with diagonal matrix D in Equation 37 are assumed to be identically distributed (i.e., Gaussian) with common location and scale hyperparameters. In this way, “NPS” model instances are hierarchical Bayesian variants of the multivariate multiple linear regression model specified by Equation 43 and Equation 44 with 18 primary parameters and six hyperparameters. The latter describe non-fixed hyperparameters that must also be inferred by the NPS BEMOS model. The “MLE” model instances described in a previous section have the same multivariate multiple linear regression framework, and indeed deliver Bayesian uncertainty estimates with the aid of MCMC sampling, but they lack prior information and, as a consequence, contain only 18 parameters—no hyperparameters.

Finally, the present work would like to engage the benefits of the t-distribution priors in the original regression shrinkage procedure described by Kruschke (2014) (e.g., Figure 15). While the advantages of the t-distribution are well-known vis-à-vis robust regression, they can be problematic in practice—especially in the context of MCMC sampling. The full t location-scale distribution implied by Equation 21 has three hyperparameters: the location, scale, and shape (i.e., degrees of freedom). Unfortunately, permitting all three hypermeters as random variables in the model can cause computational difficulties that are difficult to resolve. Vehtari (2017) suggests a compromise procedure that heuristically fixes some hyperparameter values in the prior. The modeler can then iterate, according to step three in G13's Bayesian data analysis process (i.e., Chapter II, page 34, of the present work), to tune these fixed parameters or, if necessary, keep them fixed with appropriate values. To avoid such complications in the NPS BEMOS model, a Gaussian prior was selected to engage the regression coefficient shrinkage of K14. Having the same general characteristics of the t-distribution—that is, a greater concentration of probabilistic mass near the central tendency with a symmetric presentation—the Gaussian can still engage the core distributional assumption of identically distributed betas. Moreover, the hierarchical nature of the hyperparameter inference is similarly retained; these parameters are included in the Bayesian cost function of the NPS BEMOS model. "By letting the regression coefficients be mutually informed by other predictors, and not only by the data of the single predictor each multiplies," Kruschke (2014) notes that the "the coefficients are less likely to be spuriously distorted by a rogue sample."

4. JPM Diagrams

Kruschke (2014) introduced "hierarchical diagrams" to map the structure Bayesian probability models and better visualize the relationships between various model elements. Figure 15 is an example of such a diagram, and it describes the primary elements of a hierarchical Bayesian linear regression model (i.e., Figure 15b) with a normal likelihood function and t-distribution priors on the beta regression coefficients; Gaussian priors are placed on most of the remaining parameters and hyperparameters. The present work describes these visualization aids as JPM diagrams to avoid confusion with hierarchical modeling terminology. In this way, a non-hierarchical model can also have a JPM diagram (e.g., Figure 15a) that describes the core elements of Bayesian cost functions—that is, the analytical machinery that ultimately generates our posterior parameter beliefs.

The MVN and MVT likelihood functions considered by this dissertation will produce four distinct JPM diagrams: two for the "MLE" model instances with noninformative priors and two for the “NPS” instances with hierarchical prior information. The former is indicated by Figure 23, which shows the stochastic relationship between observable predictand vector $\underline{y}$ and MVN (left; panel a) and MVT (right; panel b) likelihood functions. Raw model parameters (i.e., the “thetas”) are indicated according to their roles in the model (e.g., as regression coefficients or covariance parameters). The absence of prior information in these diagrams is consistent with $p(\underline{\theta}) \propto 1$ and, indeed, the basic motivations for the classical MLE approach to parameter estimation. As previously noted by Kruschke (2014), classical MLE merely reports a single-valued parameter estimate for each model—the one that makes the fixed data sample most probable. The Bayesian version of MLE has the same cost function, but it seeks a full posterior PDF. By comparison, the OLS/MLE modeler must append uncertainty information to their point estimates in a separate procedure. The NPS BEMOS model also fixes the degrees-of-freedom parameter ν in Figure 23 based on Vehtari (2017). This was done to facilitate MCMC convergence; the presence of all three MVT parameters in the NPS BEMOS model caused erratic Markov chain behavior. The present work chose $\nu = 5$ to maximize the probabilistic mass in the MVT tails while also ensuring the variance, skewness, and kurtosis were all defined (i.e., $\nu > 4$).

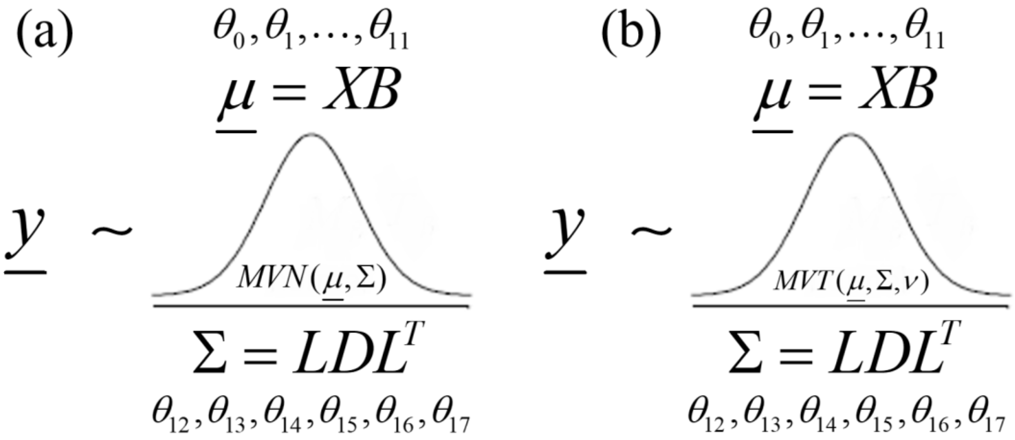

(a)
$\theta_0, \theta_1, \ldots, \theta_{11}$
$\underline{\mu} = XB$
$\underline{y} \sim$
$MVN(\underline{\mu}, \Sigma)$
$\Sigma = LDL^T$
$\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}$
(b)
$\theta_0, \theta_1, \ldots, \theta_{11}$
$\underline{\mu} = XB$
$\underline{y} \sim$
$MVT(\underline{\mu}, \Sigma, \nu)$
$\Sigma = LDL^T$
$\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}$

Figure 23. JPM diagrams for noninformative priors paired with MVN (left; panel a) and MVT (right; panel b) likelihood functions. Adapted from Kruschke (2014). The unnormalized Bayesian posteriors associated with these models comprise only the indicated likelihood functions.

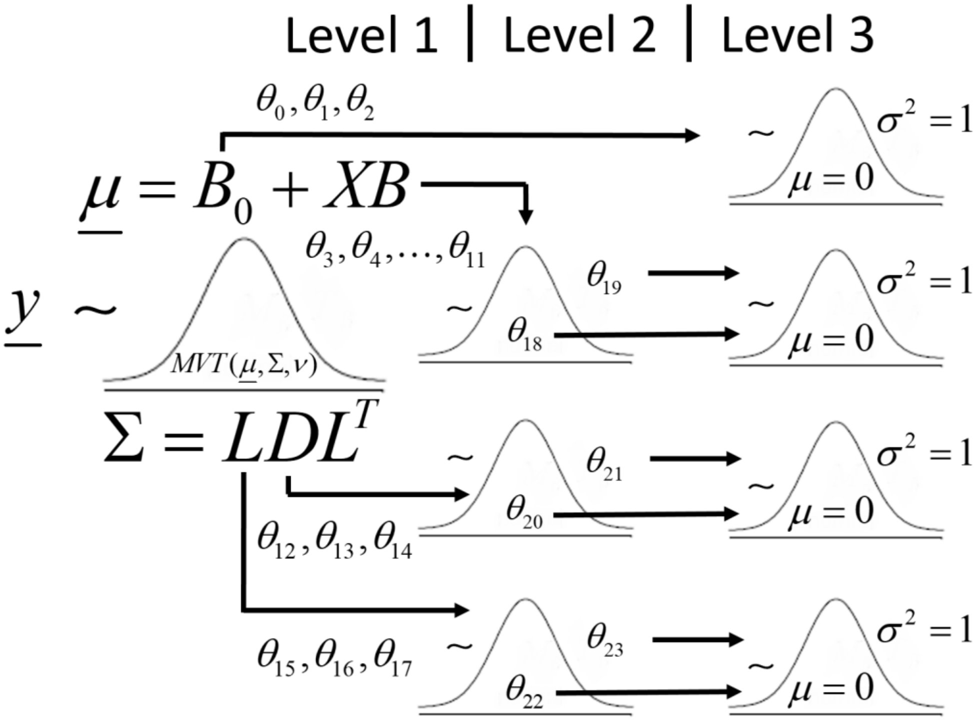

Level 1 | Level 2 | Level 3
$\theta_0, \theta_1, \theta_2$
$\underline{\mu} = B_0 + XB$
$\theta_3, \theta_4, \ldots, \theta_{11}$
$\underline{y} \sim$
$MVT(\underline{\mu}, \Sigma, \nu)$
$\Sigma = LDL^T$
$\theta_{12}, \theta_{13}, \theta_{14}$
$\theta_{15}, \theta_{16}, \theta_{17}$
$\sim$
θ_{18}
θ_{19}
θ_{20}
θ_{21}
θ_{22}
θ_{23}
$\sigma^2 = 1$
$\mu = 0$

Figure 24. JPM diagram for a hierarchical version of the NPS BEMOS model with MVN likelihood function. Adapted from Kruschke (2014). The likelihood function is in Level 1, the hierarchical hyperparameters are in Level 2, and the hyperprior distributions are in Level 3.

For “NPS” model instances with MVN likelihood functions and hierarchical Bayesian priors, the JPM diagrams are consistent with Figure 24. The stochastic relationship between observable predictand vector $\underline{y}$ and the MVN likelihood function is indicated in Level 1. The MVN location and scale parameters (i.e., $\underline{\mu}$ and Σ) each contain model parameters with prior distributions placed over them. The beta matrix containing all regression parameters has been decomposed into intercept parameters (i.e., B_0) and regression coefficients (i.e., B) to allow mutual, non-fixed Gaussian hyperparameters θ_{18} and θ_{19} to be placed on regression parameters $\theta_3, \theta_4, \ldots, \theta_{11}$. Similarly, the raw parameters that feed covariance matrix Σ via Equation 38 are decomposed into lower triangular and diagonal groups according to their analytical roles in Σ. Each group has Gaussian priors placed so that hyperparameters θ_{22} and θ_{23} are coupled with $\theta_{15}, \theta_{16}, \theta_{17}$ in lower triangular matrix L and hyperparameters θ_{20} and θ_{21} are placed on raw parameters in diagonal matrix D. As a result, the beta regression coefficients $\theta_3, \theta_4, \ldots, \theta_{11}$ are assumed to be normally distributed with mean θ_{18} and standard deviation $\exp\left(\theta_{19}\right)$; raw covariance parameters are assumed to be normally distributed with mean θ_{22} and standard deviation $\exp\left(\theta_{23}\right)$; raw variance parameters are assumed to be normally distributed with mean θ_{20} and standard deviation $\exp\left(\theta_{21}\right)$. All hierarchical hyperparameters (i.e., $\theta_{18}, \theta_{19}, \ldots, \theta_{23}$) are located in Level 2 of the model. Hyperparameters $\theta_{19}, \theta_{21}, \theta_{23}$ have been exponentiated in the aforementioned descriptions to ensure their values are valid (i.e., positive) as scale parameters; their raw values generated by a Gaussian proposal distribution in the Metropolis algorithm can be both positive and negative. This approach is consistent with the log-sigma convention described by Gelman et al. (2013) and is analogous to placing Gamma priors on variance parameters vis-à-vis Kruschke (2014) and Figure 15. Finally, the beta intercepts (i.e., $\theta_0, \theta_1, \theta_2$) and all hierarchical hyperparameters are assigned to a standard normal prior distribution in Level 3; the latter (i.e., priors for hyperparameters) are also known as hyperpriors. A similar structure is assigned to “NPS” model instances with MVT likelihood functions—only the likelihood function changes. Computational difficulties associated with MCMC convergence similarly compelled the present work to fix the degrees-of-freedom parameter as before. To this end, $\nu = 5$ was chosen to maximize the probabilistic mass found in the MVT tails to ensure the variance, skewness, and kurtosis of the likelihood function were all suitably defined (i.e., $\nu > 4$).

Both likelihood functions are non-conjugate with the normal prior distributions indicated in Figure 24. Gelman et al. (2013) notes that the MVN likelihood requires a normal-inverse-Wishart distribution for conjugacy, which places additional constraints on the selection of common-sense prior beliefs. The lack of an exponential term in the MVT PDF means that the associated log-likelihood is similarly intractable when paired with Gaussian prior information. Moreover, the hierarchical structure of both models complicates the aforementioned conjugacy considerations with no clear indications for analytical solutions in the statistical literature (e.g., Gilks et al. 1996, Gelman et al. 2013, or Kruschke 2014). While conjugacy is essential for model interpretation and inference completion by purely analytical means, the MCMC methods described in Chapter II provide a compelling motivation for the sophisticated JPM in Figure 24—especially when compared with the univariate framework considered by Gneiting et al. (2005) and Richter (2012).

5. Posterior Predictive Distributions

Chapter II introduced the concept of Bayesian PPDs. While the methods described heretofore focus on the specification of the probability model—including the parametric form of the likelihood function and the mathematical character of *a priori* assumptions—they have not detailed a procedure to convert posterior parameter beliefs about model parameters into prognostic statements. Using a modified form of the notation in Gelman et al. (2013), the latter is engaged by Equation 29, which describes the probability of unobserved observable data in random variable z conditioned on the observed observable data in random variable y as

$$p(z \mid y) = \sum_{\theta} p(z \mid \underline{\theta}) p(\underline{\theta} \mid y), \tag{45}$$

where $p(z \mid \underline{\theta})$ is simply the original likelihood function of the model and $p(\underline{\theta} \mid y)$ represents the posterior parameter beliefs obtained from our previous Bayesian data analysis. By design, the former is a parametric distribution; we can draw an arbitrary number of samples from it once we know the values of its distributional parameters. In this way, the discrete sum in Equation 45 is important to our probabilistic forecasts; it marginalizes out dependence on the posterior inferences. In other words, the Bayesian PPD is an “average of conditional predictions over the posterior distribution of θ“ (Gelman et al. 2013).

Recall again the behavior of a hypothetical Markov chain that evolves according to the rules in Equation 27. At each step in the chain, the proposal distribution draws a multiparameter sample based on its current location in the phase space of the model. In this way, each multiparameter sample is a vector of estimates—guesses, really, as the proposal distribution is doing this randomly—for the true values of the model parameters. Equation 20 describes the multiple linear regression procedure that combines betas with ensemble predictors to produce updated estimates of the location, scale, etc., of $p(z \mid \underline{\theta})$. Thus, each step in the Markov chain represents a new set of mathematical instructions that map ensemble predictors to the updated location, scale, etc., of $p(z \mid \underline{\theta})$. Equation 45 simply weights these mathematical instructions according to the probabilistic mass in $p(\underline{\theta} \mid y)$. Each sample drawn from $p(z \mid \underline{\theta})$ is therefore paired in a 1:1 ratio with a sample from $p(\underline{\theta} \mid y)$; the distribution of samples generated in this manner merely communicates a weighted average of the former conditioned on the latter.

The Bayesian PPDs constructed for the present work require random sampling from both MVN and MVT distributions—that is, the parametric distributions selected for $p(z \mid \underline{\theta})$. The procedure for sampling from the MVN distribution is straightforward; most programming languages (e.g., MATLAB, Python, etc.) have numerical routines that will complete the task directly. However, Equation 21 shows the comparatively sophisticated structure of the MVT distribution, which is actually a combination of the former with the univariate chi-squared distribution. As a result, MVT samples are constructed from sampling both of these PDFs—that is, the MVN and $\chi^2(\nu)$ distributions—separately and combining the results according to Equation 22. In the present work, this is pursued without recursive looping by pre-allocating a MxN array of $MVN(\underline{0}, I)$ samples and a $Nx1$ column vector of $\chi^2(\nu)$ samples with Python NumPy functions; N here represents the desired number of samples. Contemporary ensemble predictors are then standardized according to Equation 30 with μ_s and s_x measured from the training data. The standardized predictors are then combined with longitudinal slices of the posterior parameter samples according to

$$\mu = XB, \tag{46}$$

where X is $1xKxN$ matrix of predictors, B is a $KxMxN$ matrix of posterior parameter samples, and μ is $1xMxN$ matrix of location samples. The matrix multiplication is executed on each vertical slice of the aforementioned elements with a Python NumPy function called "matmul." Parameter samples appropriate for the elements of the covariance matrix are first reshaped according to Equation 38 and then combined with the $MVN(\underline{0}, I)$ samples shifted according to Equation 22 as

$$z = XB + \Sigma MVN(\underline{0}, I)\left[\frac{\nu}{\chi^2(\nu)}\right]^{1/2}. \tag{47}$$

The data transformations described at the beginning of this chapter must be taken into consideration before a final, complete PPD is produced. In this way, vector samples for predictands in Equation 47 are in the standardized, log-transformed coordinate space of the NPS BEMOS model. The present method applies a log transformation before standardizing each dimension of the training data separately. Vector samples obtained from Equation 45 and Equation 47 must have a reverse transformations applied to make predictions or other inferences. Figure 25 demonstrates the impact of these transformations on a joint PPD for minimum diurnal surface temperature and maximum diurnal surface wind speed. While both variables have been fit with symmetric MVN likelihood functions according to the JPM in Figure 24, the marginal PPD at top has an asymmetric presentation with positive skew; the associated joint distribution in the center of Figure 25 has a corresponding PCC that is non-zero. This asymmetric PPD is a consequence of the aforementioned data transformations and provides additional predictive fidelity when fitting parametric distributions to observable data with the NR/EMOS technique.

In this way, predictions formed from Bayesian data analysis are produced according to the modeling process indicated in Figure 26. Training data must first be selected from a suitable period in which predictions for sensible weather variables extracted from a parent model may be compared with corresponding observations (i.e., *observed* observable data). Data transformations suitable to the sensible weather variables may then be applied. The underlying shape and character of the data, as well as the details of the forecast application, can then inform the modeler’s selection of the likelihood function and prior distributions. The former describes the stochastic generative process best suited to the anticipated shape and character of *unobserved* observable data; the latter represents statistical belief from any source of information—empirical or otherwise—that might constrain or weight likely values of model parameters in Level 1 of Figure 24. These elements form the unnormalized Bayesian posterior (i.e., the JPM).

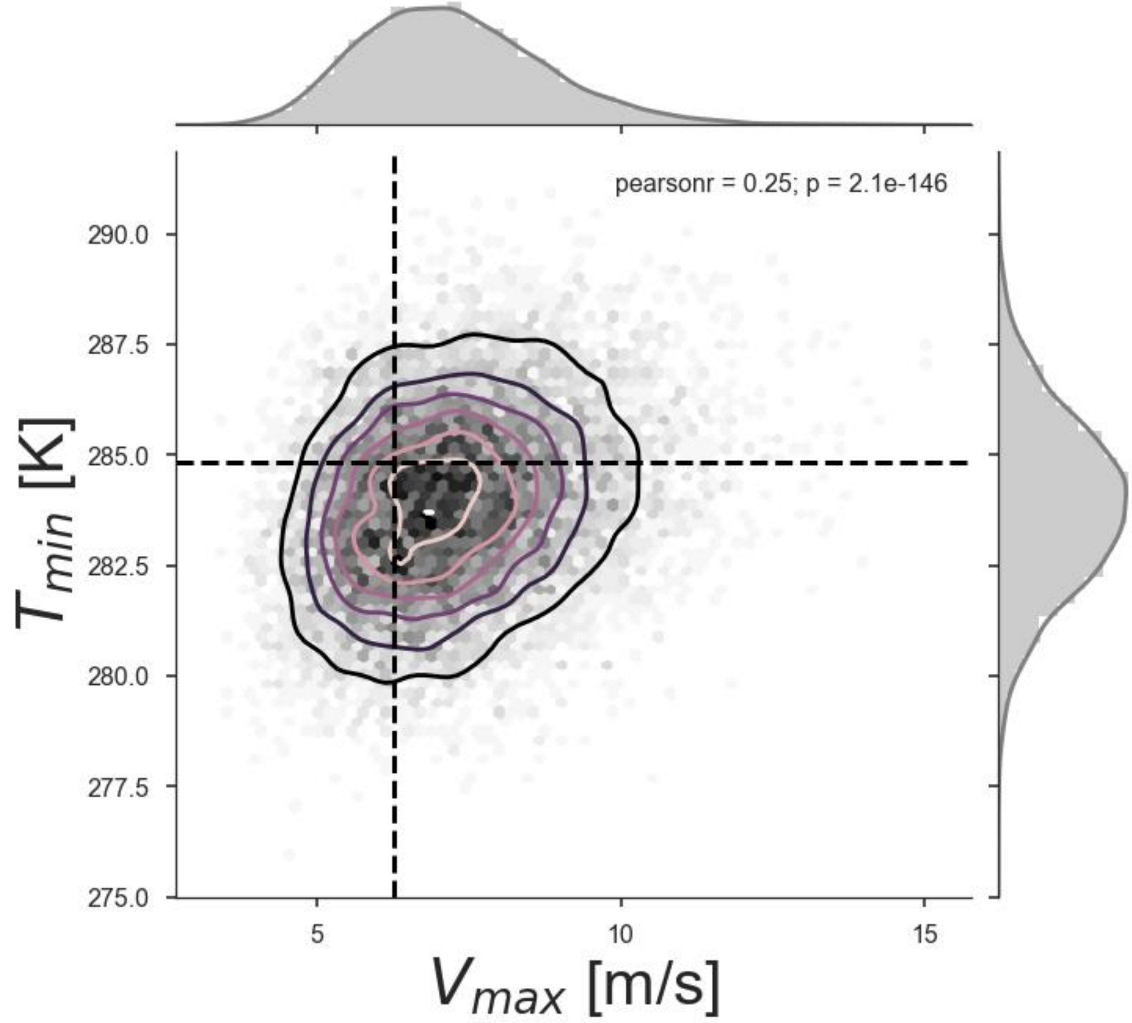


Figure 25. Joint posterior predictive distribution (PPD) with 10000 discrete samples for maximum diurnal surface wind speed (abscissa) and minimum diurnal surface temperature (ordinate). The contours superimposed on the primary plot represent isopleths of probability density (lighter is higher); the underlying hex-plot indicates the relative density of samples in local area (darker is higher). Marginal PPDs for each forecast variable are located on the appropriate border with KDE superimposed on histograms of the underlying data. The true values observed for this day are indicated with dashed black lines. This PPD corresponds to a forecast valid May 31, 2017 for Monterey, CA.

Once the JPM has been selected, the total number of raw model parameters can be determined from all model components—that is, from the primary linear regression and covariance framework imbedded within the likelihood function and, if any, the number of hierarchical hypermeters located in prior distributions on additional model levels (e.g., Level 2 in Figure 24). The Metropolis sampler must then be tuned to specify the total number of Markov chain transitions (i.e., MCMC chain length), the burn-in threshold (i.e., the transition number after which the chain is assumed to have reached equilibrium), and the number of posterior samples retained after burn-in (i.e., thinning). Markov chain length and burn-in are frequently determined in a heuristic manner to ensure model convergence—topics described in more detail in the next section. Finally, the model is initialized and posterior samples for each model parameter are collected and combined with random samples from the likelihood function according to Equation 45. Reverse transformations are then applied to form the final PPDs for the desired predictands.

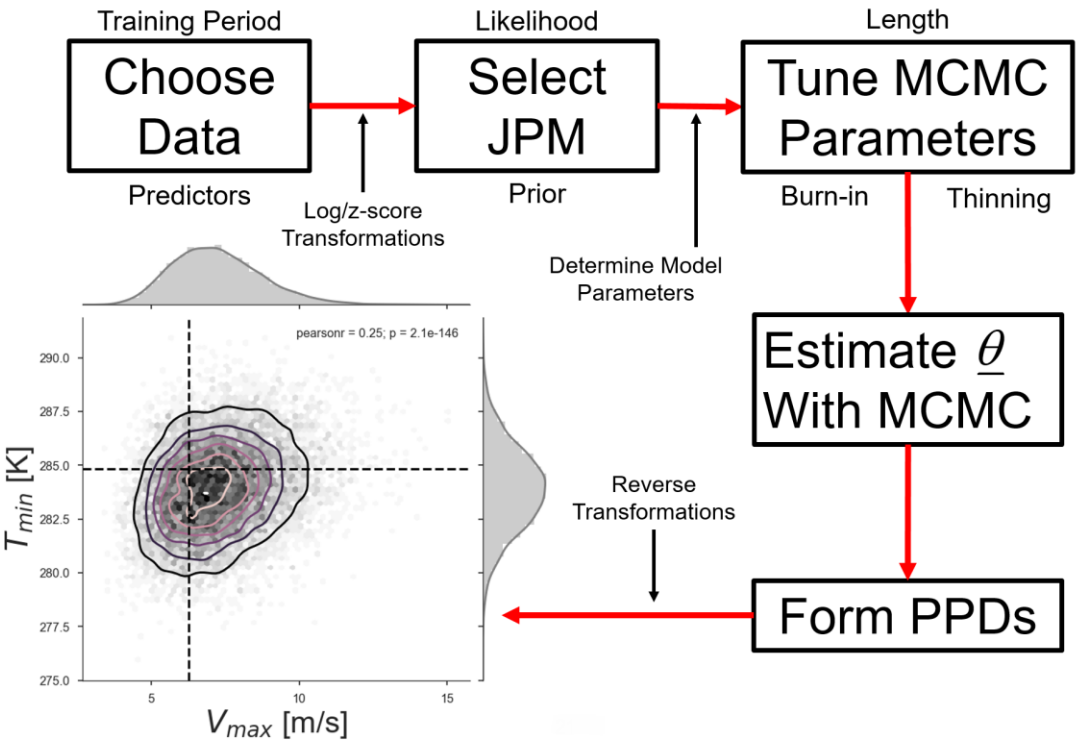


Figure 26. A pedagogical representation of the NPS BEMOS modeling process. The order of major actions is represented with red arrows. Minor actions are indicated outside of primary actions with black arrows.

D. METROPOLIS SAMPLER

With the theoretical foundation of the Metropolis algorithm established in Chapter II, this section describes the details of the method adapted for the NPS BEMOS model. This includes the adaptive components of the Metropolis sampler, multi-dimensional Markov chains initialization, details associated with the multiparameter updating scheme, the computational structure of a Bayesian log-odds ratio as a MCMC target density, and Markov chain autocorrelation and thinning vis-à-vis MCMC burn-in.

1. Adaptive Sampling

The Metropolis algorithm requires the selection of a symmetric proposal distribution. This PDF determines the stochastic manner by which proposals are made and, if accepted, appended to the evolving Markov chain. In this way, Equation 27 determines only if proposals are accepted or rejected; it has no influence on the locations proposed by the sampler. The latter is engaged by the proposal distribution. However, standard techniques render it oblivious to the frequency with which proposals are accepted; it merely uses the current location of the Markov chain, and the width of its distribution (i.e., jumping kernel variance), to draw samples from itself. Wiecki (2015) notes that this form of sampling requires iterative tuning to ensure the phase space of the model is adequately and efficiently explored. Inappropriate proposal widths can unnecessarily delay MCMC convergence or, in the case of the unstandardized parameters described earlier in this Chapter (i.e., Figure 20b), produce Markov chains that are poorly suited to the probability structure of the joint posteriors.

Wiecki (2015) demonstrated the impact of proposal width on Metropolis sampling in Figure 27. A Gaussian jumping kernel has been selected with two step sizes: “large” and “small” proposal widths. In the former, a large number of samples are clustered in a high-probability region near the primary mode of the distribution (blue; left). However, the impact of small kernel variance is evident in the smaller peak (blue; right); the chain is effectively stuck in a narrow region of the model’s phase space, because the sampler is unable to generate proposals that significantly differ from the current location. In this way, the proposal distribution is unable to efficiently sample the model space when the proposal width is too small; the result is an implausible multimodal distribution (blue). A larger step size (green) can generate proposals that are more displaced from the current location of the Markov chain. However, some of the fine structure of the analytical posterior distribution (i.e., Figure 19) is lost. In this way, the proposal distribution must be iteratively tuned by the modeler to ensure posterior samples adequately represent the desired inferences. The impact of proposal width on Markov chain sample autocorrelation is indicated in Figure 28. The autocorrelation function measures the “clumpiness” of the posterior samples retained from MCMC sampling and can be an important convergence diagnostic. The determination of the latter frequently requires low sample autocorrelation consistent with the “large” and “medium” step sizes in Figure 28.

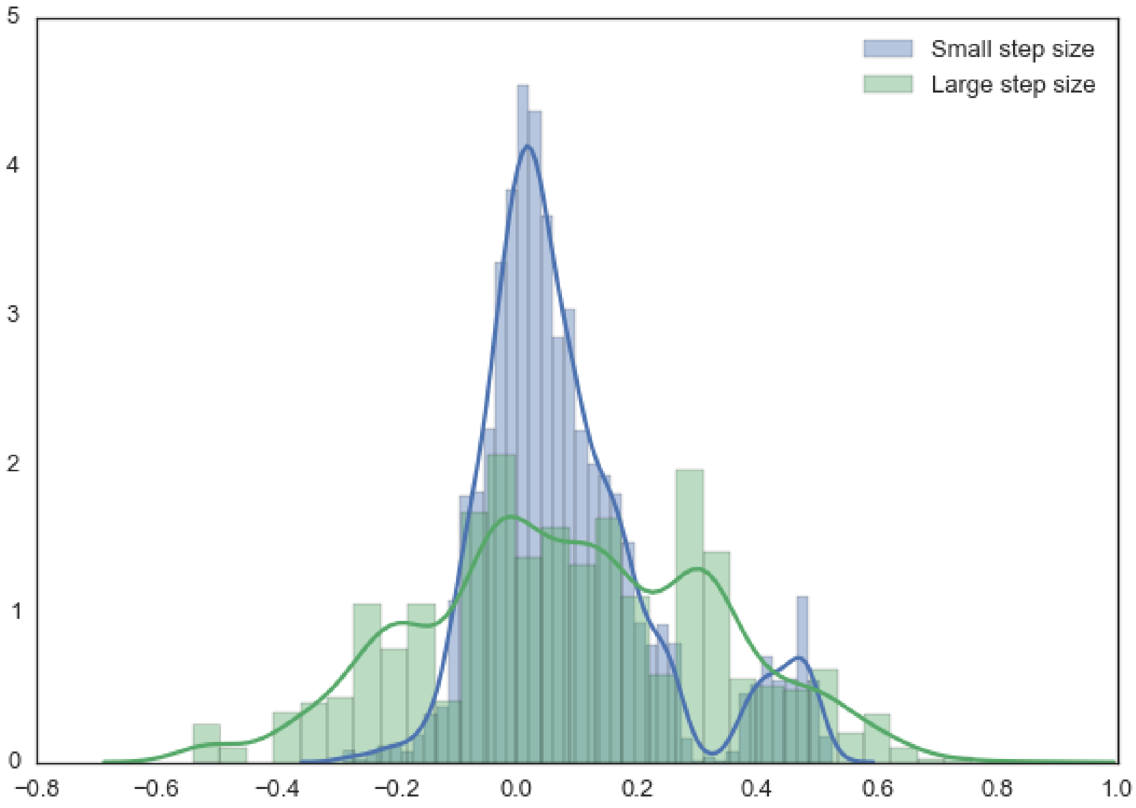


Figure 27. Posterior parameter samples obtained from a Bayesian parameter estimation scheme completed with the Metropolis algorithm. Source: Wiecki (2015). The variance of the Gaussian proposal distribution is indicated with binary step sizes. Both representations would be considered undesirable in MCMC sampling relative to the true solution Figure 19.

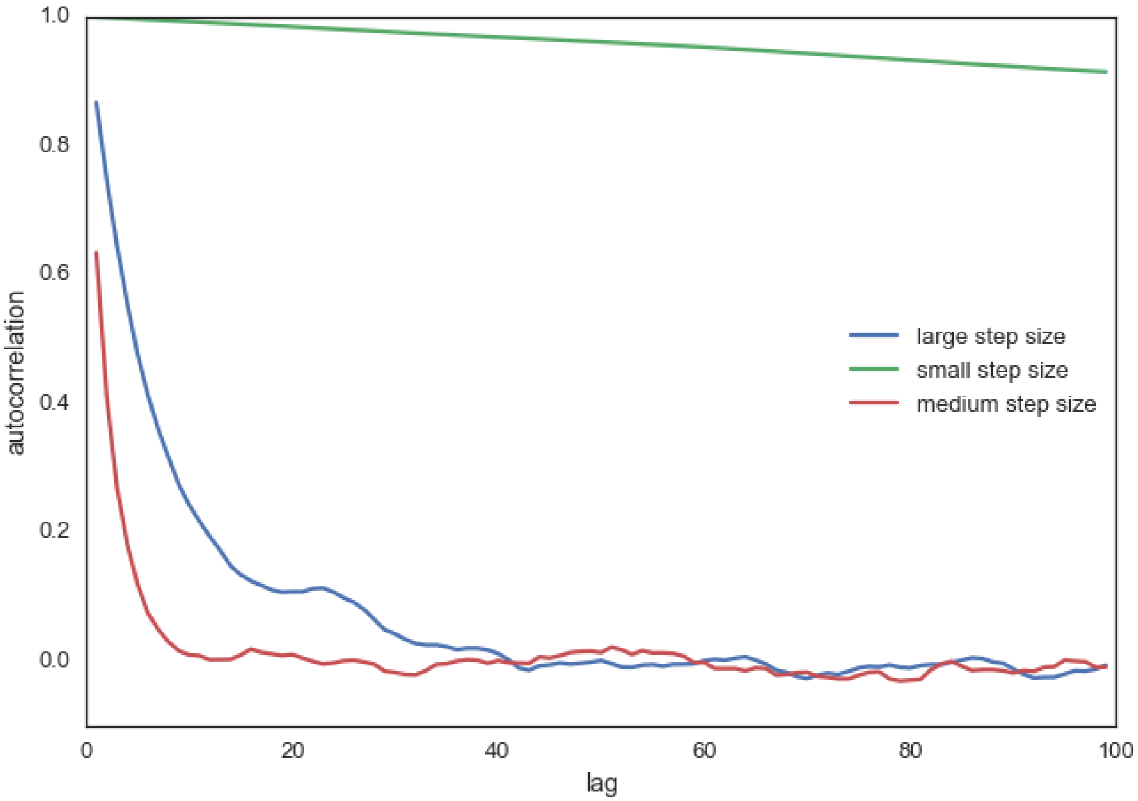


Figure 28. Visualization of MCMC sample autocorrelation obtained from a pedagogical Bayesian parameter estimation scheme. Source: Wiecki (2015). Each curve represents a distinct width of a proposal distribution used to generate samples from the Gaussian target distribution in Figure 19. Low(er) autocorrelation is desirable in MCMC sampling.

While manual tuning might be practical for simple, pedagogical Bayesian inference schemes, complicated JPMs (e.g., hierarchical) require more efficient Markov chain sampling. To this end, Roberts et al. (1997) proved that an “asymptotically optimal acceptance rate is 0.234 under quite general conditions” in their paper “Weak Convergence and Optimal Scaling of Random Walk Metropolis algorithms.” This classic and important result suggests that a modeler should “tune the proposal variance so that the average acceptance rate is roughly $^1/_4$“ (Roberts et al. 1997). Their “useful heuristic” is critical for adaptive sampling techniques, which seek programmatic tuning of proposal distribution variance. To this end, the present work has developed a modified form of the Metropolis algorithm—one that tunes the proposal width programmatically.

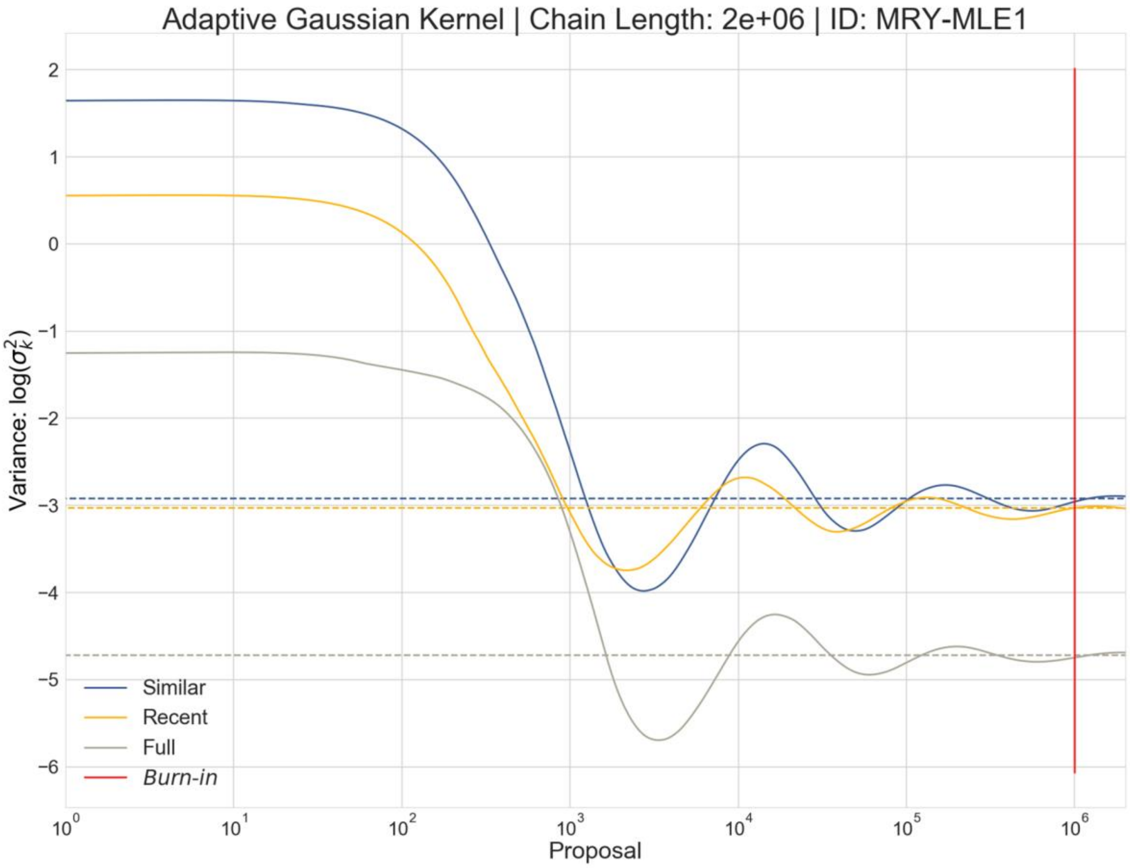


Figure 29. Semi-log visualization of NPS BEMOS proposal width variability (ordinate) as a function of proposal number (abscissa). The red vertical line indicates model "burn-in," which describes a pre-defined boundary between valid and invalid posterior parameter samples. The horizontal dashed lines indicate average proposal distribution variances found for each training period.

For the NPS BEMOS model, the moving average of the proposal acceptance rate is tracked at each step in the evolving sequence of Markov chain transitions and, perhaps more importantly, compared with the optimal value provided by Roberts et al. (1997). The proposal of the next step in the Markov chain is computed iteratively according to

$$\ln\left[\sigma_{i+1}\right] = \ln\left[\sigma_i\right] + f(a_i), \tag{48}$$

where σ_i is the standard deviation of the proposal distribution at the current step, σ_{i+1} is the standard deviation of the proposal distribution to be used in the next step, a_i is the current acceptance rate, and $f(a_i)$ is a custom model function that depends on the current acceptance rate and the predetermined length of the full Markov chain. For the present work, $f(a_i)$ was heuristically specified as

$$f(a_i) = \frac{\ln(j) - \ln(i) - \ln(a_T)}{(0.25N - 1000)(i/N) + 1000} \tag{49}$$

where j represents the number of accepted proposals, i represents the current proposal number, N is the predetermined full length of the Markov chain, and $a_T = 0.234$ was developed from theory. The numerator in Equation 49 is formed by the natural logarithm of the ratio of the current acceptance rate, a_i , and the desired theoretical rate a_T. The denominator represents a scaling factor that permits larger adaptive corrections with earlier Markov transitions. This is designed to ensure adaptive tuning is suitably damped.

Figure 29 and Figure 30 demonstrate the impact of this adaptive approach to proposal width tuning on the log-variance of a customized MVN proposal distribution. The latter is analytically described as

$$\underline{\theta}_{i+1} \sim MVN(\underline{\theta}_i, \sigma_k^2(i)I), \tag{50}$$

where $\underline{\theta}_i$ is the multiparameter location of the Markov chain at sequence number i, $\underline{\theta}_{i+1}$ is the multiparameter location of the Markov chain at next step in the chain (i.e., the proposed multiparameter location), $\sigma_k^2(i)$ is the variance of the jumping kernel at sequence number i, and I is an identity matrix that matches the dimensionality of $\underline{\theta}_i$ and $\underline{\theta}_{i+1}$. Over three distinct training periods, which are represented by the blue, gold, and grey curves, the JPM structure in Figure 23a was paired with the NPS BEMOS sampler to complete the inference for 18 model parameters. A time series of the proposal widths developed by Equation 48 and Equation 49 is indicated with a semi-log plot of σ_k^2 and a as a function of the proposal number in Figure 29. The longest training period (i.e., "Full") required a distinct proposal width from the remaining two training periods; however, all were found to converge to the optimal acceptance rate in Figure 30.

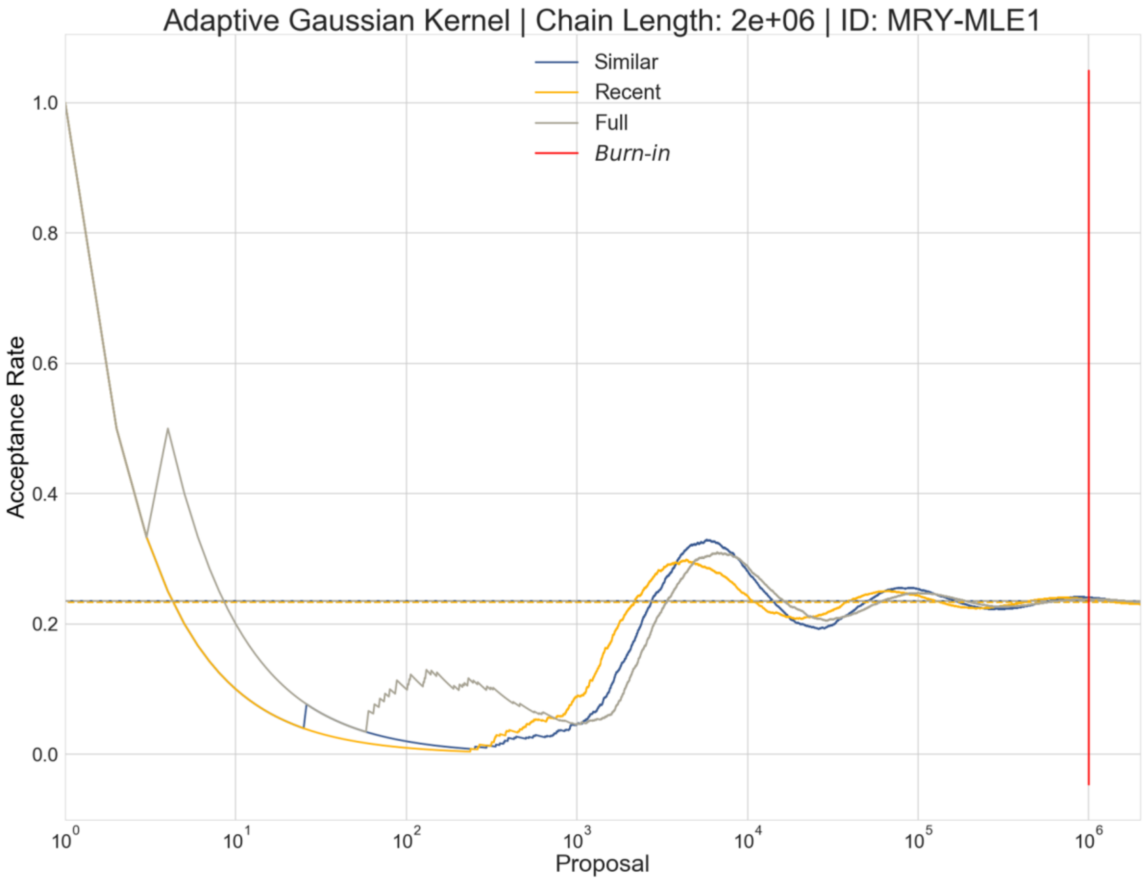


Figure 30. Semi-log visualization of NPS BEMOS proposal acceptance rate (ordinate) as a function of proposal number (abscissa). The red vertical line indicates model "burn-in," which describes a pre-defined boundary between valid and invalid posterior parameter samples. The horizontal dashed lines indicate the average acceptance rate found for each training period; the indicated values are consistent with the optimal theoretical value.

These figures are also notable for demonstrating the damped oscillator behavior of the heuristic approach developed by the present work. This follows from the form of Equation 48: the magnitude of the correction applied at each step is proportional to deviation between the current and optimal acceptance rates and the proposal number—similar to the restoring spring-force applied in a damped system. Unlike their analogues from classical Physics, however, no analytical solutions are known to exist for an *a priori*, optimal solution to this problem for arbitrary Bayesian probability models; the results depend on the details of the JPM and the training data. While generalized methods for adaptive sampling are common in data science (e.g., Gelman et al. 2013), the details of the approached developed by the present work were pursued somewhat independently and specialized to the scientific application presented in Chapter IV. Moreover, they provide a customized, heuristic solution to the problem of optimal proposal width articulated by Wiecki (2015).

2. Convergence and Thinning

Chapter II noted that MCMC convergence was guaranteed if the evolving chain was found to be irreducible, aperiodic, and non-transient (Gilks et al. 1996 and Gelman et al. 2013). In this way, the utility of the Metropolis algorithm is revealed when our posterior parameter inferences are structured so that the equilibrium distribution of the Markov chain matches the unnormalized Bayesian posterior distribution—that is, our "target" distribution—according to Equation 39. Conveniently, the adaptive sampling procedure described in the previous section can help diagnose model convergence. In conjunction with the aforementioned autocorrelation metric, convergence can be informally verified when a time series of these quantities reaches a plausible equilibrium (e.g., the quasi-steady behavior of the time series data in Figure 29 and Figure 30). While model convergence isn't guaranteed from the former, the quasi-steady behavior of σ_k^2 and a is consistent with model convergence. To be sure, one must also examine the behavior of the Markov chain as it evolves in the phase space of the model. Data scientists commonly use these so-called trace diagrams (e.g., Stansbury 2012a and Wiecki 2015) to examine the parameter proposals generated by a Metropolis sampler as a function of the Markov chain sequence (i.e., proposal) number.

Figure 31 is an example of such a trace diagram; it extends the inference considered by Figure 29 and 30 to show the actual location of the associated multiparameter Markov chain (i.e., $\underline{\theta}_i$) along a single dimension (e.g., θ_5) of the full 18-parameter vector for each transition (i.e., proposal) in the chain. Moreover, three training periods are indicated (i.e., the blue, gold, and grey curves); each corresponds to a one-dimensional slice of an 18-dimensional Markov chain associated with the indicated period of training data (i.e., "Full," "Similar," and "Recent"). All three figures (i.e., Figure 29, 30, and 31) correspond to a multiparameter (i.e., 18-dimensional) Markov chain realized with two million state transitions (i.e., proposals). To this end, Figure 32 further considers a particular "view" of this Markov process that is more relevant to the ultimate posterior parameter inferences we seek. If one were to stand at the end (far right) of the "Similar" (blue) chain in Figure 31 with a longitudinal perspective, so that the vertical variability observed in parameter location is aligned perpendicular to one's line of sight, then Figure 32 provides a visualization of random samples taken from the rear half, or "burn-in" region, of this chain. This plot is analogous to Figure 19 and represents a "machine-learned" solution to a Bayesian parameter estimation scheme formed by the JPM in Figure 23a—that is, posterior belief $p(\underline{\theta} \mid y)$ in model parameter θ_5 conditioned on model structure, prior information, and training data. Kruschke (2011) and Gelman et al. (2013) discuss data "thinning" so that posterior samples are drawn only from the burn-in region. MCMC proposals generated after burn-in are described as "warmed" samples by the present work. If the Markov chain has been constructed with sufficient length so that equilibrium is achieved prior to burn-in, then the warmed samples may be considered IID draws from the target distribution—that is, the unnormalized Bayesian posterior. In this way, sample autocorrelation can be reduced with longer Markov chains and thinning from the burn-in region. Figure 32 has been constructed in this manner; random samples taken from the warmed portion of the chain form an empirical distribution that corresponds to the Bayesian posterior. The posterior is visualized with a grey histogram and, optionally, KDE (continuous black curve) of the underlying distribution.

The steady-state behavior of the moving average of each one-dimensional slice of the chain in Figure 31 is desirable and highly consistent with model convergence. In this way, the present work assesses model convergence heuristically with a combination of trace diagrams for each parameter dimension (e.g., Figure 31), autocorrelation metrics (e.g., Figure 28), diagnostic metrics associated with the adaptive sampler (e.g., Figure 29 and Figure 30), and, indeed, plausible posterior parameter distributions (e.g., Figure 32).

For comparison, the implausible posterior parameter beliefs in Figure 27 suggest that model convergence may not have been obtained. In this scenario, additional convergence diagnostics would be required. To ensure convergence, the model may simply require more transitions so that the total chain length and burn-in region are suitably extended. While more formal methods of assessing model convergence are available (e.g., Gelman and Rubin 1992; Brooks and Gelman, 1998), they are beyond the scope of the present work. The eager reader is encouraged to consider Gilks et al. (1996), Gelman et al. (2013), and Kruschke (2014) for more information.

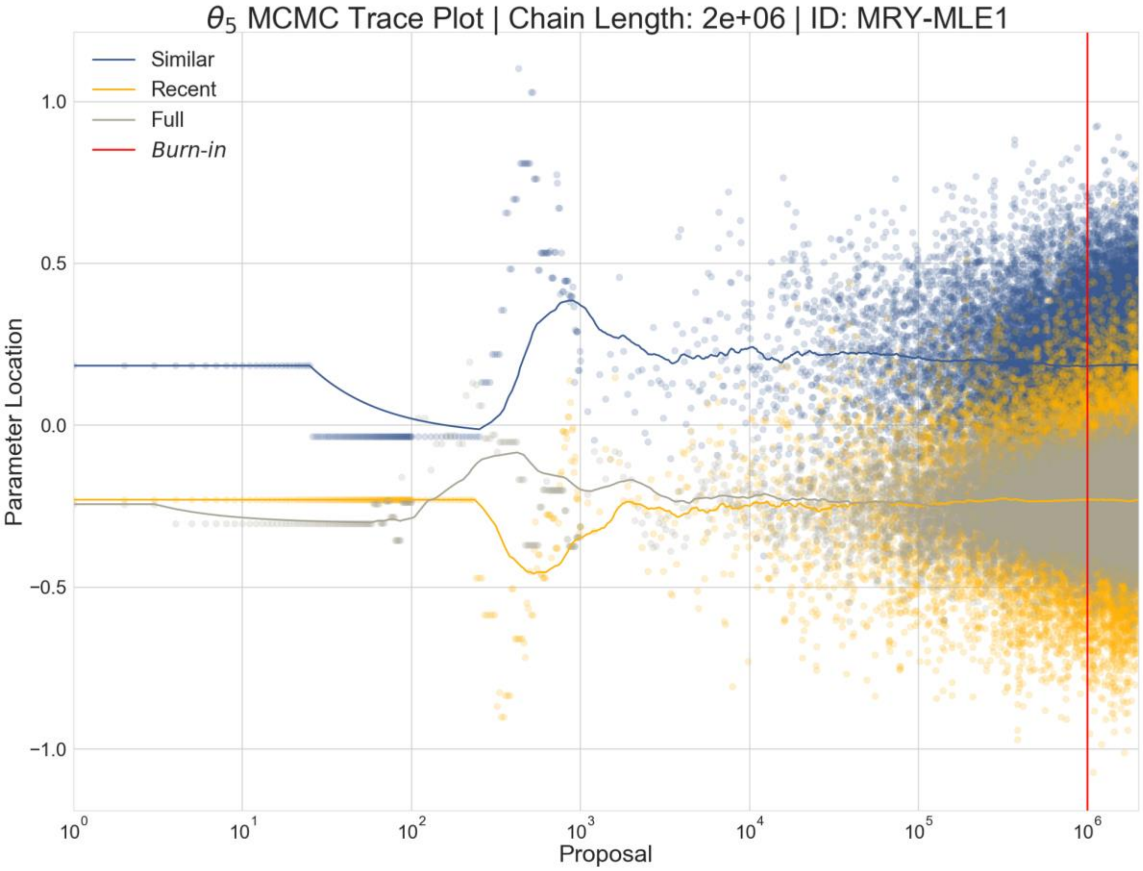


Figure 31. Semi-log trace diagram of an NPS BEMOS Markov chain. MCMC samples generated from an adaptive multiparameter variant of the Metropolis algorithm are depicted with filled circular markers. The ordinate describes the location of the parameter proposal in the phase space of the model as a function of proposal number (abscissa). The red vertical line indicates model "burn-in," which describes a pre-defined boundary between valid and invalid posterior parameter samples. The solid blue, gold, and grey horizontal lines describe the moving average of the chain along this parameter dimension.

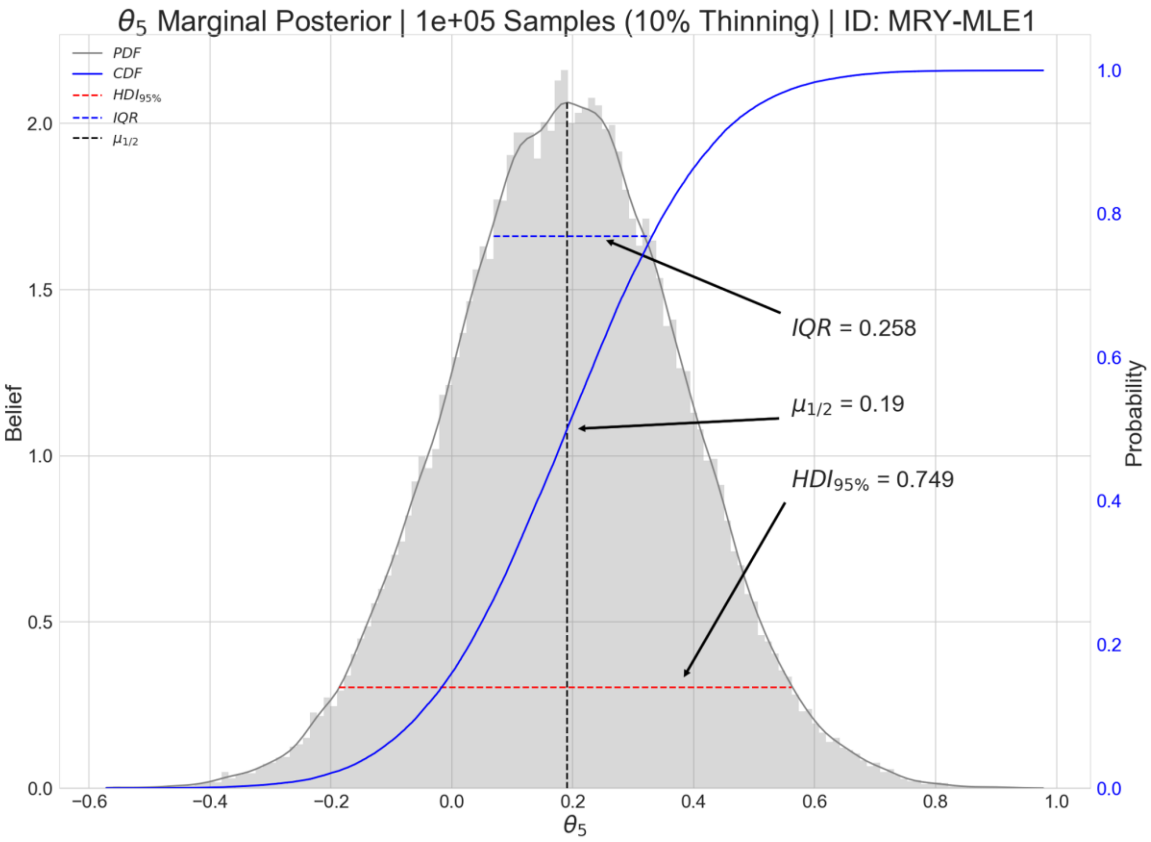


Figure 32. Distribution of discrete posterior parameter samples obtained from "thinning" the "Similar" (blue) Markov chain in Figure 31. Posterior samples are indicated with a grey histogram and superimposed KDE for the PDF; the cumulative distribution function (CDF) is plotted in solid blue over the model parameter. The sample median is indicated with a dashed black line; the narrowest interval containing (50%) 95% of the probabilistic mass is indicated with a dashed (blue) red line.

3. Posterior Odds and Multiparameter Updates

Gelman et al. (2013) notes that the Metropolis algorithm actually views the Bayesian target distribution according to Equation 27—the posterior odds ratio. The latter is a product between the prior odds and the likelihood ratio, and is frequently modified in data science with another log transformation to produce

$$\ln\left[\frac{p(\underline{\theta}_{i+1}\,|\,y)}{p(\underline{\theta}_i\,|\,y)}\right]=\ln\left[p(y\,|\,\underline{\theta}_{i+1})\right]-\ln\left[p(y\,|\,\underline{\theta}_i)\right]+\ln\left[p(\underline{\theta}_{i+1})\right]-\ln\left[p(\underline{\theta}_i)\right], \qquad (51)$$

where each quantity is as before—now only transformed to benefit from a sum of natural logarithms. In this way, the log transformation of the posterior odds provides a complete description of the cost function evaluated by the Metropolis algorithm and, indeed, the models produced by the present work. Moreover, this modification is valid in the same way that log-likelihood optimization is valid; the logarithmic transformation preserves the order of the original data while significantly reducing computational issues associated with the products and ratios of exceptionally large and small numbers. The present work experienced significant instances of computational "blow up" without this log transformation and is consistent with the work of other Bayesian/MCMC modelers (e.g., Wiecki 2015). It should also be noted that the details of the Bayesian JPM specify the right-hand side of Equation 51. More specifically, proposals generated by Equation 50 are now accepted with probability $\min\left[\ln(r), 0\right]$ so that location $\underline{\theta}_{t+1}$ is accepted and appended to the sequence if a random number drawn from the open interval $\ln\left[(0,1)\right]$ is less than $\ln(r)$ vis-à-vis Chapter II page 56. The benefits of the aforementioned log transformations are consistent with the data transformations introduced previously in this chapter. However, they are applied to model elements and, as a result, do not require inverse transformations.

The MVN proposal distribution indicated in Equation 50 is conceptually straightforward. However, it also assumes that parameter updates are performed simultaneously for each step in the chain. In simple terms, the model must specify how the individual model parameters (e.g., θ_5) are updated. Stansbury (2012b) describes two common multiparameter updating schemes: block-wise and component-wise updates. The former engages a full multivariate proposal distribution consistent with Equation 50 where each element of $\underline{\theta}$ is updated in parallel—that is, with each step in the evolving Markov chain. The component-wise updating scheme only updates a subset of $\underline{\theta}$—often only one dimension—with each Markov transition. Stansbury (2012b) similarly describes the advantages and disadvantages of each scheme: block-wise updates are complex and reject many proposals because it can be difficult to satisfy $\min\left[\ln(r), 0\right]$ according to

Equation 51 with each dimension of $\underline{\theta}$ changing at every step in the Markov chain. Component-wise updates address this problem, but develop fewer transitions along each dimension of $\underline{\theta}$, since the total number of transitions must now be allocated between the parameter dimensions. As a result, component-wise updating can require more total transitions to reach equilibrium (i.e., model convergence) when compared with the former scheme.

As indicated by Equation 50, the present work has selected block-wise component updates to improve the efficiency of the adaptive sampler (i.e., require fewer total proposals to reach convergence). Moreover, the NPS BEMOS model pre-allocates vector samples from $MVN(\underline{\theta}_i, \sigma_k^2 I)$ to avoid the computational cost of successive Python function calls. In this way, an NxL matrix of samples are drawn from $MVN(\underline{0}, I)$ for an L-dimensional parameter vector and subsequently normalized along each row so that each sample of $\underline{\theta}$ created in this manner represents a random jump from the current chain location (i.e., $\underline{\theta}_i$) along the boundary of an L-dimensional unit hypersphere—that is, a unit L-sphere. Finally, the adaptive updates to σ_k^2 in Equation 50 are recombined with this pre-allocated matrix of unit L-sphere jumps and the current chain location (i.e., $\underline{\theta}_i$) to produce an unstandardized jump to a new location (i.e., $\underline{\theta}_{i+1}$) with each transition.

E. REFERENCE METHOD

Chapter II developed a portion of the theoretical framework that connects MLE and Bayesian parameter estimation schemes. However, the original method of least squares, or OLS, introduced by Gauss, Legendre, and Laplace serves as a ubiquitous point estimation scheme in science today—one that ultimately led to Fisher’s work with MLE (Feigelson 2015). Conveniently, the multivariate multiple linear regression approach to statistical post-processing can similarly be pursued with OLS techniques. That is, an OLS cost function can provide point estimates of the regression parameters in Equation 43. To this end, Neter et al. (1996) offers a complete analytical solution to Equation 4 with the Moore–Penrose pseudoinverse as

$$B = (X^T X)^{-1} X^T Y, \qquad (52)$$

where X is the NxK design matrix of predictor variables from the training period, Y is the NxM predictand matrix of observations from the training period, and B contains the regression parameters from Equation 20 and Equation 43. As previously noted, Bayesian models with Gaussian likelihood functions and noninformative priors will produce posterior distributions that are similar, but not identical to, the point estimates generated by classical OLS methods when errors are assumed to be normally distributed. More specifically, Neter et al. (1996) describes this for one predictand as

$$y = Xb + \varepsilon, \qquad (53)$$

where y is a $Nx1$ column vector, X is as before, b is an arbitrary column of betas from Equation 43, and ε is a $Nx1$ column vector of normally-distributed errors. For this scenario, OLS is minimizing the SSR consistent with Equation 8—a quantity that is similar to right-hand side of Equation 12. The presence of the variance parameter does add additional information into the MLE/Bayesian cost function.

Neter et al. (1996) offers a classical frequentist approach to variance estimates associated with the OLS inferences in Equation 52 and Equation 53. More specifically, Neter et al. (1996) expresses the estimated variance of unobserved observable data as

$$s_p^2 = MSE\left[1 + X_h^T (X^T X)^{-1} X_h\right], \qquad (54)$$

where s_p^2 is the variance of the univariate prediction, X_h is a $1xK$ row vector of predictors for the current forecast instance (i.e., not from the training period), X is the NxK design matrix from before. The MSE describes the error mean square, or residual mean square, as

$$MSE = \frac{\sum_{i=1}^{N}\left(Y_i - Y_i\right)^2}{N-2}, \qquad (55)$$

where the numerator is the SSR from Equation 8 and N describes the number forecast instances in the training period. To produce reference distributions that provide meaningful comparisons with NPS BEMOS model output, the present work extends the variance estimates in Equation 54 to a multivariate framework. More specifically, the single-valued parameter estimates are combined with contemporary predictors vis-à-vis Equation 4 to produce point estimates for the desired predictands. A multivariate Gaussian predictive distributed is therefore centered on these OLS point estimates and combined with diagonal variance elements from Equation 54 as

$$z \sim MVN(X_h B, \Sigma_p), \tag{56}$$

where z is the multivariate predictand, X_h is as before, and B is determined by Equation 52. For $M = 3$ and the application in Chapter IV, Σ_p is therefore properly visualized as

$$\Sigma_p = \begin{bmatrix} s_{p,1}^2 & 0 & 0 \\ 0 & s_{p,2}^2 & 0 \\ 0 & 0 & s_{p,3}^2 \end{bmatrix}. \tag{57}$$

F. PERFORMANCE METRICS

Scoring rules appropriate for measures-oriented and distributions-oriented forecasts have been considered for this dissertation. While there are numerous approaches to evaluating forecast performance (e.g., Murphy and Winkler 1987; Murphy 1993; Wilson et al. 1999), the present work focused on the PIT visualizations of Dawid (1984) and Diebold et al. (1998), the calibration and sharpness analysis of Gneiting et al. (2007), and the proper scoring rules for continuous forecast distributions articulated by Gneiting et al. (2005), Bröcker and Smith (2007), Gneiting and Raftery (2007), and Bröcker (2012). This special class of metrics is concisely defined in Gneiting and Raftery (2007) by the following description:

> A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than G [not equal to] F. It is strictly proper if the maximum is unique. In prediction problems, proper scoring rules encourage the forecaster to make careful assessments and to be honest. In estimation problems, strictly proper scoring rules provide attractive loss

> and utility functions that can be tailored to the problem at hand. (Gneiting and Raftery 2007)

Examples of strictly proper scoring rules in Bröcker and Smith (2007) include the ignorance score of Good (1952) in Equation 58, the Brier score in Equation 59, and the proper linear score in Equation 60; the error in the mean of Equation 61 is non-strictly proper. These quantities are respectively defined by Bröcker and Smith (2007) as

$$S[p(x), X] = -\ln[p(X)], \tag{58}$$

$$S(p, X) = (X - p)^2, \tag{59}$$

$$S[p(x), X] = \int p^2(z)dz - 2p(X), \tag{60}$$

$$S[p(x), X] = (X - m_p)^2, \tag{61}$$

where $S[p(x), X]$ corresponds to the score expectation associated with the specified rule, X is an *observed* forecast quantity (i.e., the realized predictand) associated with the a random variable x (i.e., the predictand), m_p is the mean of probabilistic forecast distribution $p(x)$, z is the associated dummy integration variable, and $p(X)$ describes the probability density assigned by $p(x)$ to outcome X. The notation in Equation 59 differs from the other metrics in evaluating a discrete probability p when X is constrained to assume binary verification values—that is, $X = 0$ or $X = 1$ (Bröcker and Smith 2007). Finally, the scoring rules detailed in the following sections have been selected by the present work based on their relevance to the NR/EMOS and BEMOS analyses of Gneiting et al. (2005) and Richter (2012), and, indeed, the forecast application presented in Chapter IV,

1. Measures-Oriented Scores

This dissertation used three primary scoring rules for single-valued estimates of a desired forecast quantity. The first is the classical, ubiquitous mean absolute error (MAE). Richter (2012) summarizes it according to Gneiting et al. (2005) as

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | f_i - t_i | \tag{62}$$

where f_i and t_i are the prediction and truth on the i^{th} forecast trial. Consistent with notation and theory introduced by Taylor (2001), the present work has also adopted a variant of the canonical mean squared error (MSE)—that is, the centered MSE as

$$E^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\left(f_i - \overline{f}\right) - \left(t_i - \bar{t}\right)\right]^2, \tag{63}$$

where f_i and t_i are as before and overbar quantities describe the mean of each variable. Additionally, the linear correlation between the forecasts and observations is described by the PCC as

$$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[\left(f_i - \overline{f}\right)\left(t_i - \bar{t}\right)\right]}{\sigma_f \sigma_t}, \tag{64}$$

where σ_f and σ_t are the standard deviations of the forecasts and observations. The latter metrics will not be computed and displayed separately, but rather combined with the variance of the forecasts and observations (i.e., the quantities in the denominator of Equation 64) in an eponymous Taylor diagram (e.g., Figure 33).

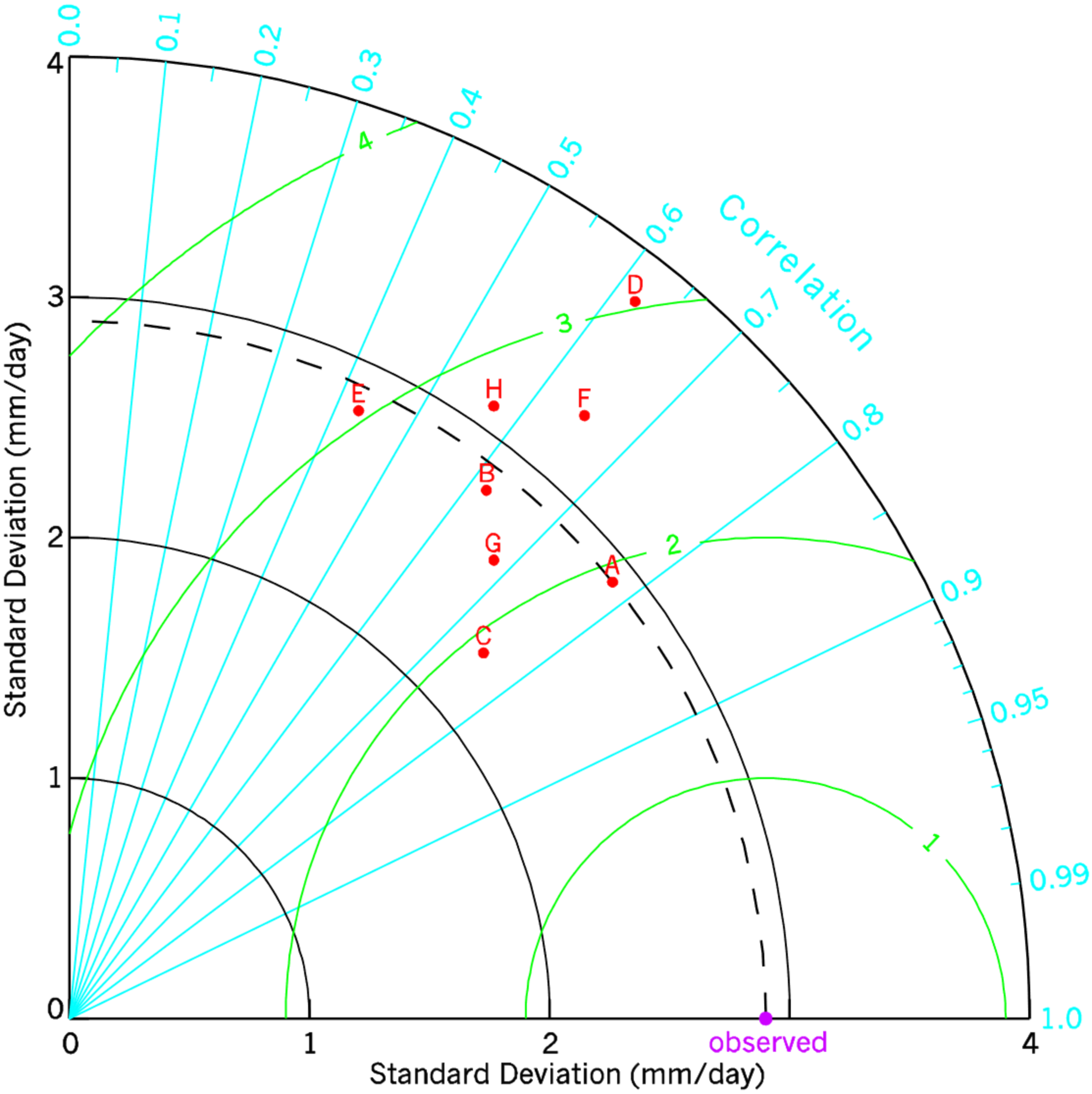


Figure 33. Taylor diagram indicating three distinct measures of forecast performance for eight single-valued predictions of annual mean precipitation. Source: Taylor (2005). Forecast variability is indicated by the radial distance between the indicated forecasts and the true origin (black); the variability of the observations is indicated with the dashed, constant-radius arc marked "observed" (purple). The linear correlation (i.e., PCC) between forecasts and observations is visualized by the azimuthal coordinate (teal) of the indicated forecasts. The centered RMS error for each of the indicated forecasts is indicated by the radial distance from the adjusted, reference origin (purple). Generalized forecast performance is indicated by the location of each forecast relative to green isopleths of centered RMS error.

Taylor (2001) recognized that the canonical Law of Cosines could be reformulated to express the quantities in Equation 63 and Equation 64 as a single expression. In particular, Taylor (2001) observed that the Law of Cosines, which may be written as

$$c^2 = a^2 + b^2 - 2ab\cos\left(\phi\right), \tag{65}$$

could be adapted to express the centered MSE as a function of the PCC and forecast and observation variances as

$$E^2 = \sigma_f^2 + \sigma_t^2 - 2\sigma_f \sigma_t \rho, \tag{66}$$

where ρ is the linear correlation (i.e., PCC) between the forecasts and observations from Equation 64 and all other quantities are as before. In this way, Taylor diagrams provide three distinct performance metrics, which are connected by the Law of Cosines, on a single two-dimensional graph. Using a polar coordinate system as a reference, the linear correlation ρ is interpreted as $\cos\left(\phi\right)$ and, therefore, associated with the azimuthal coordinate; the variability of the forecasts, σ_f, is associated with the radial coordinate. Using the variability of the observations, σ_t, as a reference, Equation 66 then creates a new radial dimension with a displaced origin at $p(r,\phi) = (\sigma_t, 0)$.

The radial distance between any estimate in the plane and this new origin describes the square root of the centered MSE—that is, the centered RMS error. As a result, Taylor diagrams provide a convenient method of visualizing measures-oriented forecast performance; one can quickly assess comparative model quality over a set of forecast trials by the separation between each forecast estimate and the center of the “bullseye” pattern of green isopleths in Figure 33. Models with better centered RMS error will appear closer to the center of this “bullseye” pattern. Moreover, the metrics that compose the centered RMS error are still visible on the same plot. In this way, one can verify that a predictive scheme is producing forecast variability consistent with the observations—a common deficiency with statistical forecasts—and, indeed, that the forecasts are well-correlated with the observations—all while assessing traditional MSE.

2. Distributions-Oriented Scores

Hersbach (2000), Gneiting et al. (2005), Juban et al. (2007), Gneiting and Raftery (2007), Bröcker (2012), and Richter (2012) give additional consideration to the continuous ranked probability score (CRPS), which Gneiting and Raftery (2007) define as

$$CRPS(F,x) = -\int_{-\infty}^{\infty} \left[F(y) - H(y-x)\right]^2 dy, \tag{67}$$

where y is a random variable for the predictand with realized x, $F(y)$ is the cumulative distribution function for forecast distribution $p(y)$, and $H(y-x)$ is the Heaviside step function. According to Hersbach (2000), this proper scoring rule is derived from the ranked scoring rules considered by Matheson and Winkler (1976), which have special utility to the evaluation of Bayesian methods. To this end, Bröcker (2012) notes that the CRPS permits direct comparisons of EPS performance without the need for additional kernel dressing. Gneiting and Raftery (2007) similarly observe its utility with MCMC methods and its relation to the discrete Briar score. Juban et al. (2007) also describes its relation to MAE for discrete forecasts. Grimit et al. (2006) note that the CRPS for linear forecast variables may be reported for an ensemble of discrete forecasts as

$$CRPS(F_{ens},x) = \frac{1}{M}\sum_{m=1}^{M}\left|x_m - x\right| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\left|x_m - x_n\right|, \tag{68}$$

where F_{ens} is the predictive distribution comprised of $x_1, x_2, \ldots, x_M$ single-valued estimates from an EPS or PPD, and x is the true value observed for this forecast trial; the right-hand side of Equation 68 describes the absolute difference between each discrete element of F_{ens} and every other over a double sum of sequences. In this way, Grimit et al. (2006) observe the utility of the CRPS:

> The first term on the right-hand side of [Equation 68] is the expected value of the absolute error, and the second terms is a correction factor that measures the sharpness of probabilistic forecast F and renders it proper. The linear CRPS generalizes the absolute error, to which it reduces if F is a deterministic forecast…Thus, the linear CRPS provides a direct way of comparing deterministic and probabilistic forecasts with a single metric

that is proper and reported in the same unit as the observations, making it a tangible measure of predictive performance. (Grimit et al. 2006)

As with Gneiting et al. (2005), Bröcker and Smith (2007) and Gneiting and Raftery (2007), Grimit et al. (2006) also note the resurgence in popularity of the dimensionless ignorance score introduced by Good (1952). It represents an intuitive metric for predictive distributions, as it measures the *a posteriori* probability density assigned to truth. As mentioned in Chapter II, this does require KDE to generate a PDF for the associated predictive distribution. However, the CRPS and ignorance score are both strictly proper scoring rules well-suited to a rigorous evaluation of the predictive distributions produced by the OLS, MLE, and Bayesian methods in this dissertation. To this end, the present work will report the ignorance score (or logarithmic score) alongside the aforementioned CRPS.

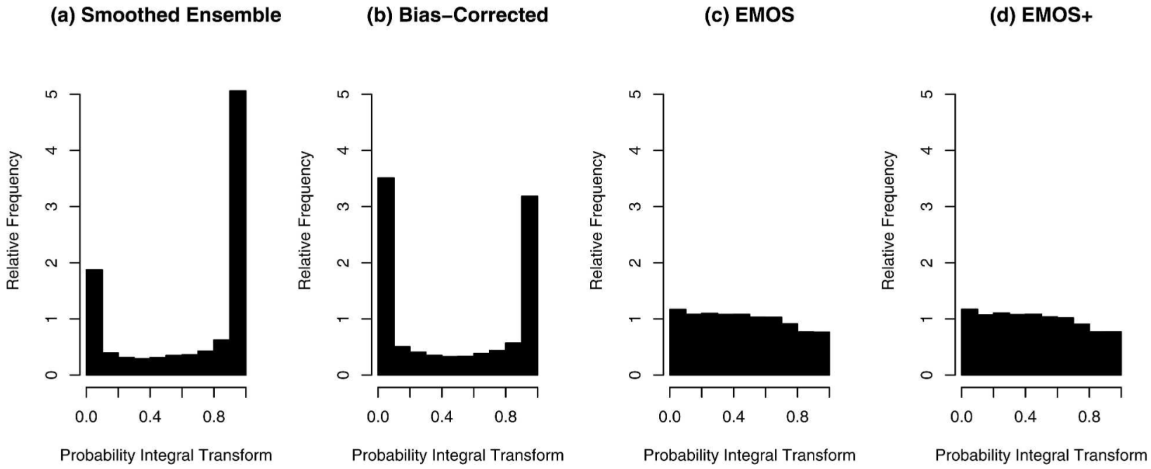


Figure 34. PIT histograms for four probabilistic forecast schemes of maximum diurnal surface temperature. Source: Gneiting et al. (2005). Calibrated PPDs have a uniform appearance; biased and underdispersive PPDs have skew and curvature.

Probability integral transforms (PIT) have been used to quickly assess the calibration of forecast distributions according to Dawid (1984), Diebold et al. (1998), Gneiting et al. (2005), and Richter (2012). The PIT of a probabilistic forecast is $PIT = F(y)$, where F describes the cumulative distribution function (CDF) associated with predictive distribution $p(y)$ and truth y. In this way, PIT histograms represent a discrete distribution of probabilities associated with predictive CDFs. Calibrated PPDs will therefore have a relatively uniform PIT histogram (e.g., Figure 34c/d); biased PPDs will have a skewed presentation consistent with Figure 34a; underdispersive PPDs will appear as in Figure 34b.

The final measure of ensemble calibration considered for this dissertation is provided by reliability diagrams. They compare the relative frequency of observed outcomes against the probabilities assigned to them by a forecast scheme (CAWCR 2017). The present work has modified the traditional reliability diagram and, indeed, the coverage of the central prediction intervals considered by Gneiting et al. (2005) and Richter (2012) to examine the reliability of Bayesian highest density intervals (HDI). These intervals may be computed for any predictive distribution regardless of the generative process; they describe the narrowest interval that contains a specified fraction of the probabilistic mass (Kruschke 2014). In this way, a 95% HDI would refer to the narrowest interval that contained 95% of a continuous distribution's probabilistic mass. These intervals are often equivalent to central prediction intervals (e.g., symmetric Gaussian PDF), but may be different when the PPD is, for example, skewed or multimodal. To this end, the present work computed the HDI endpoints for 19 evenly-spaced probabilistic mass fractions (i.e., $0.05, 0.1, \ldots, 0.95$) for every one of the 172,935 forecast distributions evaluated in this dissertation. Each HDI can then be checked against the corresponding observations to examine the relative frequency these intervals contained the true value of the predictand. In this way, a calibrated forecast distribution will produce probability statements (i.e., predictive HDIs) that match the frequency these intervals contained the observations.

IV. ANALYSIS

A. MOTIVATION

The present work was initially motivated by the University of Oklahoma Weather Challenge (WxC)—a North American collegiate weather forecasting competition available to alumni, faculty, staff, and students (i.e., undergraduate and graduate) of institutions of higher education (WxC 2017). More specifically, NPS BEMOS was conceived in the search for a programmatic scheme that could consistently convert ensemble predictors into reliable forecast distributions that outperform raw objective guidance and, indeed, other WxC competitors. In this way, the competition compels roughly 1200 human participants to use a combination of objective guidance—that is, *real* data from operational NWP models—and their dynamical reasoning to submit daily meteorological predictions for diurnal extrema of select sensible weather variables. These predictions are made in parallel with National Weather Service (NWS) official forecasts for designated American cities and are designed to provide a rigorous assessment of real-world forecast skill.

The competition requires participants to predict diurnal maximum surface temperature, diurnal minimum surface temperature, diurnal maximum wind speed, and diurnal cumulative liquid precipitation for a 24-hour forecast window. Predictions are submitted no later than six hours prior to this window; however, most objective guidance available for the relevant forecast is initialized 12 to 24 hours before the valid period. Verification is facilitated by observational stations in the Automated Surface Observing System (ASOS), which are typically located at suitable airports near cities selected for the competition. In this way, each forecast season comprises 11 cities with functional ASOS stations. Eight forecasts are allocated to each non-tournament city over a piecewise continuous period of two calendar weeks; that is, forecasts are submitted Monday through Thursday on the first week with an inclusive break between Friday and Sunday. Forecasting for the current city resumes on Monday of the second week as before. The next city becomes active on the following Monday after the final four days of the current city have been completed.

The WxC preserves the fundamental challenges of any operational forecast and illustrates the difficulties of meteorological prediction—even over relatively short timescales. Human participants quickly learn that objective model guidance is difficult to consistently outperform. They are also likely to note the skill of the WxC consensus solution, which represents the central tendency of all competition participants; it also serves as a practical demonstration of the efficacy of filtered solutions vis-a-vis Leith (1974) and Kalnay (2003). Nevertheless, the short time horizon of WxC forecasts accommodates the strengths of subjective (i.e., human) guidance and high-resolution, limited-area dynamical models. As previously stated, the former is well-positioned to monitor model performance, add prognostic detail in the mesoscale domain, and communicate forecast uncertainty—elements that become more difficult to engage with longer forecast lead times. Limited-area (i.e., regional) models, by comparison, are specialized to outperform global models by adding fine structure to meteorological prediction at short timescales. Kalnay (2003) notes some of the primary advantages and disadvantages:

> Because of their higher resolution, regional models have the advantage of higher accuracy and the ability to reproduce smaller-scale phenomena such as fronts, squall lines, and much better orographic forcing than global models. On the other hand, regional models have the disadvantage that, unlike global models, they are not "self-contained" because they require lateral boundary conditions at the borders of the horizontal domain. For this reason, regional models are used only for short-range forecasts. After a certain period, which is proportional to the size of the model, the information contained in the high-resolution initial conditions is "swept away" by the influence of the boundary conditions, and the regional model becomes merely a "magnifying glass" for the coarser model forecast in the regional domain. (Kalnay 2003)

In this way, the WxC is reasonably regarded as a forecast application that emphasizes the strengths of human forecasters and high-resolution dynamical guidance. With the same reasoning, the competition format diminishes the efficacy of statistical forecast methodologies relative to the former. Gneiting et al. (2005) and Richter (2012) considered a similar forecast application for their research; the former examined "48-h forecasts of sea level pressure and surface temperature over the northwestern United States and British Columbia, using phase I of the University of Washington ensemble"

(Gneiting et al. 2005). Richter (2012) focused on the same geographic area with the University of Washington Mesoscale Ensemble, but only considered surface temperature forecasts as a predictand. While prosaic from a research perspective, Gneiting et al. (2005) notes that these sensible weather variables are consistent with the practical interests of the public vis-à-vis Murphy and Winkler (1979). As a result, the present work considers the WxC format an appropriate forecast environment for the NR/EMOS technique and the NPS BEMOS model developed for this dissertation. Moreover, shorter lead-times and diurnal extrema provide additional difficulties for statistical post-processing methods relative to high-resolution dynamical guidance. To facilitate a more tractable analysis, NPS BEMOS performance will therefore be compared to raw ensemble guidance and the reference method with diurnal maximum surface temperature, diurnal minimum surface temperature, and diurnal maximum wind speed predictands.

B. DATA

1. Parent Ensemble

The daily 15Z run of the NCEP Short Range Ensemble Forecast (SREF) system was selected for training and comparison with the NPS BEMOS model. This is the last SREF model run available before the WxC submission deadline at 00Z and corresponds to a 15-hour lead time before the start of the next 24-hour forecast window (i.e., 06Z). The SREF system was recognized as a convenient, publically-available source of high-resolution ensemble model output with a 16-km horizontal resolution and 40 vertical levels (Du et al. 2015). It also enables efficient programmatic collection of predictors at most WxC forecast locations for the past 12 months. The present work focused on SREF plume output (e.g., Figure 35), which describes ensemble flows from model initialization through 84 hours. After a significant model upgrade on October 21, 2015, SREF plumes are derived from the Weather Research and Forecasting (WRF) model, which comprises two dynamical solvers or cores: the Advanced Research WRF (WRF-ARW) and the Nonhydrostatic Multiscale Model on B-Grid (NMMB) model (NWS 2015). These SREF plumes provide a total of 26 ensemble predictors; this includes six positive perturbations, six negative perturbations, and one control solution from each model core (NWS 2015).

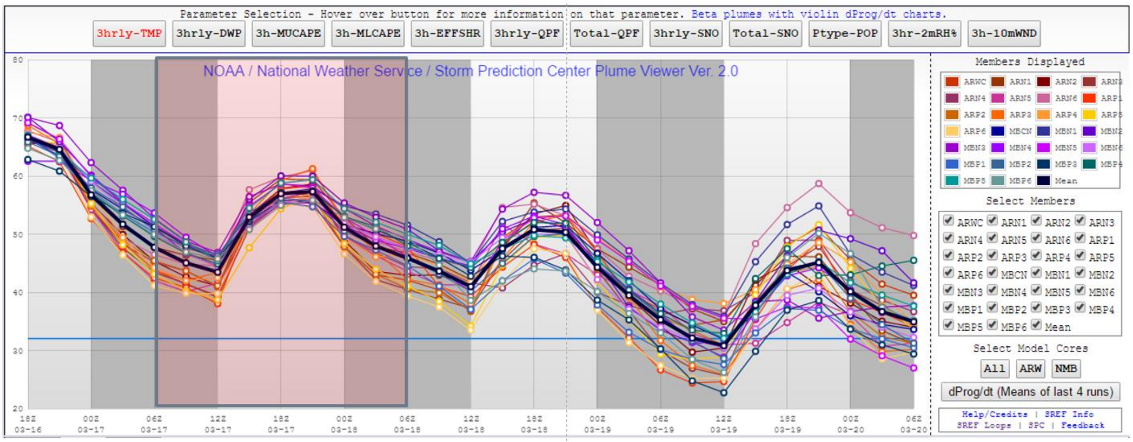


Figure 35. A times series of NCEP SREF ensemble flows for surface temperature at Beckley, WV for March 16–20, 2016. Adapted from SREF 2016. Each discrete flow represents a member of the 26-member SREF plume. Flows with a "warm" hue are associated with the ARW core; "cold" hues correspond to the NMMB core. The area shaded in red corresponds to a 24-hour forecast window for the WxC forecast competition. Extrema associated with each flow are extracted and used as training data by the present work.

2. Verification

The predictors extracted from SREF plumes must be combined with corresponding observations to produce a full set of training data. Data collected from the aforementioned ASOS stations is transformed into observed weather reports by NWS weather forecast offices (WFOs) and hosted by the Climate Services Division for public distribution. The present work selected preliminary monthly climate (CF6) data for verification of designated WxC cities. These reports convert daily observations into monthly summaries of relevant sensible weather variables. The NWS provides the following description of CF6 data: "daily weather statistics for the month, including temperatures, precipitation, degree days, wind and sky cover. In addition, monthly statistics such as average temperatures and departures from normal, degree days, and rainfall are also included. This product is available for up to 5 years" (NWS 2017). All variables (i.e., SREF and CF6) were converted to the International System of Units (SI) so that surface temperatures are reported with the Kelvin scale [K] and surface wind speeds are reported in meters per second $[m\ s^{-1}]$.

3. Research Design

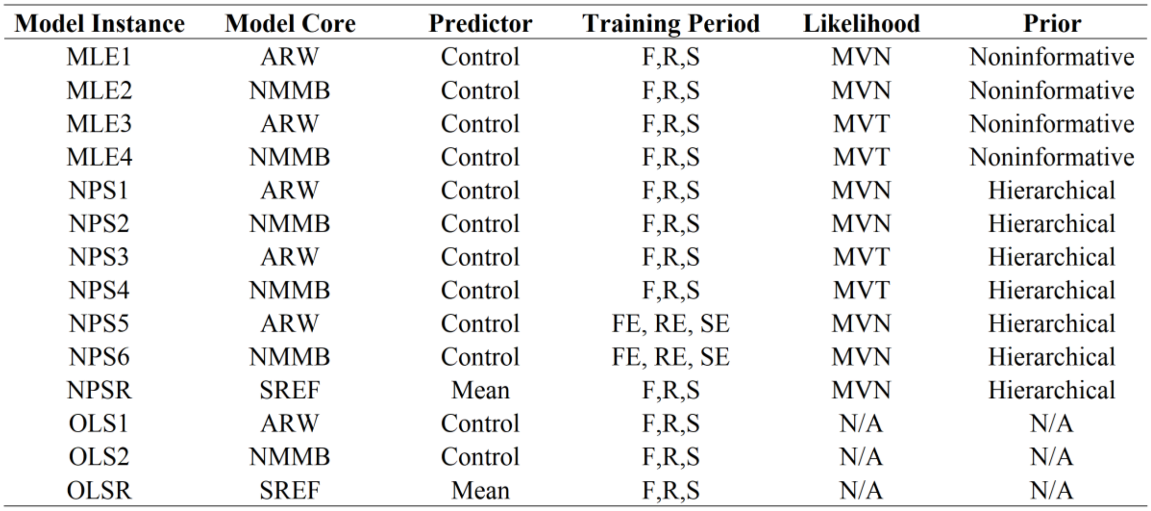
This dissertation generated 14 primary variations (i.e., Table 2) of the multivariate multiple linear regression model described in Chapter II. As previously stated, model instances labeled “MLE” describe a Bayesian probability model with a stochastic regression structure specified by Equation 9 and noninformative prior information (i.e., $p(\underline{\theta}) \propto 1$) indicated by Figure 23; instances labeled as “NPS” share the stochastic regression structure of Equation 9 and add full hierarchical prior information according to the JPM in Figure 24; “OLS” instances have a deterministic regression structure consistent with Equation 3 and, as a result, have no likelihood or prior distributions. It should be noted that “MLE” and “NPS” inferences were completed with the adaptive multiparameter Metropolis sampler developed by this dissertation. The OLS reference method was solved directly with the linear algebra specified by Equation 52. In this way, each model instance was developed to explore a disparate element of the research questions posed at the beginning of Chapter III. In no particular order, these model conjectures considered the impact on predictive performance from

- training period length and character
- the type of ensemble predictor(s) used to train the model
- the distributional form of the parametric likelihood function selected
- the presence of Bayesian prior information.

These 14 primary model instances were designed *a priori* to express measureable variability in forecast performance according to the assumptions of the aforementioned research questions. To this end, this dissertation considered a number of additional perturbations to the MLE, NPS, and OLS comparisons described above; these elements are indicated by the additional columns included in Table 2. The “Model Core” and “Predictor” columns, for example, describe the source of the predictor variables used to train the indicated model instance. All model instances use three predictor variables consistent with $K-1=3$ in Equations 3 and 9. More specifically, the stochastic regression framework for “MLE” and “NPS” instances is now properly visualized as

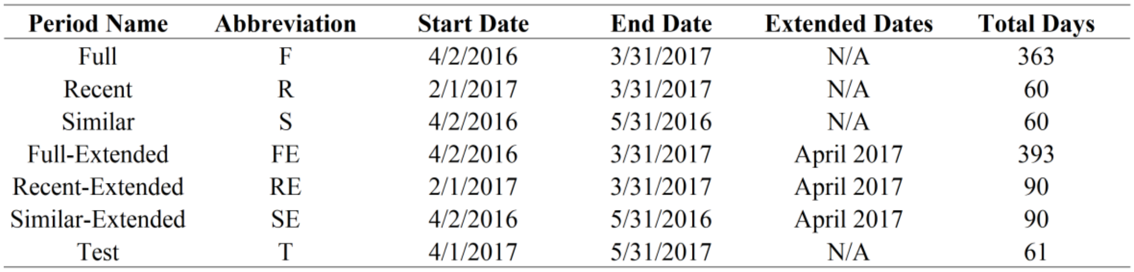
$$\left[\mu_1,\mu_2,\mu_3\right]=\left[1,x_1,x_2,x_3\right]\begin{bmatrix}\beta_0 & \beta_1 & \beta_2\\ \beta_3 & \beta_4 & \beta_5\\ \beta_6 & \beta_7 & \beta_8\\ \beta_9 & \beta_{10} & \beta_{11}\end{bmatrix}, \tag{69}$$

where μ_1,μ_2,μ_3 describe the updated multiparameter location of the parametric likelihood function with predictors x_1,x_2,x_3 and regression coefficients $\beta_1,\beta_2,\ldots,\beta_{11}$. The deterministic regression framework for OLS estimation is similar; however, the left-hand side of Equation 69 becomes observable predictands—that is, y_1,y_2,y_3. The critical distinction made by the present work, for both deterministic and stochastic approaches, is that x_1,x_2,x_3 correspond specifically to maximum diurnal surface temperature, minimum diurnal surface temperature, and maximum diurnal surface wind speeds, respectively, from the dynamical solvers indicated in column two of Table 2. Column three clarifies further whether the control member or ensemble mean was used for x_1,x_2,x_3.



Model Instance	Model Core	Predictor	Training Period	Likelihood	Prior
MLE1	ARW	Control	F,R,S	MVN	Noninformative
MLE2	NMMB	Control	F,R,S	MVN	Noninformative
MLE3	ARW	Control	F,R,S	MVT	Noninformative
MLE4	NMMB	Control	F,R,S	MVT	Noninformative
NPS1	ARW	Control	F,R,S	MVN	Hierarchical
NPS2	NMMB	Control	F,R,S	MVN	Hierarchical
NPS3	ARW	Control	F,R,S	MVT	Hierarchical
NPS4	NMMB	Control	F,R,S	MVT	Hierarchical
NPS5	ARW	Control	FE, RE, SE	MVN	Hierarchical
NPS6	NMMB	Control	FE, RE, SE	MVN	Hierarchical
NPSR	SREF	Mean	F,R,S	MVN	Hierarchical
OLS1	ARW	Control	F,R,S	N/A	N/A
OLS2	NMMB	Control	F,R,S	N/A	N/A
OLSR	SREF	Mean	F,R,S	N/A	N/A

Table 2. Summary of various NPS BEMOS model instances evaluated by this dissertation. Each column describes a research design element that was perturbed to explore the conjectures outlined in Chapter III. Training period details are examined further in Table 3.

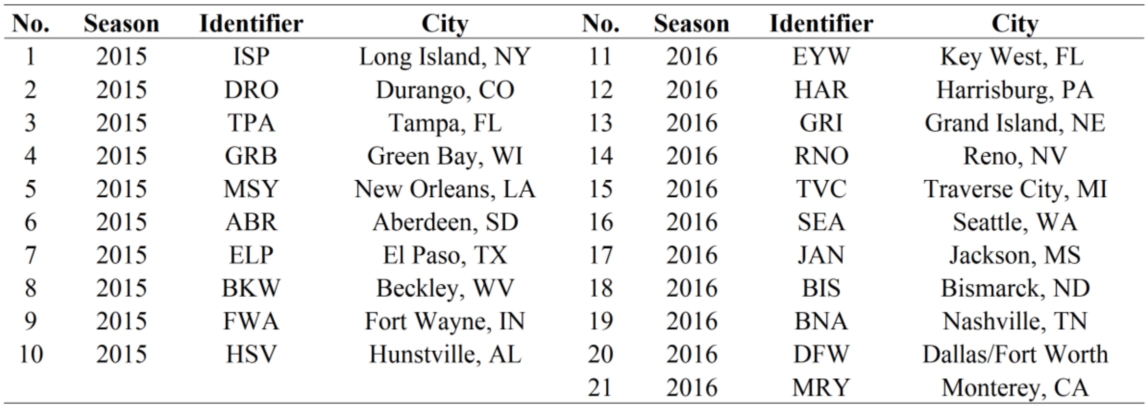


Period Name	Abbreviation	Start Date	End Date	Extended Dates	Total Days
Full	F	4/2/2016	3/31/2017	N/A	363
Recent	R	2/1/2017	3/31/2017	N/A	60
Similar	S	4/2/2016	5/31/2016	N/A	60
Full-Extended	FE	4/2/2016	3/31/2017	April 2017	393
Recent-Extended	RE	2/1/2017	3/31/2017	April 2017	90
Similar-Extended	SE	4/2/2016	5/31/2016	April 2017	90
Test	T	4/1/2017	5/31/2017	N/A	61

Table 3. Summary of various training and test periods considered by this dissertation. All intervals are continuous except for the “Similar-Extended” period, which is piecewise continuous with a break between the indicated end date and the start of the extended training period (i.e., April 2017).

Column four of Table 2 describes the training period given to the indicated model instances according to the abbreviations indicated in Table 3. Each period corresponds to a different set of training data and, depending on the details of the parameter estimation specified (i.e., OLS, MLE, and NPS), potentially distinct predictive distributions during the test period. It should be noted that the extended periods include a portion of the test period and are, therefore, not valid indicators of predictive performance for April 2017. However, they do provide valid comparisons for May 2017 and a difficult reference for non-extended models during April 2017 test days. Table 3 also communicates that the 14 primary model instances in Table 2 were further conditioned into 42 distinct models variations. The final two columns in Table 2 describe the structure of the JPM selected for each instance. The JPM comprises the likelihood function and the prior; the perturbations in these columns engage assumptions associated with the efficacy of MVT distributions in robust regression and, indeed, the ostensible importance of beta regression shrinkage through hierarchical prior information. In this way, each column of Table 2 represents a dimension along which the various OLS, MLE, and NPS statistical models were perturbed.

Comparisons between odd and even “MLE” and “NPS” model instances communicate the impact of potentially disparate predictor information coming from each dynamical solver (i.e., model core) in the SREF system. For the present work, odd instances always use the ARW core; even instances use the NMMB core. Model instances ending in a “3” or “4” similarly describe the influence of MVT likelihood functions vis-à-vis robust regression; all other model variations have a MVN likelihood function. Finally, the present work considers the research design details associated with NPS1 and NPS2 to represent “default” configurations of the NPS BEMOS model. They include hierarchical prior information to make them fully Bayesian in character, and they also use a standard MVN likelihood function. As a result, this default configuration was used as a foundation for the research questions focusing on the influence of additional training data (i.e., the extended training periods in the “April update” for NPS5 and NPS6) and the importance of SREEF ensemble mean predictors given to NPSR and OLSR. In this way, NPS5, NPS6, and NPSR share identical JPMs with NPS1 and NPS2—only training data has been modified between them. Model instances ending with an “R” use the ensemble means of the entire SREF ensemble as predictors for each sensible weather variable. This expresses an explicit desire to demonstrate that NPS BEMOS instances using ensemble control information can meet or exceed the performance of all other models considered by the present work.



No.	Season	Identifier	City	No.	Season	Identifier	City
1	2015	ISP	Long Island, NY	11	2016	EYW	Key West, FL
2	2015	DRO	Durango, CO	12	2016	HAR	Harrisburg, PA
3	2015	TPA	Tampa, FL	13	2016	GRI	Grand Island, NE
4	2015	GRB	Green Bay, WI	14	2016	RNO	Reno, NV
5	2015	MSY	New Orleans, LA	15	2016	TVC	Traverse City, MI
6	2015	ABR	Aberdeen, SD	16	2016	SEA	Seattle, WA
7	2015	ELP	El Paso, TX	17	2016	JAN	Jackson, MS
8	2015	BKW	Beckley, WV	18	2016	BIS	Bismarck, ND
9	2015	FWA	Fort Wayne, IN	19	2016	BNA	Nashville, TN
10	2015	HSV	Hunstville, AL	20	2016	DFW	Dallas/Fort Worth
				21	2016	MRY	Monterey, CA

Table 4. Summary of forecast cities selected for evaluation by this dissertation. Cities 1–10 (left) correspond to competition cities from the 2015–2016 WxC season. Cities 11–20 (right) correspond to competition cities from the 2016–2017 WxC season. The final city (i.e., No. 21; Monterey, CA) selection was based on the location of the author’s university.

Table 4 describes the 21 cities selected for model training and evaluation by the present work. In this way, the two most recent WxC competition seasons were sampled to produce a collection of cities sufficiently large to minimize the influence of local topography, mesoscale dynamics (e.g., coastal metrological considerations), and seasonal synoptic forcing. The quasi-uniform geographic distribution of these locations is indicated by Figure 36. The number of cities selected was also comparable with the original research design of Gneiting et al. (2005) and Richter (2012); however, these studies only considered the Pacific Northwest region. For the periods indicated in Table 3, these 21 cities generated a total of 30,240,000,000 MCMC transitions—the equivalent of 15,120 univariate Markov chains with 2,000,000 transitions. With 1% thinning, these MCMC samples produced a total of 126,819 point forecasts and 1,268,190,000 discrete PPD samples during the 61-day test period when summed over all predictands, cities, and training periods.

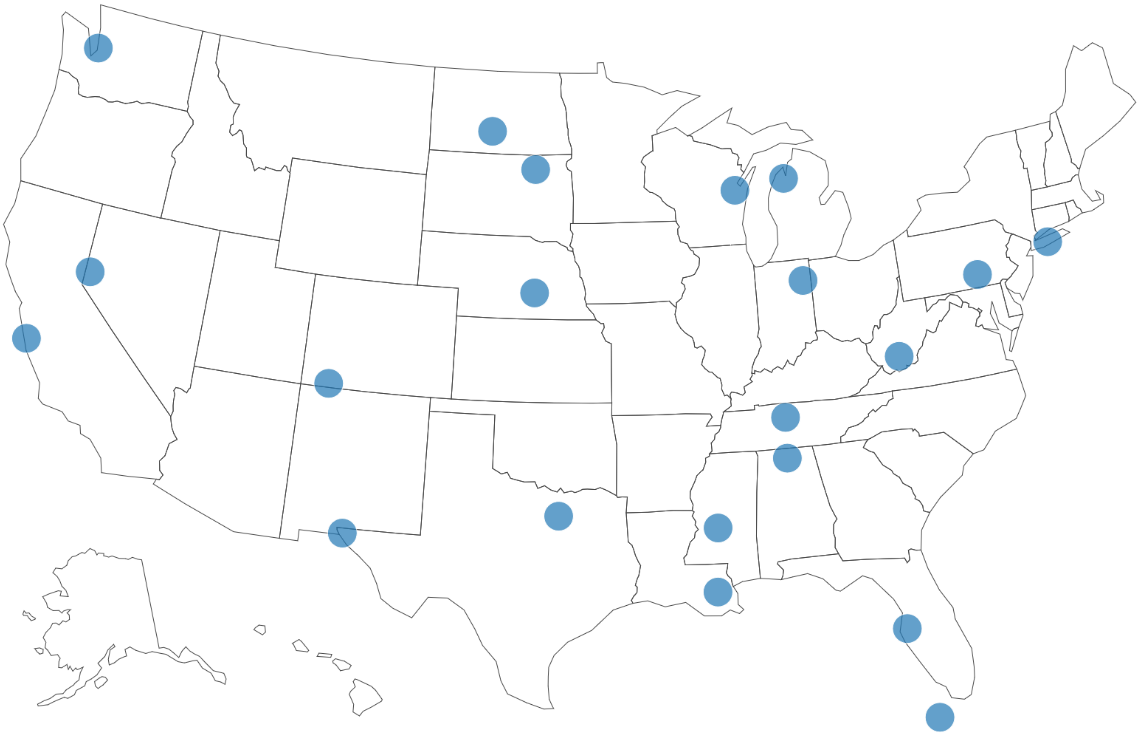


Figure 36. The geographic distribution of the 21 cities selected for a comparative model evaluation by this dissertation.

Finally, the parameter estimation schemes detailed in Chapters II and III can be applied to the present forecast application with a local or regional format vis-à-vis Richter (2012). The former describes an inference framework where posterior model parameter beliefs are conditioned on data from individual cities—no regional or national mixing of training data. In simple terms, this means each city in Figure 36 has unique regression parameters according to the details of the SREF forecasts and observations at that location. The regional approach mixes training data from neighboring cities according to shared topographic or synoptic characteristics, so that multiple cities will share the same regression coefficient beliefs; only the local predictor variables supplied by the SREF system will vary between them. While a regional approach might have merit in data- and time-constrained training environment, Richter (2012) found that the local approach was generally more effective. The local approach was selected by the present work based on these findings and the overdispersive distribution of cities in Figure 36.

C. RESULTS

1. Parameter Beliefs

Posterior parameter beliefs obtained from Bayesian data analysis and MCMC sampling techniques explicitly incorporate the uncertainty of the inference. As previously stated in Chapters II and III, no additional computations are required to produce uncertainty intervals for the parameter estimates we seek. In this way, a visual inspection of Bayesian posterior distributions reveals the impact of the model perturbations in Table 2. To this end, Figure 37 depicts a pairplot matrix for posterior parameter beliefs estimated by the NPS1 model for Monterey, CA conditioned by training period. It corresponds to the true PPC terms (i.e., ρ_{ij}) in the covariance matrix specified by Equation 44—not the raw covariance parameters in Equation 37. Diagonal elements of Figure 37 describe the marginal univariate PDFs for each parameter according to training period. Upper triangular elements describe pairwise joint scatter plots for the posterior samples collected for each parameter. Lower triangular elements depict the associated two-dimensional PDFs and isopleths of joint probability density for these pairwise relationships.

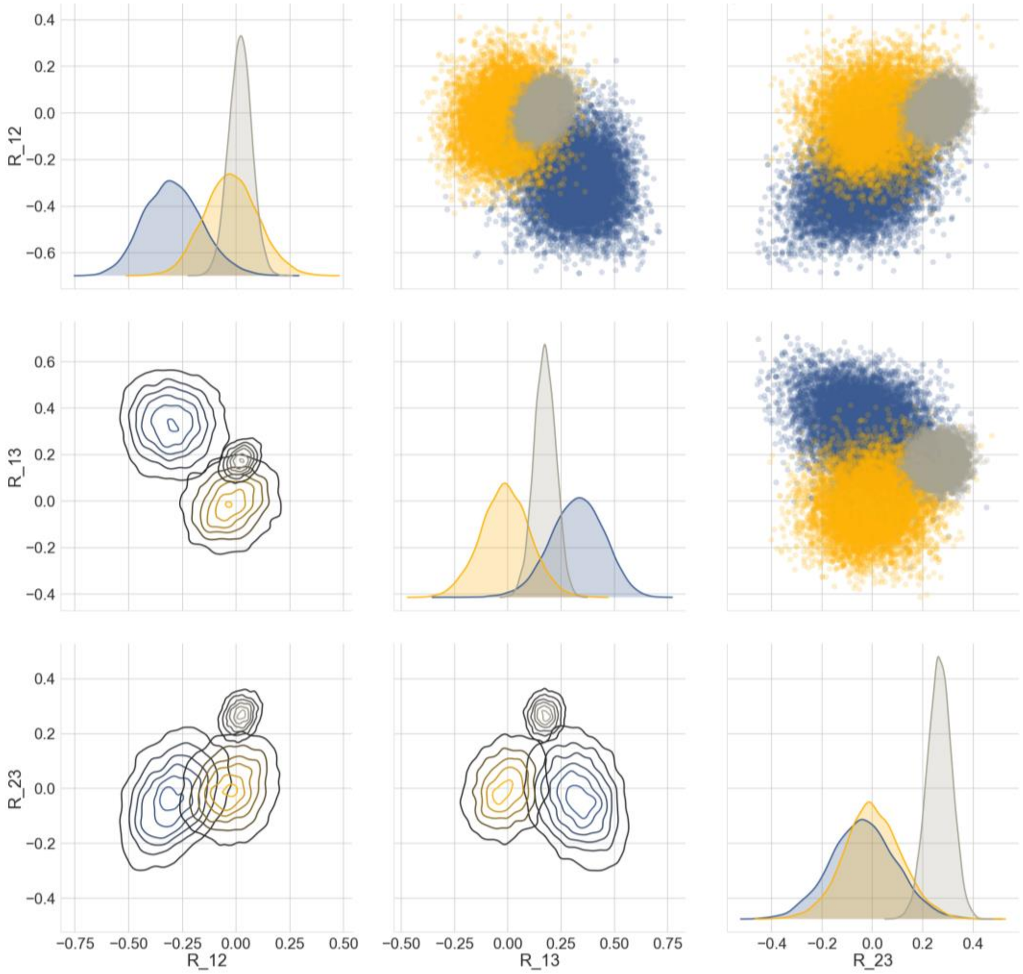


Figure 37. Pairplot matrix for NPSR and Monterey, CA showing posterior parameter beliefs for true covariance parameters. Results have been conditioned by training period; “Similar” data are indicated in blue; “Recent” data in gold; “Full” data in grey.

In this way, Figure 37 represents a multiparameter extension of Figure 32 that also communicates the sensitivity of Bayesian parameter estimation to the size and character of the training data. The blue and gold data correspond to shorter training periods (i.e., 60 days); the grey data are associated with the “Full” training periods and contains 363 multivariate forecast trials. The impact of the longer training period is indicated by sharper posterior distributions for the grey PDFs. Moreover, we observe that the central tendency of each marginal univariate distribution in the diagonal plots depends on the training period as well. The Univariate PDF for ρ_{13} suggests that the

“Full” parameter estimates represent a mixture of beliefs from the shorter periods; the grey PDF is centrally located between their corresponding blue and gold distributions. Conversely, the NPS1 model found (i.e., “learned”) that optimal ρ_{12} and ρ_{23} beliefs for the “Full” training period are generally higher (i.e., right) for larger values of the posterior beliefs obtained from the shorter training periods. The uncertainty associated with all of these estimates is indicated by the dispersion of the PDFs and, perhaps more importantly, by the spread of the associated joint scatter plots in the upper triangular panels of Figure 37. The off-diagonal panels emphasize the distinct central tendencies produced for these model parameters by each training period and, indeed, disparate confidence levels (i.e., dispersion) in the associated inference.

Figure 38 reinforces this result. It depicts the parameter trace for raw model parameter θ_4, which corresponds to the regression coefficient for maximum diurnal surface temperature predictors associated with the minimum diurnal surface temperature predictand. The running average of the posterior parameter samples generated by the Metropolis algorithm is indicated with solid, blue, gold, and grey lines. The multiparameter Markov chain was seeded so that all parameters begin at their associated OLS solutions indicated by Equation 52. However, the equilibrium distributions for the “Similar” and “Recent” training data, which are found at the end (i.e., right) of each chain, are distinct from their OLS starting locations. This indicates that the NPS BEMOS model produced a parameter estimation solution that differed from the classical OLS solution. Moreover, the differences in central tendency for each training period are evident by the equilibrium location of the running averages (i.e., solid curves). As noted with Figure 37, this result reinforces the unique posterior parameter estimates produced by each training period. The confidence of the associated inference in similarly apparent with the dispersion of the samples (i.e., filled circular markers) in Figure 38. The “Full” training period produced the narrowest parameter interval estimates. As previously noted, the convergence of the model run is partially indicated by the steady state behavior of the running mean. This suggests that posterior parameter samples collected after the burn-in threshold are representative of the target Bayesian posterior distributions sought for each model parameter.

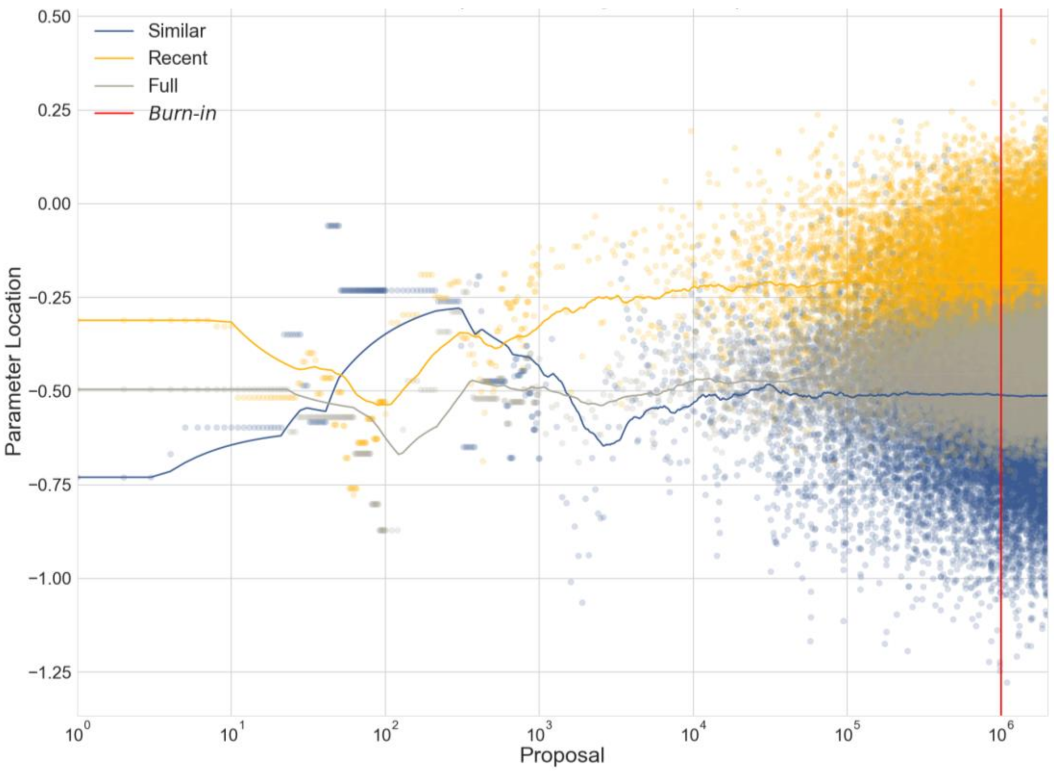


Figure 38. Semi-log trace plot for the NPS1 model at Monterey, CA. The ordinate describes the location of the parameter proposal in the phase space of the model as a function of proposal number (abscissa). The red vertical line indicates model "burn-in," which describes a pre-defined boundary between valid and invalid posterior parameter samples. The solid blue, gold, and grey horizontal lines describe the moving average of the chain for the indicated parameter.

A broader examination of the disparities in parameter estimation produced by classical OLS methods and the NPS BEMOS model instances is considered in Figure 39. It depicts the OLS margin—that is, the difference between the Bayesian and OLS parameter estimates—for each of the regression coefficients in Equation 69 associated with four statistical model perturbations from the ARW dynamical model core. The box-and-whisker plots—or, more simply, box plots—show the distribution of OLS margins for each parameter over the 21 cities and three training periods considered for each model. In this way, each box plot corresponds to 63 OLS margins for each model parameter. An OLS margin of zero corresponds to equivalence between the two parameter estimation methods; values away from zero indicate distinctions. The results from Figure 39 depict consistent disparities between NPS BEMOS and OLS parameter estimates. While the MLE1 model produced almost no margins, due to its MVN likelihood and noninformative prior, the MLE3, NPS1, and NPS3 models all produced meaningful variability from the reference OLS solutions. This was caused by the details of their JPMs—either the presence of a MVT likelihood function or hierarchical priors. NPS3 was the most different from its OLS counterpart, due to the presence of both a MVT likelihood function and hierarchical priors. Both MVT models—that is, MLE3 and NPS3, produced non-zero beta intercept parameters. This result is notable based when compared with Kruschke (2014), which posits that intercept parameters will be zero with standardized parameters. Finally, the magnitude of the OLS margins in Figure 39 should be considered non-trivial for standardized regression coefficients.

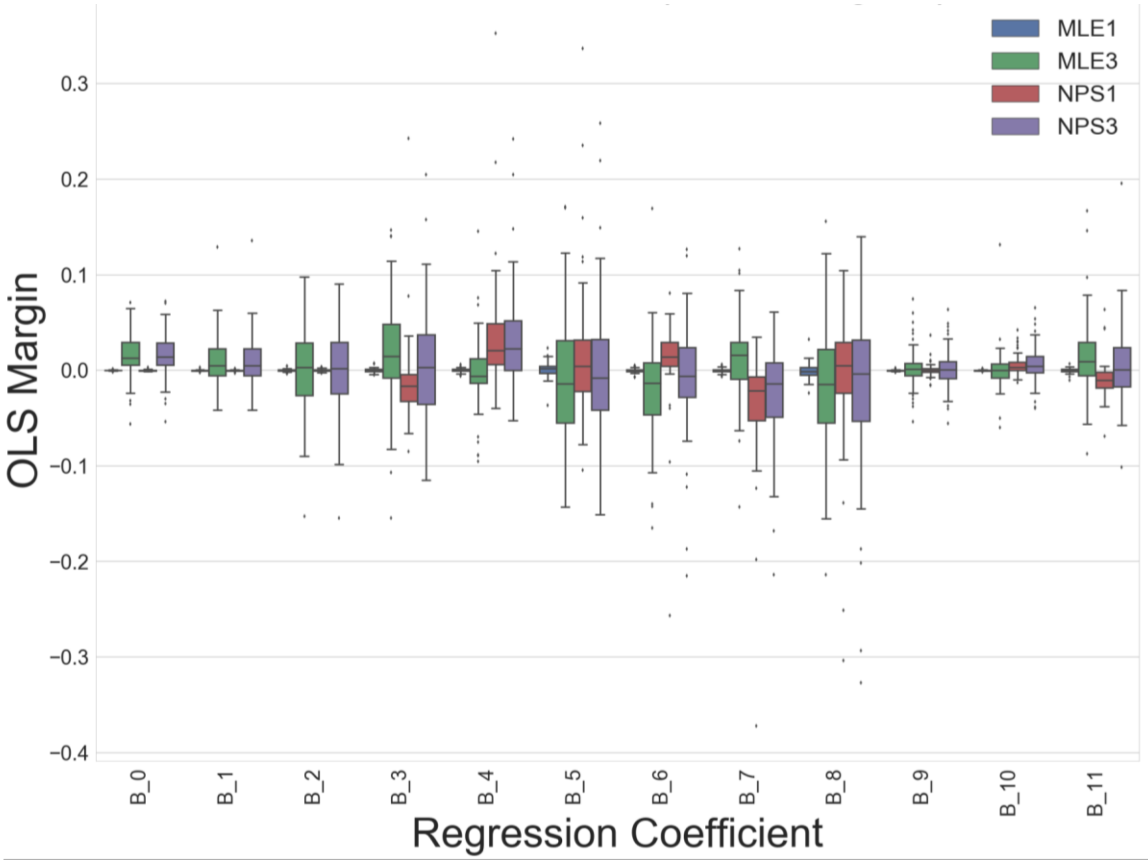


Figure 39. Box plot of OLS margins for linear regression coefficients associated with four NPS BEMOS models.

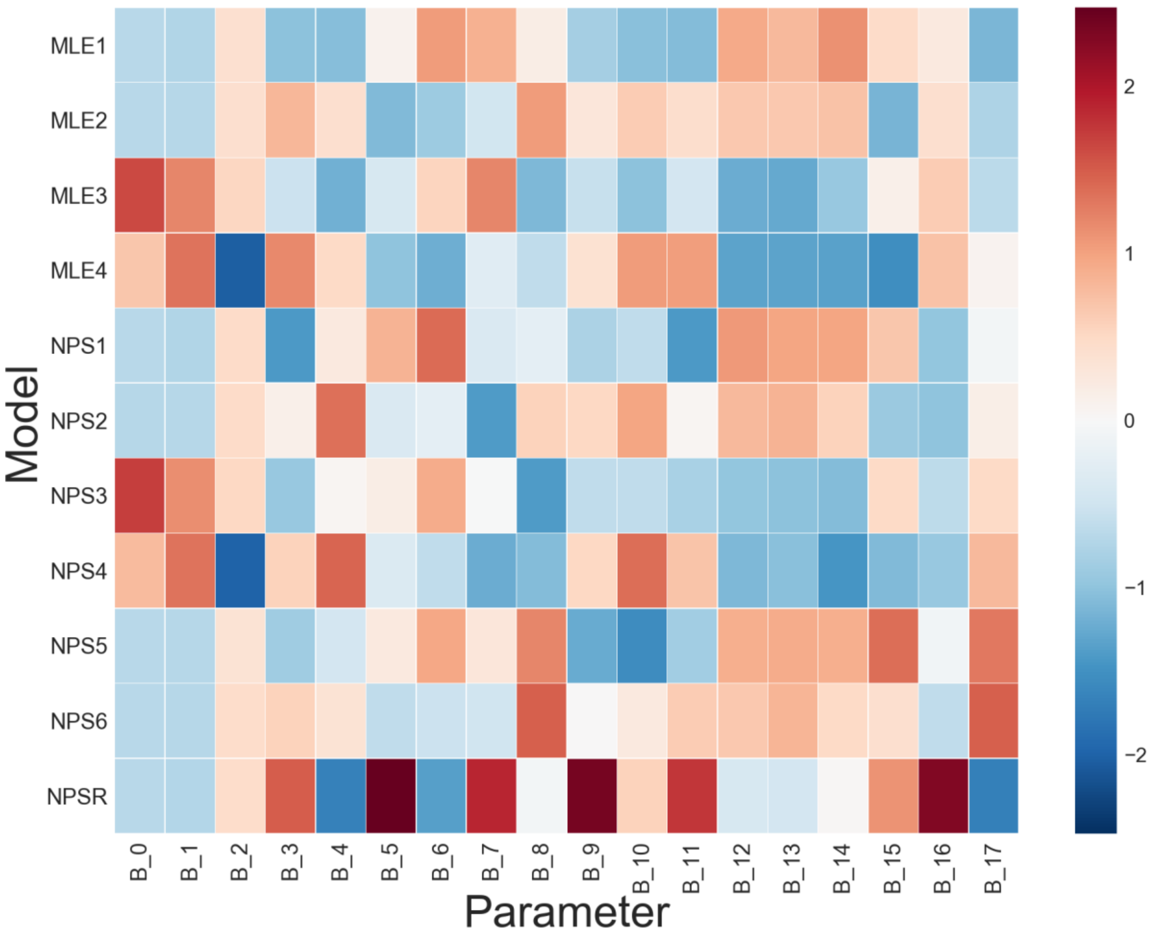


Figure 40. Mean posterior parameter beliefs averaged over all training periods, cities, and MCMC samples for each regression coefficient. Results have been standardized within each regression parameter to emphasize differences between statistical model perturbations. OLS parameter estimates have been excluded. Cooler colors indicate smaller mean posterior parameter values relative to the beta-group mean.

Figure 40 depicts a heatmap of posterior parameter means conditioned to show the sensitivity of model parameters to the JPM structures indicated in Table 2. This information has been standardized with z-scores for each parameter to produce a stronger visual contrast within each column. In this way, the color palette now indicates the magnitude of mean posterior parameter belief centered and scaled according to the group mean and standard deviation for each model parameter. As previously noted, MVT likelihood functions produced non-zero intercept coefficients. Moreover, the distinct signature of each ensemble core is evident in the differences between even and odd model instances. The ensemble mean predictors from the full SREF ensemble also produced distinct regression coefficient beliefs as evident in the bottom row of Figure 40 for NPSR.

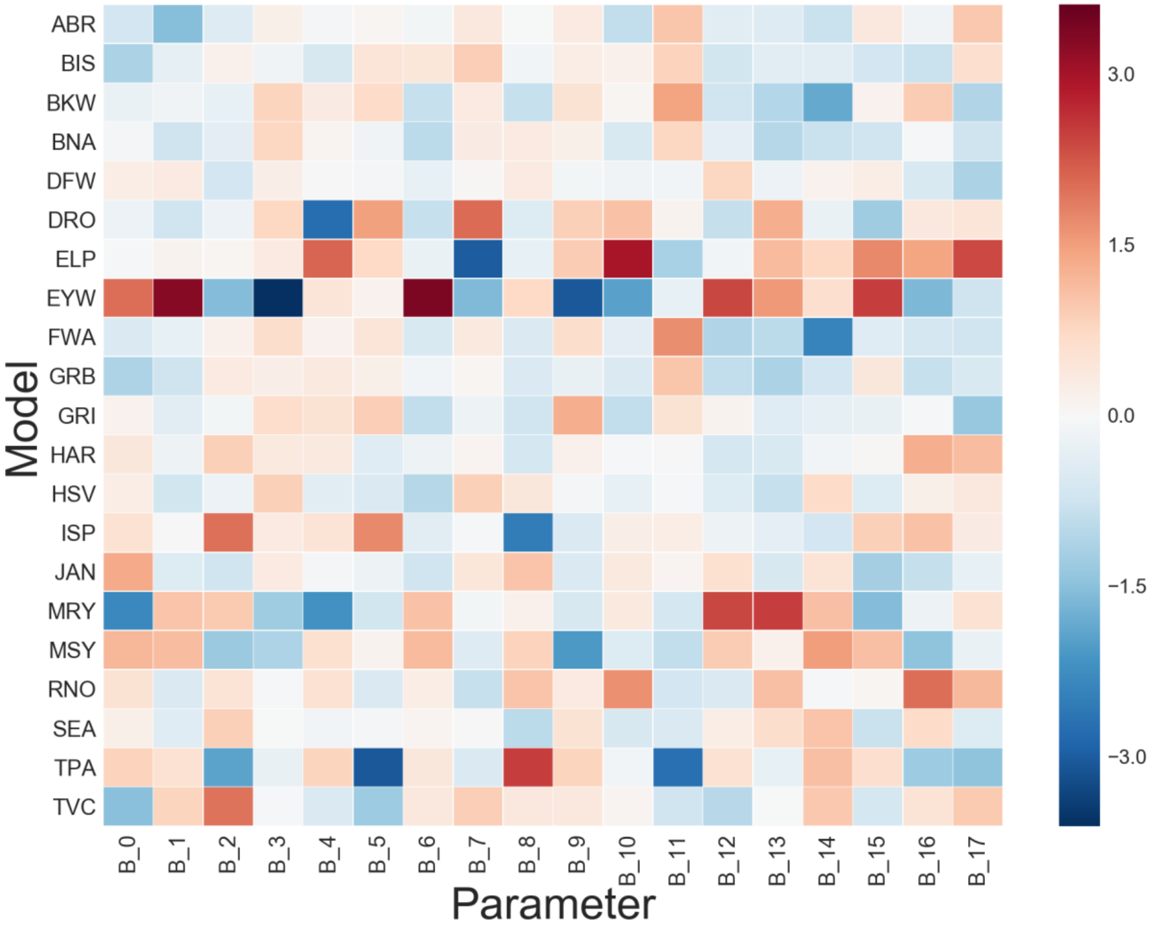


Figure 41. Mean posterior parameter beliefs averaged over all training periods, models, and MCMC samples. Results have been standardized within each beta group to clarify differences between cities. OLS parameter estimates have been excluded. Cooler colors indicate smaller mean posterior parameter values relative to the beta-group mean.

The same standardized heatmap format has been repeated in Figure 41 to express the sensitivity of parameter estimation to the data collected from each city in Table 4. The present work has trained the statistical models in Table 2 so that each city is free to produce unique regression parameter beliefs. In this way, local topography and mesoscale dynamics can be captured and combined with location-specific predictor variables. Figure 41 reveals locations that produced anomalous posteriors averaged over all models and training periods. Consistent outliers (e.g., Key West, FL) and variability within each beta group suggests that the local approach was justified. Nevertheless, the full impact of disparate posterior parameter beliefs will be assessed through forecast performance.

2. Measures-Oriented Performance

The dynamical solvers in the SREF EPS form the predictive basis of the statistical forecasts produced by the models in Table 2. However, Richter (2012) notes that ensemble perturbations (i.e., members) frequently exhibit linear correlation. To examine this for the diurnal extrema considered by the present work, the PCC from Equation 64 was applied between all members of the SREF system in the diagonal correlation matrix heatmap for diurnal maximum surface temperatures in Figure 42. The sequential colormap indicates the magnitude of covariance between respective ensemble members; warmer colors correspond to a higher relative linear correlation. Redundant covariance information in the upper triangular portion of the matrix has been masked. The top left portion of visible matrix describes the linear correlation of ARW core members; similarly, the bottom right portion of the visible matrix indicates the linear correlation between NMMB members. The covariance between dynamical solvers is represented by the primary block at the bottom left of the matrix.

As with Richter (2012), PCCs were found to be high with a minimum coefficient above 0.93. However, the ARW core demonstrated more linear correlation between its members than the NMMB core. This suggests that ARW perturbations are more correlated (i.e., less distinct) than their NMMB counterparts. Moreover, the covariance between ARW and NMMB perturbations was found to be of comparable magnitude to the covariance within NMMB members—save for the NMMB control. This is evident in Figure 42 by the block of cooler colors in the lower half of the visible matrix; the higher relative covariance of the NMMB control perturbation with other elements of its core is indicated by the vertical strip of warmer colors at the bottom-center of the visible matrix.

A similar pattern was observed for the other predictands; however, covariation between SREF perturbations for maximum diurnal surface wind speeds were found to have a larger range (i.e., Figure 43). Warmer colors in Figures 42 and 43 indicate higher relative linear correlation between respective ensemble members; cooler colors indicate lower relative linear correlation between respective ensemble members. The colorbar ranges for each figure are distinct and based on the extrema found for with each predictand. In this way, the variability in correlation coefficient between SREF EPS forecasts of maximum diurnal surface temperature was notably lower than that found for maximum diurnal surface wind speed predictions.

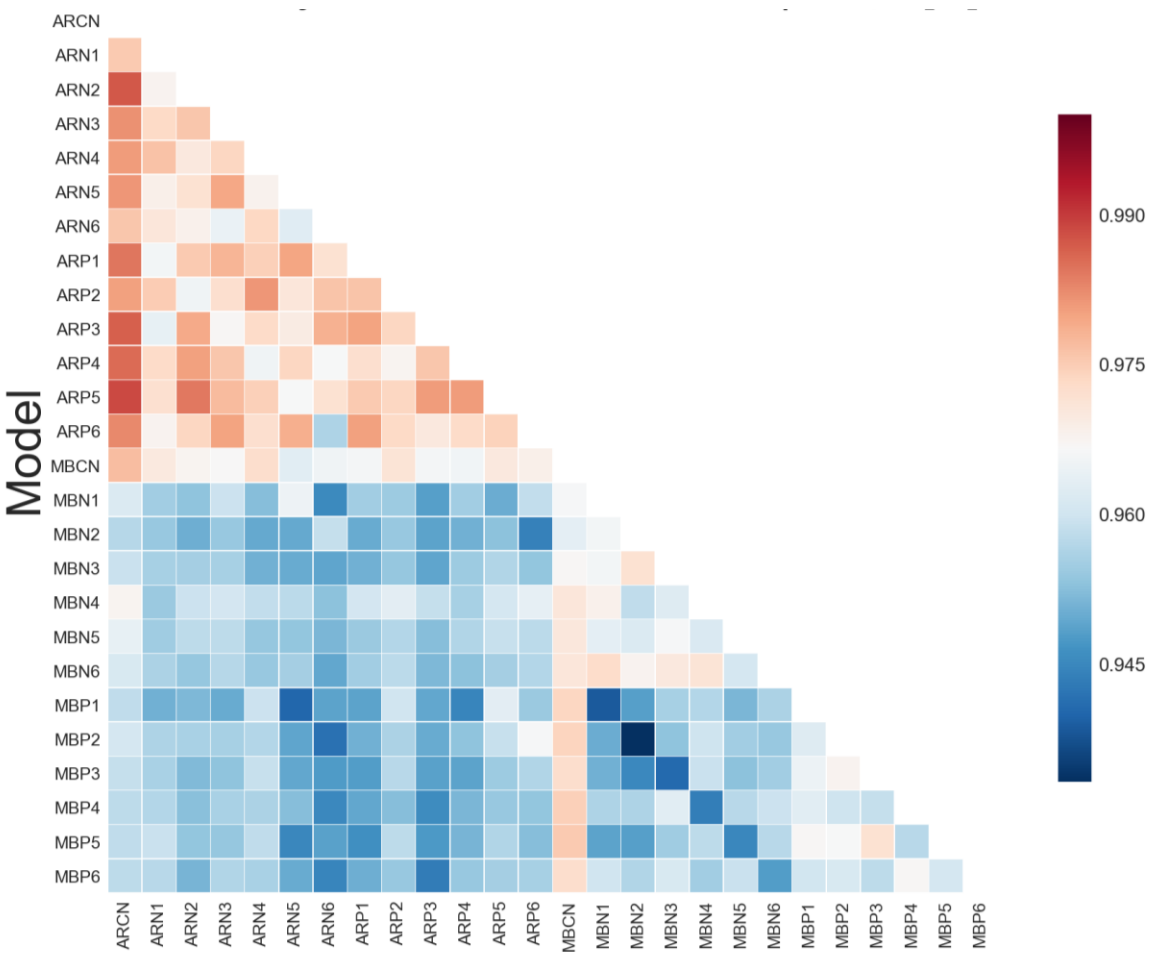


Figure 42. Diagonal correlation heatmap for diurnal maximum surface temperature predictions with the NCEP SREF EPS. Indicated values correspond to the correlation coefficient between forecast from all 26 members of the SREF system extracted during the 61-day test period.

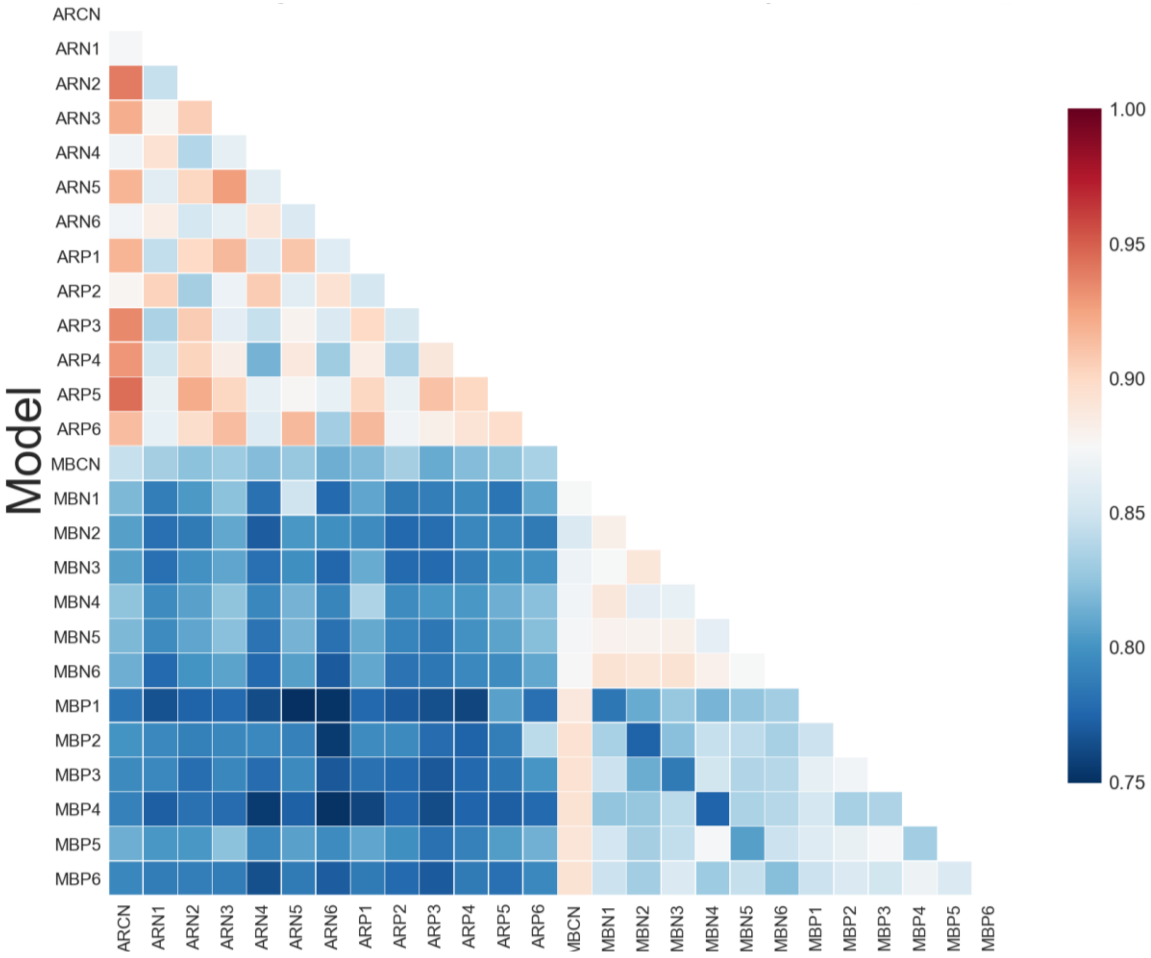


Figure 43. Diagonal correlation heatmap for diurnal maximum surface wind speed predictions with the NCEP SREF EPS. Indicated values correspond to the correlation coefficient between forecast from all 26 members of the SREF system extracted during the 61-day test period.

Figure 44 compares the covariance between the statistical forecasts models in Table 2 with the dynamical SREF system. To reduce the total number of matrix elements, only the means from each core and the full ensemble are reported (i.e., AREF, MBEF, and SREF). Linear correlations between statistical models was found to be similarly high; PCCs typically exceeded 0.9 across all predictands and model perturbations. Moreover, a larger range in covariance was observed for maximum diurnal wind speed predictions. The “checkerboard” structure observed along various matrix diagonals in Figure 44 represents the information and structure shared between model instances (i.e., model core, predictors, training period, and JPM). In this way, models using the same dynamical solver and predictor variables are likely to experience more covariance than more disparate instances. The relative covariance between the dynamical and statistical forecasts was found to be notably distinct. This is recognized from the vertical strip of cooler values found in the first three columns. This finding was repeated in all three predictands to varying degrees; however, the diurnal maximum surface wind speed results were the most pronounced and, as a result, are reported here. This covariance structure suggests that statistical model instances in Table 2 have unique predictive character—they are not superficial bias adjustments that mimic the generalized tendencies of their parent predictor variables.

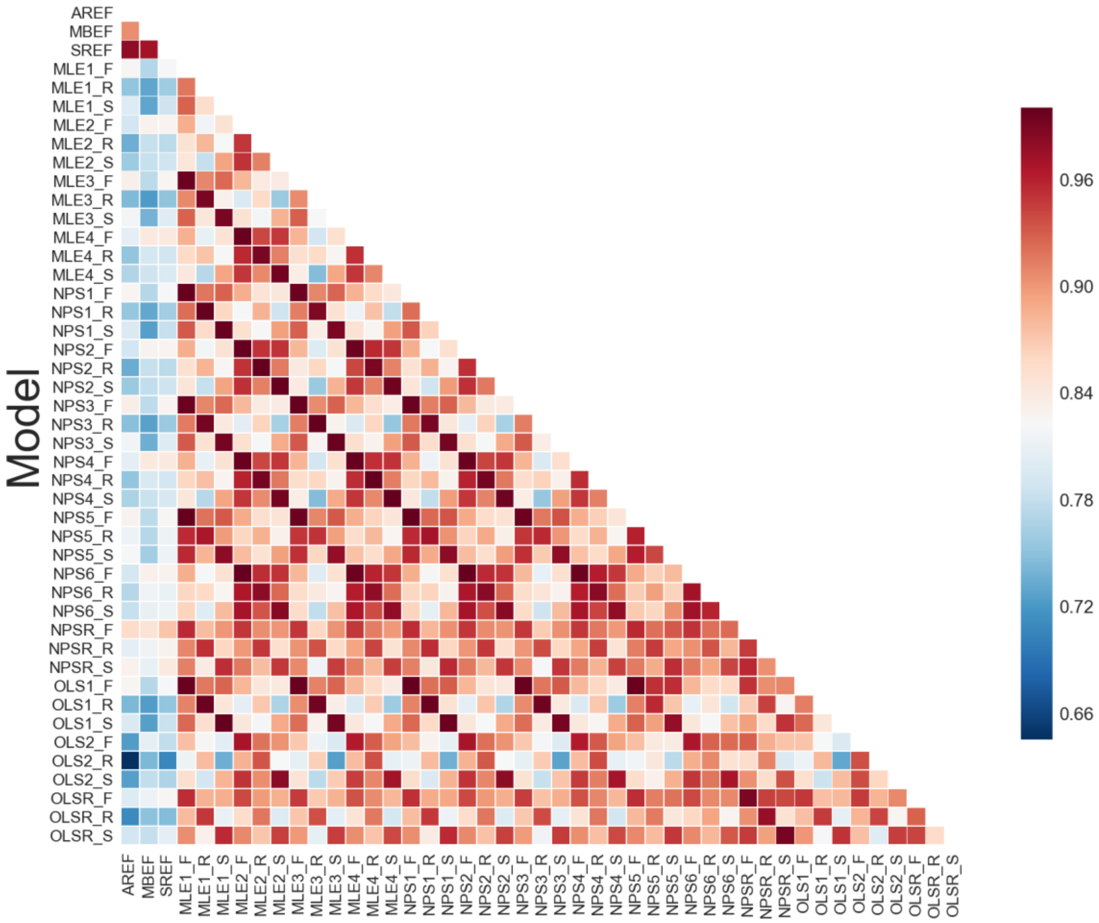


Figure 44. Diagonal correlation heatmap depicting covariance between statistical and dynamical models for diurnal maximum surface wind speed predictions. This includes the 42 model instances in Table 2 and the means of each SREF core extracted during the 61-day test period.

To evaluate the nature of the low relative correlation found between statistical and dynamical forecasts (i.e., the cooler strip of covariance in Figure 44), the typical variability observed in the former should be compared against the latter and, indeed, the observations themselves. If the statistical models in Table 2 produce forecast variability that is inconsistent with the natural variability of the sensible weather variables, then these models would be recognized to have poor resolution (i.e., in a verification sense). In this way, we desire statistical forecasts that have a standard deviation that better matches the natural variability of the observations when compared with its parent dynamical source. We also seek forecasts with higher linear correlation with the observations themselves. Taylor diagrams for each model perturbation and predictand permit a graphical comparison of these performance metrics and conveniently add the centered MSE—a modified form of the classic metric for measures-oriented forecast performance.

To this end, Figure 45 depicts a Taylor diagram for diurnal maximum surface temperature predictions over all 21 cities and 61 days in the test period. All model instances in Table 2 are compared with the ensemble mean from each core and the full ensemble (i.e., the central tendency of the members from each dynamical solver and the full ensemble). The PPD mean (median) of each MVN (MVT) model instance was used to form the relevant single-valued forecast estimates. OLS estimates were originally determined as point estimates and required no descriptive statistics. In this way, each point in Figure 45 (i.e., model number) corresponds to three dimensions of forecast performance for the indicated models. The forecast data represented in Figure 45 has also been normalized relative to σ_t—that is, the standard deviation of the observations—so that all indicated performance metrics are dimensionless. Taylor (2005) notes that each performance dimension describes some aspect of agreement between the forecasts and the observations; that is, the linear correlation, the relative variance, and the centered RMS error. The covariance (i.e., linear correlation) between model forecasts and the observations is indicated by the logarithmic azimuthal scale on the outboard arc. The radial dimension expresses the standard deviation of the forecasts normalized by the standard deviation of the observations. The radial distance from the displaced origin (black pentagon) indicates the normalized magnitude of the centered RMS error.

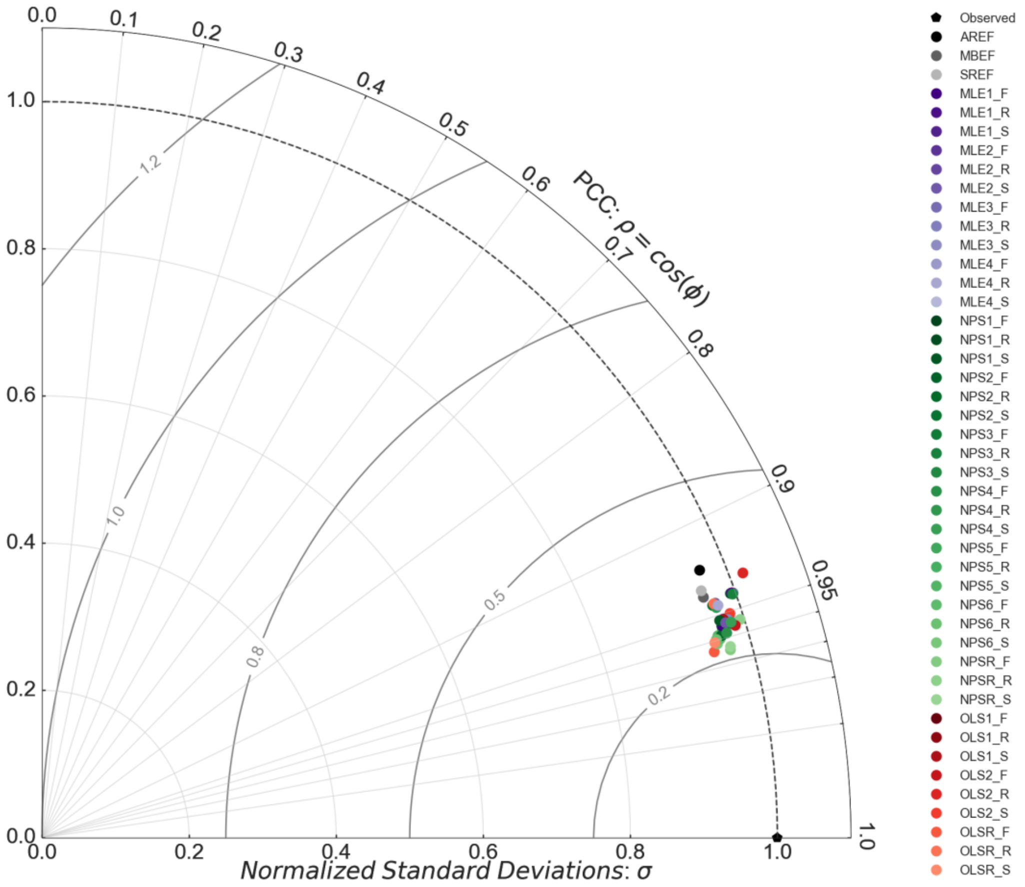


Figure 45. Taylor diagram for maximum diurnal surface temperature predictions for 45 models over 21 cities and 61 test days. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE estimates are indicated in purple; NPS estimates are in indicated in green; OLS estimates in red. The black pentagon indicates the normalized variance of the observations and the displaced origin for centered RMS error comparisons.

The tight radial clustering near $\sigma_f = 1$ in Figure 45 indicates little disparity between model forecast variance when compared with the observations. In simple terms, this means all models demonstrated meaningful forecast resolution with predictive variability that matched the observations. However, there is non-trivial spread along the azimuthal dimension. Raw ensemble output (grey) demonstrated lower linear correlation with the observations and, indeed, higher centered RMS error. The best generalized performers (i.e., the models closest to the displaced origin) were produced by OLS (red) and hierarchical Bayesian methods (green) trained with ensemble mean predictors extracted from the full SREF ensemble. However, the Bayesian models had forecast variability that was closer to the observations. In this way, the NPSR model trained over the “Full” and “Similar” periods demonstrated the best measures-oriented performance in these Taylor diagram metrics. It should be noted that all Bayesian (i.e., non-OLS) post-processing methods in Figure 45 improved upon their raw predictors; one of the OLS instances showed poorer PCC and centered RMS error performance.

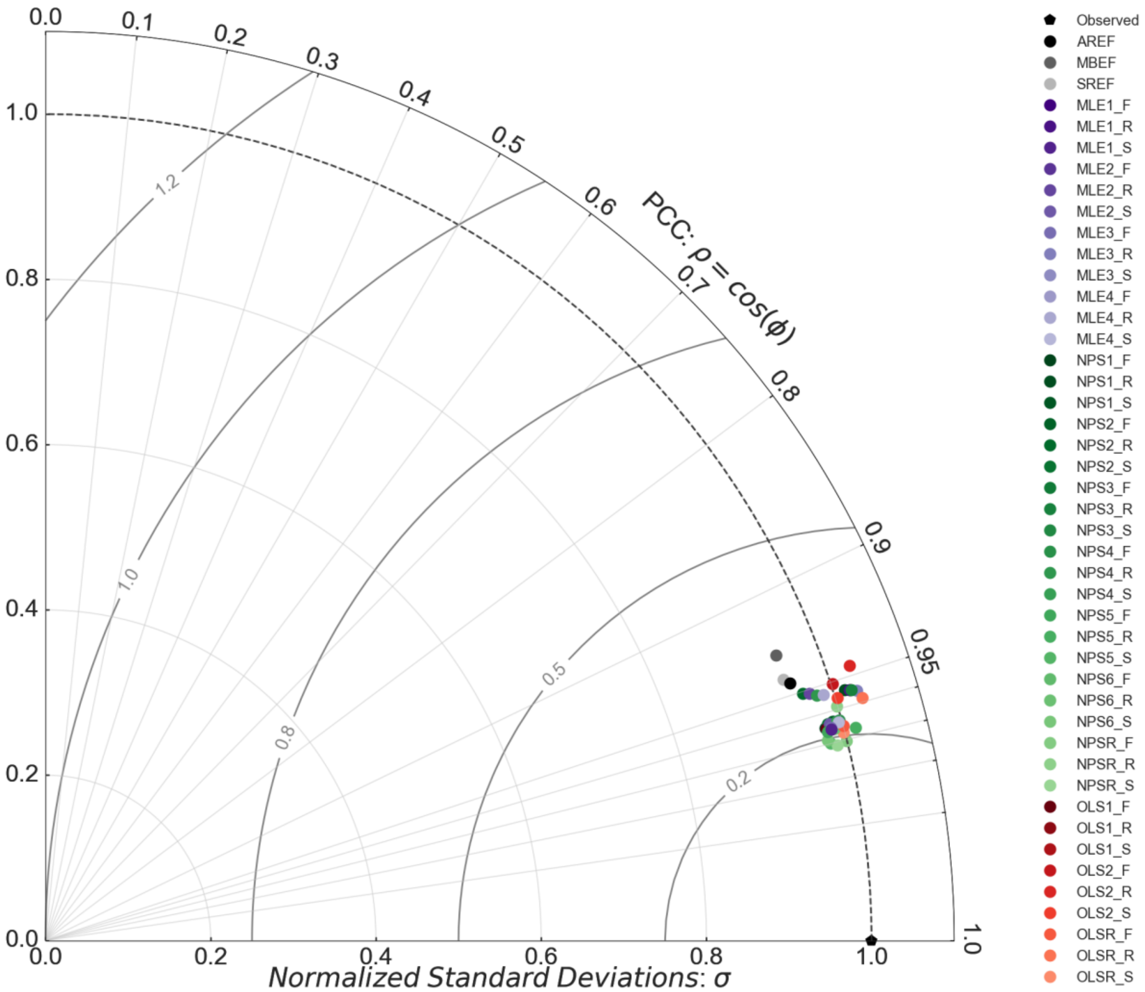


Figure 46. Taylor diagram for minimum diurnal surface temperature predictions for 45 models over 21 cities and 61 test days. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red. The black pentagon indicates the normalized variance of observations and the displaced origin for centered RMS error.

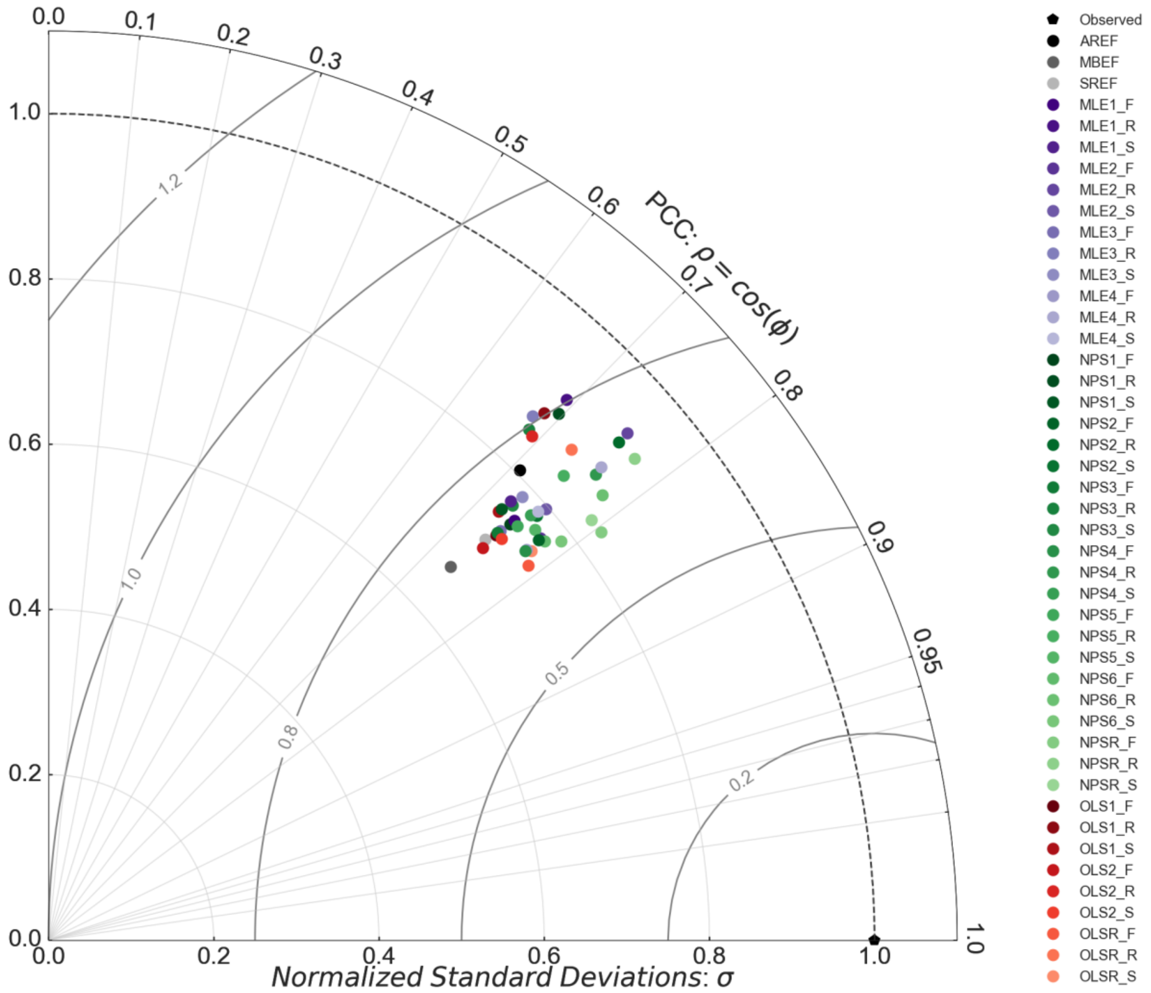


Figure 47. Taylor diagram for maximum diurnal surface wind speed predictions for 45 models over 21 cities and 61 test days. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red. The black pentagon indicates the normalized variance of observations and the displaced origin for centered RMS error.

Figure 46 indicates a similar pattern of Taylor diagram results for minimum diurnal surface temperatures. The radial distribution of forecast variability is clustered near $\sigma_f = 1$ and raw ensemble output showed the poorest performance along all three Taylor diagram dimensions. While all statistical models showed better forecast variability than the central tendency of the dynamical model cores, NPS (green) and MLE (purple) estimates showed better performance relative to the reference OLS methods (red). This is demonstrated by a tight grouping of green instances closest to the displaced origin. The

NPSR model performed well for both temperature predictands, with performance meeting or exceeding all other instances. The performance distribution for maximum diurnal surface wind speed predictions is provided in Figure 47. This predictand produced a notable increase in radial spread, which is believed to result from the natural positive skew in wind speed data. Moreover, all models demonstrated poorer normalized covariance with the observations and, as a result, higher centered RMS error. Nevertheless, the NPS BEMOS forecasts in green (NPS) and purple (MLE) demonstrated larger relative performance gains over the OLS reference solutions (red). All three of the NPSR instances performed well, with the "Full" and "Similar" training periods leading the group.

All three Taylor diagrams validate the low covariance observed between statistical and dynamical forecasts (i.e., Figure 44), because the former were found to demonstrate enhanced forecast variability and were more closely related to the observations when compared with the latter. This suggests that the unique predictive character of the statistical post-processing solutions is beneficial vis-à-vis measures-oriented forecast performance. However, these comparisons included NPS5 and NPS6, which were trained with the "April update" and had some knowledge of the true values in the full test period (Table 3). To properly consider the value of the extended training periods on NPS5 and NPS6—which are identical to NPS1 and NPS2, save for the additional training information—the test period must be constrained to the last 31 days in the interval. To this end, Figures 48, 49, and 50 repeat the information from Figures 45, 46, and 47, respectively—now with a focus on the "Recent" and "Recent-Extended" training periods. The latter has been updated to include April 2017 training data and both periods are compared with test data from May 2017 only. These updated figures provide a more granular view of the performance comparisons in the previous three images. For the maximum diurnal surface temperature predictions in Figure 48, the NPS BEMOS models in green and, to a lesser extent, purple demonstrate a performance margin relative to the majority of raw ensembles (grey) and OLS estimates (red). NPS5 and NPS6 also demonstrated better performance along the indicated metrics when compared with NPS1 and NPS2—that is, models with identical JPMs but different training data.

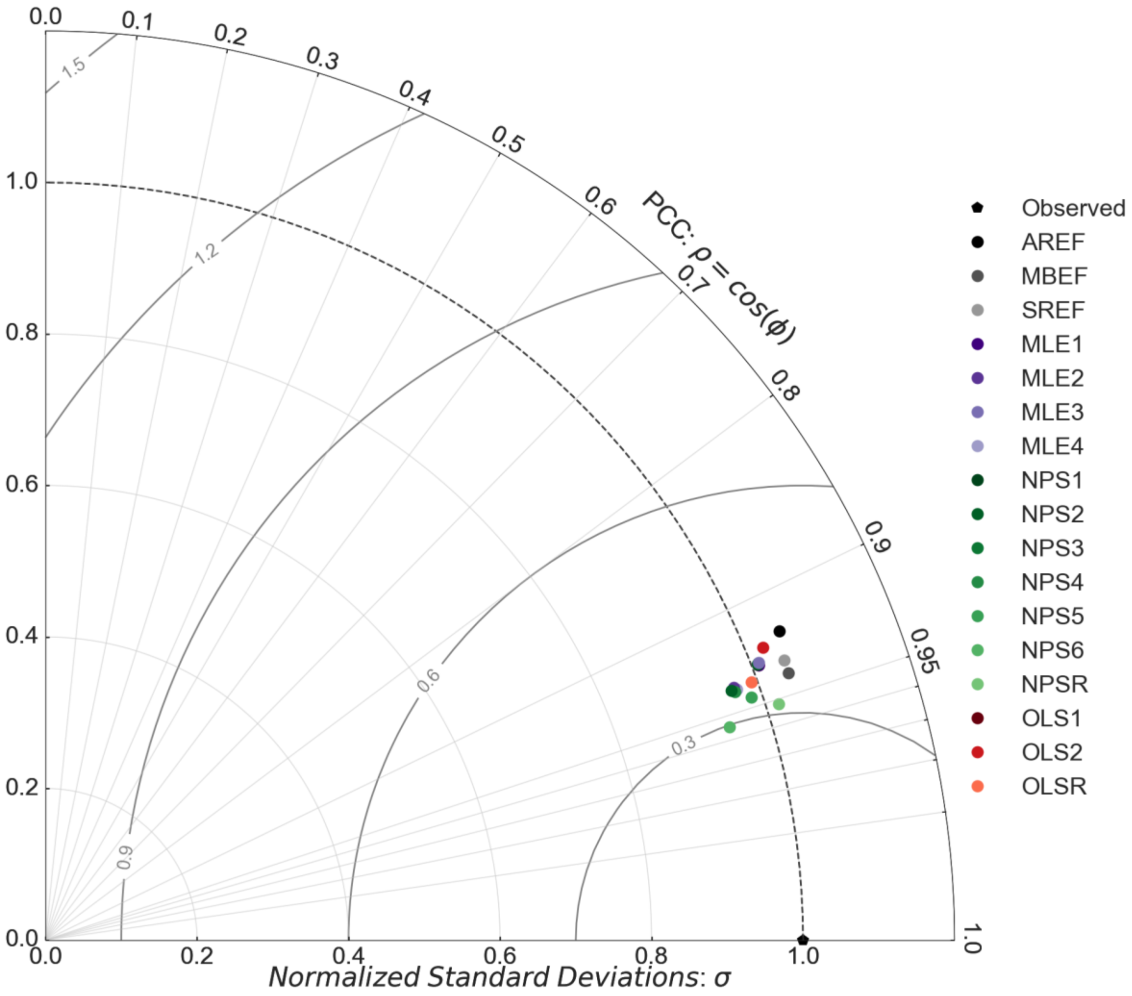


Figure 48. Taylor diagram for maximum diurnal surface temperature predictions produced by the "Recent" and "Recent-Extended" training periods. This analysis covers 17 models over 21 cities during a May 2017 test period. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red. The black pentagon indicates the normalized variance of observations and the displaced origin for centered RMS error.

Figure 49 communicates a stronger performance margin for NPS BEMOS models in diurnal minimum surface temperature predictions. NPS5 and NPS6 lead the group of 17 model instances trained on "Recent" and "Recent-Extended" data, with a majority of the Bayesian/MCMC models outperforming their OLS counterparts. The raw ensembles demonstrated poorer normalized centered RMS error compared with all statistical estimates. NPS6 and NPSR exhibited forecast variance that was remarkably consistent with the natural variability of the observations during this test period.

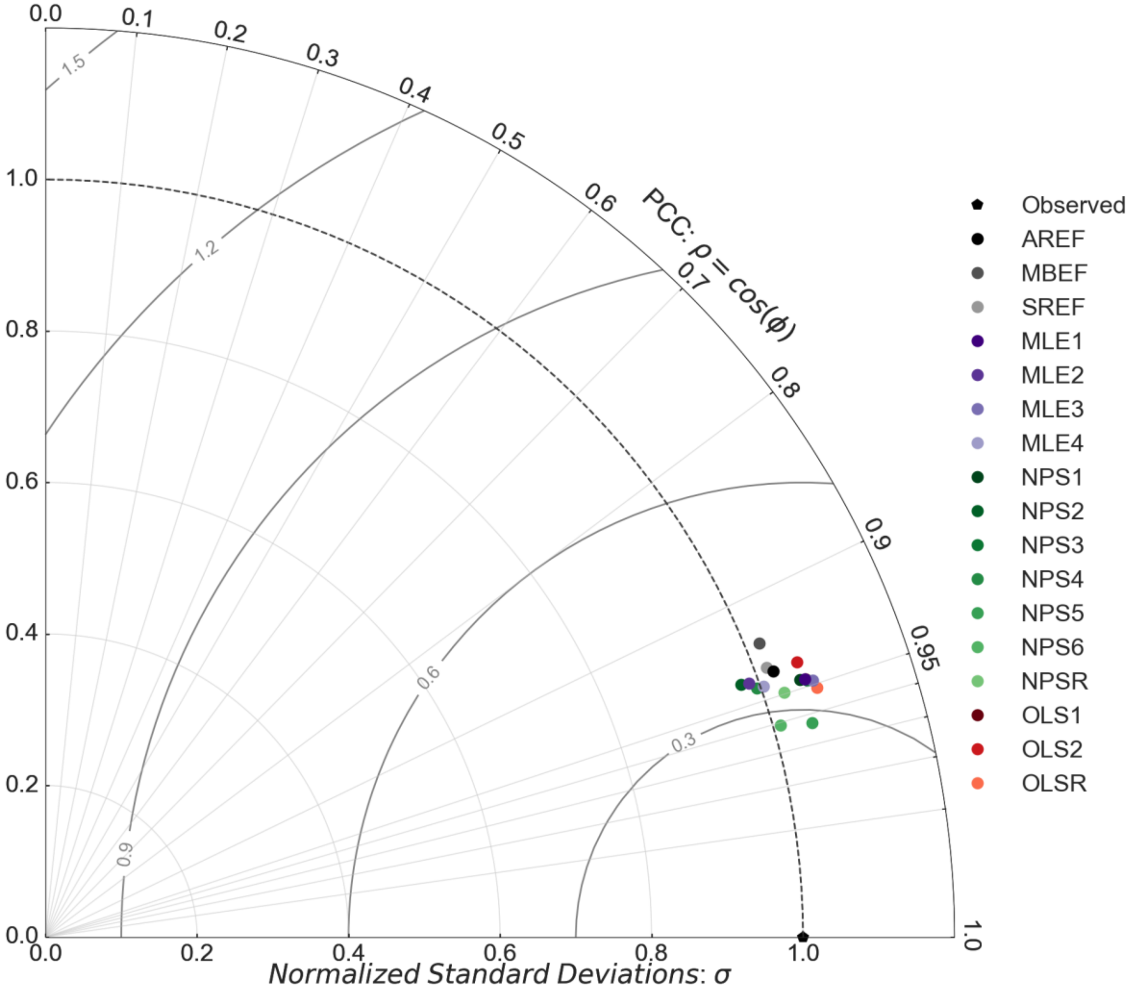


Figure 49. Taylor diagram for minimum diurnal surface temperature predictions produced by the "Recent" and "Recent-Extended" training periods. This analysis covers 17 models over 21 cities during a May 2017 test period. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red. The black pentagon indicates the normalized variance of observations and the displaced origin for centered RMS error.

Figure 50 describes the same Taylor diagram performance metrics for maximum diurnal surface wind speed predictions with "Recent" and "Recent-Extended" training data. Half of the NPS BEMOS model instances beat out all raw ensemble (grey) and OLS (red) competitors; NPS6 and NPSR were the best performers. In most cases, the Bayesian/MCMC solutions provide better forecast resolution when compared with their peers. This was especially true for the raw ensemble output, which demonstrated little more than 50% of the standard deviation of the observations in May 2017.

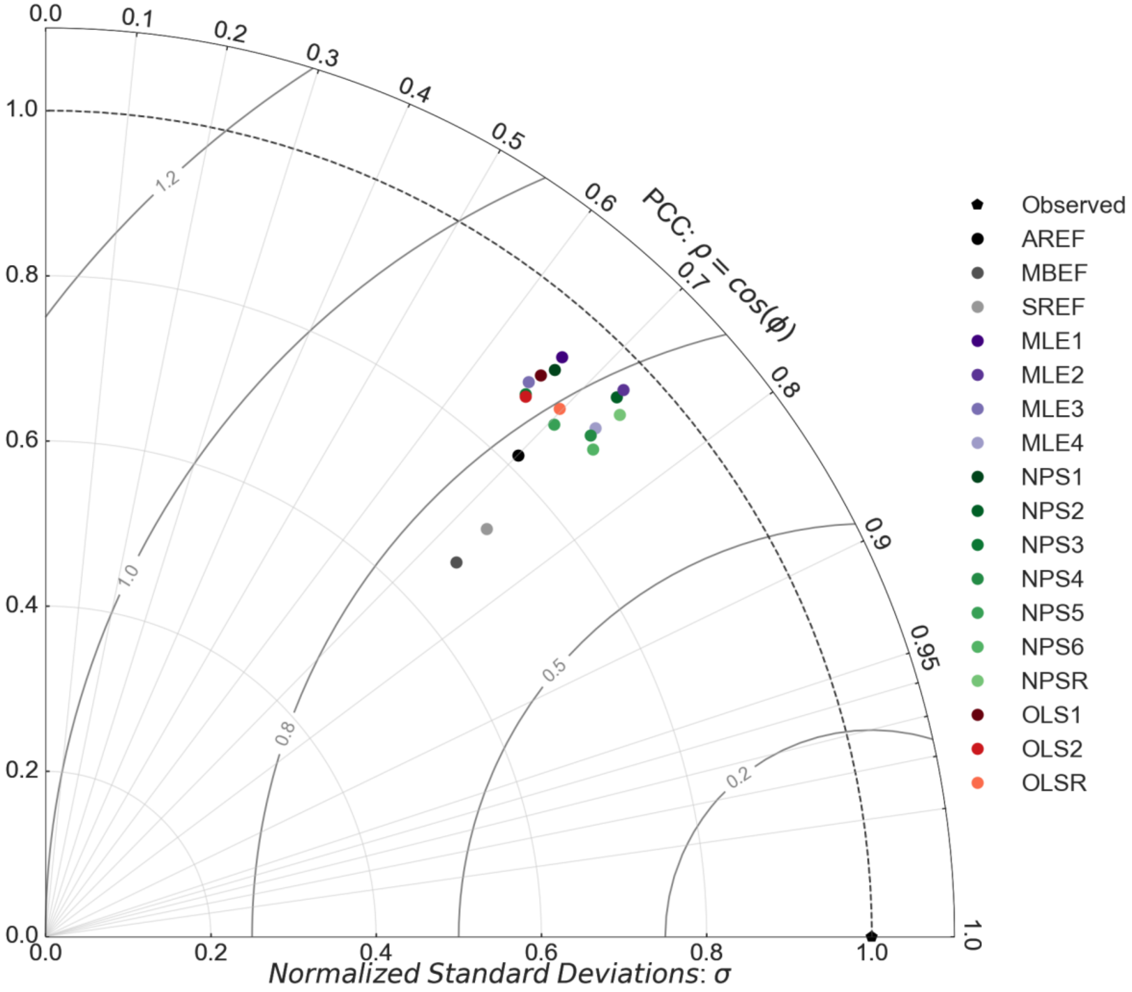


Figure 50. Taylor diagram for maximum diurnal surface wind speed predictions produced by the "Recent" and "Recent-Extended" training periods. This analysis covers 17 models over 21 cities during a May 2017 test period. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red. The black pentagon indicates the normalized variance of observations and the displaced origin for centered RMS error.

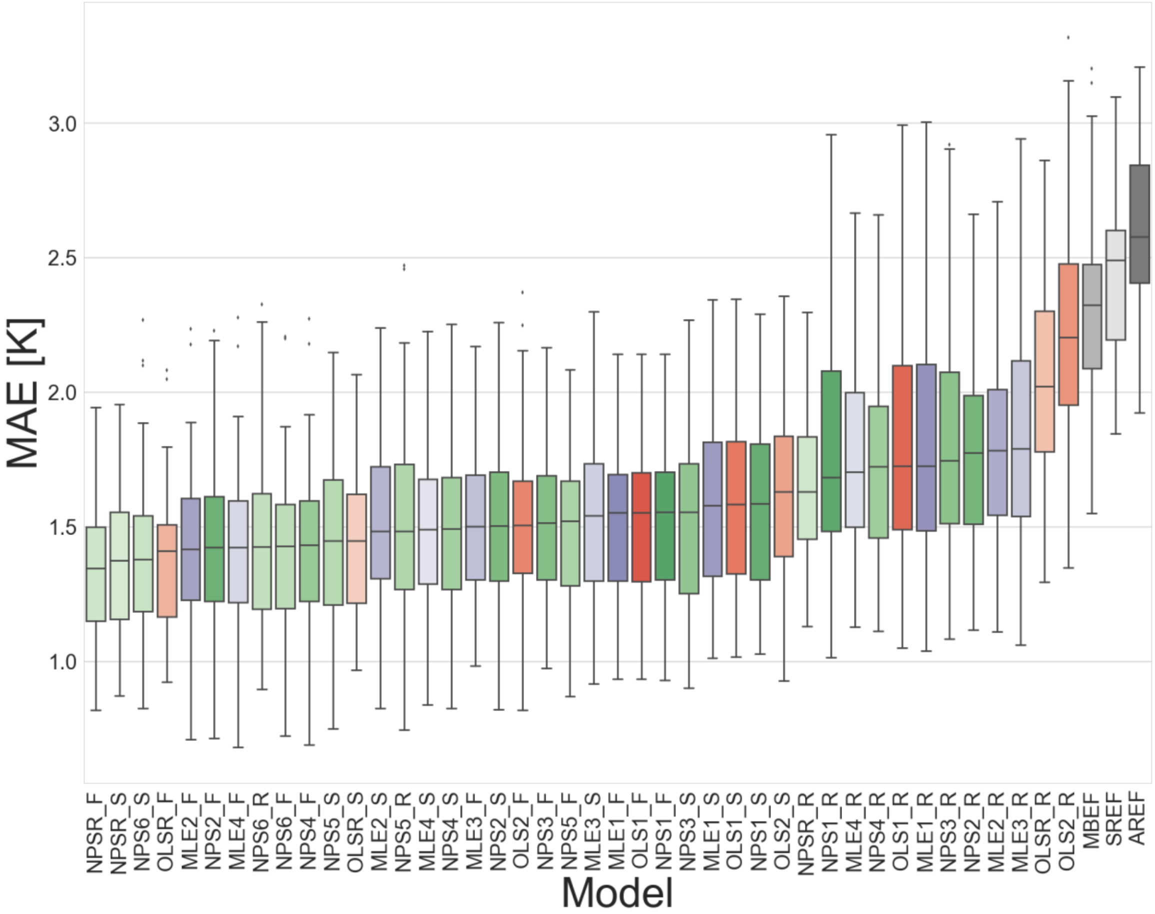


Figure 51. Box plots of inter-city MAE distributions for diurnal maximum surface temperature predictions with 45 models over 21 cities and 61 test days. Models are sorted in ascending order according to median inter-city MAE (better performance on the left). Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red.

Similar—albeit smaller—performance margins for NPS BEMOS model instances in the May 2017 test period were observed with the “Similar-Extended” and “Full-Extended” training periods. However, these figures were omitted for brevity. A summary of the measures-oriented forecast performance can be alternatively engaged through MAE comparisons. As described in Chapter III, MAE is closely related to the CRPS—that is, a strictly proper scoring rule that has become something of an unofficial standard for distributions-oriented forecast performance in the statistical post-processing community. MAE comparisons provide a convenient transition to CRPS comparisons and other scoring rules appropriate for full predictive distributions—a primary focus of this dissertation vis-à-vis ensemble calibration and forecast uncertainty.

To this end, Figure 51 depicts box plots of the MAE produced by each of the models compared in Figures 45, 46, and 47 for diurnal maximum surface temperatures. Absolute errors for each model, city, and day are averaged over the 21 cities to produce a 61-day distribution of inter-city MAEs for each model. The interquartile range and median of these MAE distributions is identified by a box and solid horizontal line; the former is bounded by the first and third quartiles. Figure 51 has been sorted in ascending order according to median inter-city MAE so that relative performance is expressed with the order of the model instances (i.e., left is better). The color palette provided in the Taylor diagrams has been reproduced in Figure 51 to provide consistent model group comparisons. NPSR instances associated with the "Full" and "Similar" training periods were the best MAE performers. However, the performance margins for this predictand are not significant and are sometimes trivial compared with their OLS counterparts. This is demonstrated by the NPSR_F and OLS_F instances; while the former had smaller first, second (i.e., the median), and third MAE quartiles, the differences are indeed small. Moreover, the performance gradient for the statistical post-processing forecasts was similarly small. Nevertheless, the poor performance of the raw ensemble is notable in Figure 51. Also evident is the comparatively poor performance of model instances trained over the "Recent" training period, which primarily occupy the right side of Figure 51. It should be noted that NPS BEMOS instances generally outperformed OLS models trained with identical data (e.g., NPS2F vs OLS2F). However, the research conjectures focused on JPM structure were not validated. NPS BEMOS models with noninformative priors (purple) sometimes outperformed hierarchical models (green); a similar pattern was sometimes observed with MVN and MVT likelihood functions.

To obtain a more granular view of the MAE performance for maximum diurnal surface temperature predictions, we can remove the dynamical models and focus on the estimates produced by statistical post-processing. To this end, Figure 52 shows the same information as in Figure 51—now conditioned by trained period to better reflect its impact on forecast performance. The "Full" training data (grey) indicate that the model perturbations explored in Table 2 show very little impact through MAE scores alone. Models trained with SREF ensemble mean predictors lead this group (e.g., NPSR and OLSR), but other comparisons yield no meaningful differences. The "Similar" training period (blue) exhibited similar patterns; NPSR_S and OLS_S lead their group with a narrow range of MAE performance distributed over the remaining model perturbations. It should be noted that the NPSR and OLSR models were able to meet, and in some cases exceed, the performance of NPS5 and NPS6—models trained with a partial knowledge of the test data. As with the Taylor diagram comparisons, however, the "Recent" training period (gold) expressed more meaningful performance variability. The OLS reference forecasts performed particularly poorly in this training period—even OLSR.

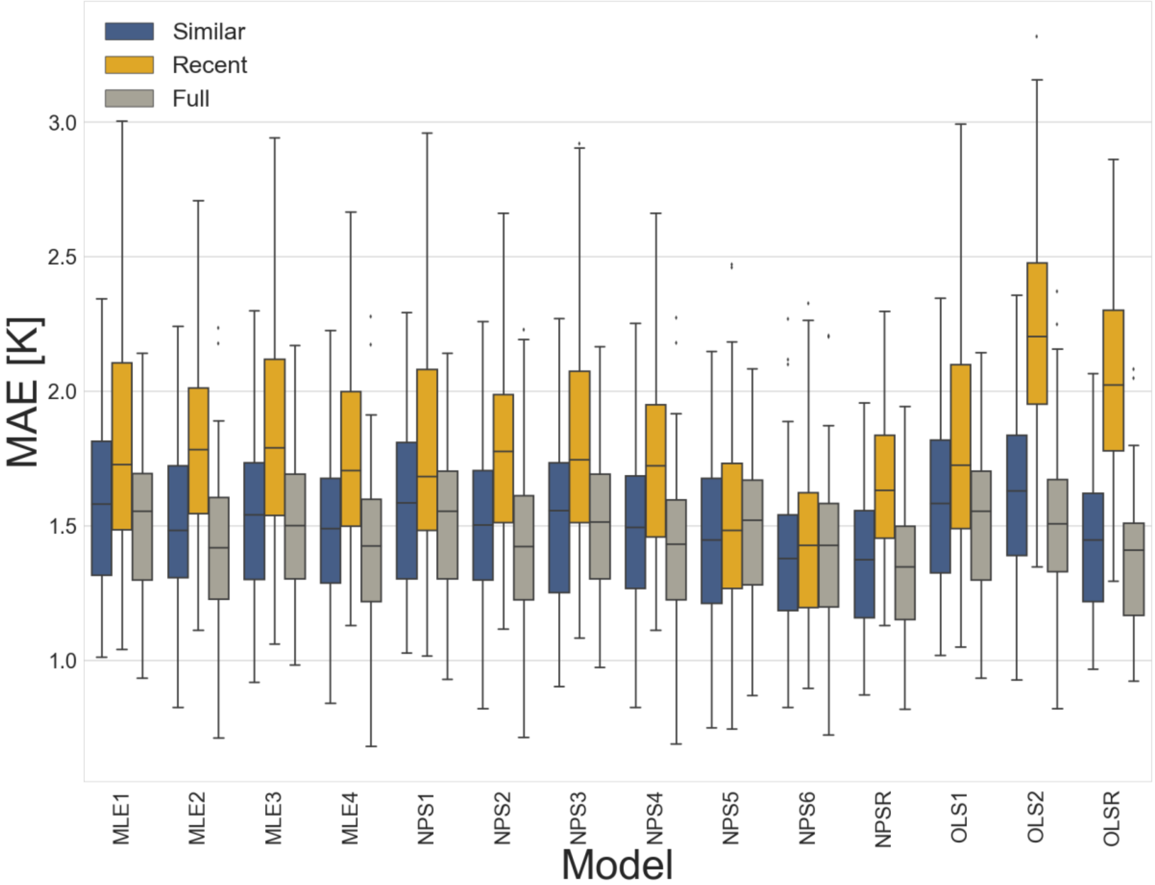


Figure 52. Box plots of inter-city MAE distributions for maximum diurnal surface temperature predictions. Performance has been conditioned by training period for 21 cities and 61 test days.

Similar patterns were observed for the remaining two predictands when conditioned by training period. Focusing again on the "Recent-Extended" training period to properly examine the impact of the April 2017 update, Figure 53 provides box plots for inter-city MAE distributions associated with maximum diurnal surface temperature predictions during the May 2017 test period. The models benefiting from the update—that is, NPS5 and NPS6—rise to the top of the MAE performance comparison. While the performance gradient is small amongst NPS BEMOS models in this case, two of the three OLS models performed poorly. Moreover, Bayesian models with hierarchical priors and MVT distributions outperformed their noninformative and MVN counterparts.

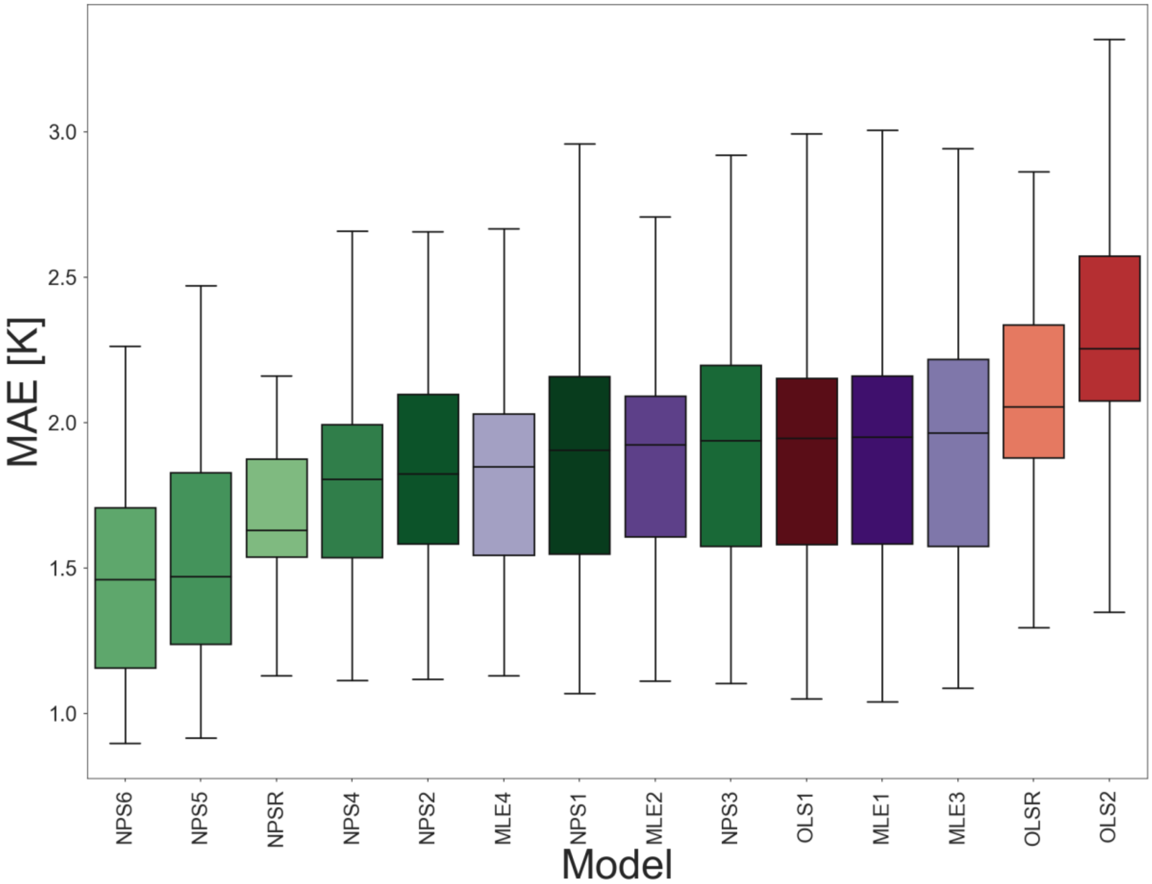


Figure 53. Box plots of inter-city MAE distributions for maximum diurnal surface temperature predictions. "Recent" and "Recent-Extended" training data for 14 model perturbations over 21 cities during the May 2017 test period are depicted. Models are sorted in ascending order according to median inter-city MAE.

Figure 54 is identical to Figure 53 but depicts MAE performance for minimum diurnal surface temperature predictions. NPS5 and NPS6 show the benefit of the updated training information relative to peer models trained with “Recent” data. However, the narrow performance range for the remaining model perturbations is evident here too. Performance differences between noninformative (purple) and hierarchical Bayesian models (green) is generally negligible. However, models with MVT likelihoods demonstrated a slight performance advantage. Only the OLS2 reference estimate performed poorly with this predictand; NPS BEMOS instances using NMMB control predictors made non-trivial improvements over the reference bias correction scheme.

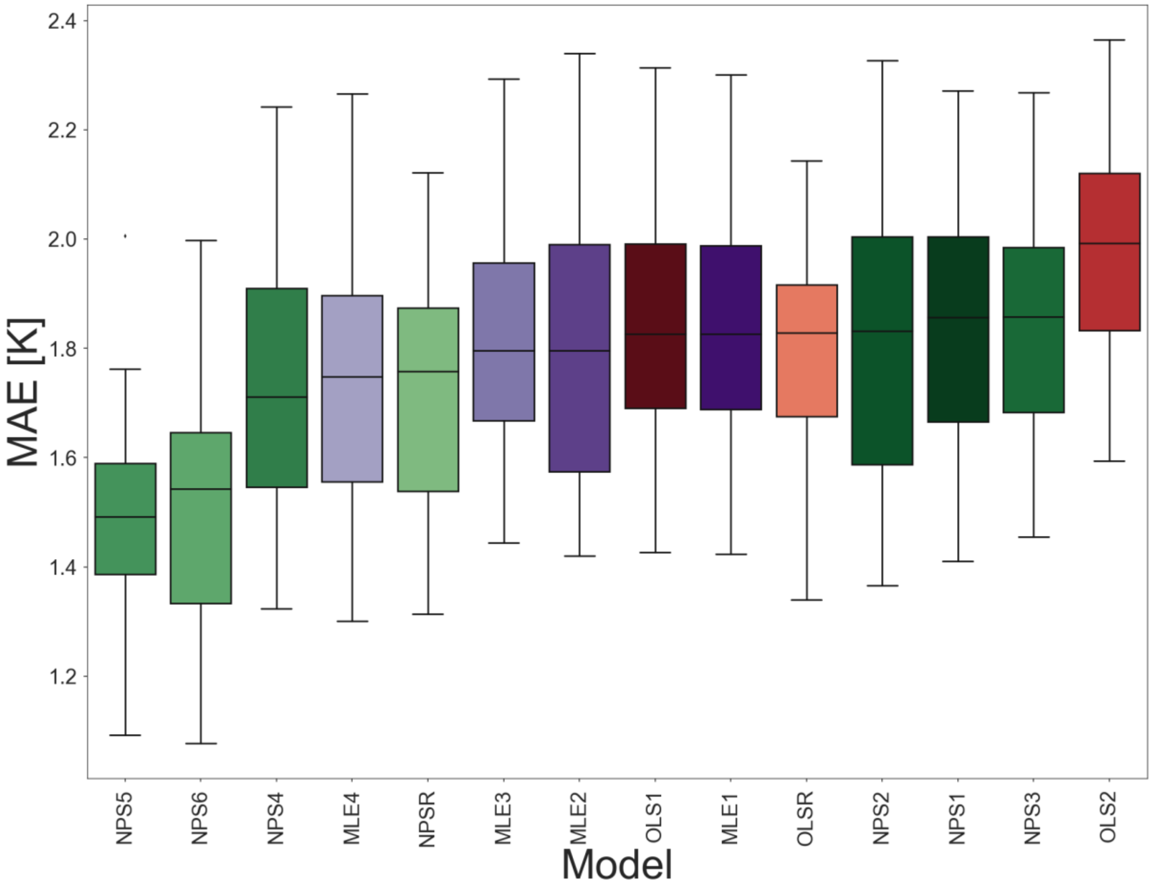


Figure 54. Box plots of inter-city MAE distributions for minimum diurnal surface temperature predictions. “Recent” and “Recent-Extended” training data for 14 model perturbations over 21 cities during the May 2017 test period are depicted. Models are sorted in ascending order according to median inter-city MAE.

Figure 55 depicts the inter-city MAE distributions for maximum diurnal surface wind speed predictions during the May 2017 test period. As with the previous two figures, NPS5 and NPS6 demonstrate the impact of April 2017 training data for “Recent” model perturbations. For this predictand, however, the April update did not provide the best MAE performance; MVT model perturbations faired marginally better than the rest of the group with a modest performance gradient delineating the bulk of the models. Hierarchical priors (green) also provided a slight performance advantage over noninformative priors (purple). All three OLS models occupy the bottom half of the group again; OLS1 and OLS2 demonstrated the poorest performance in this comparison.

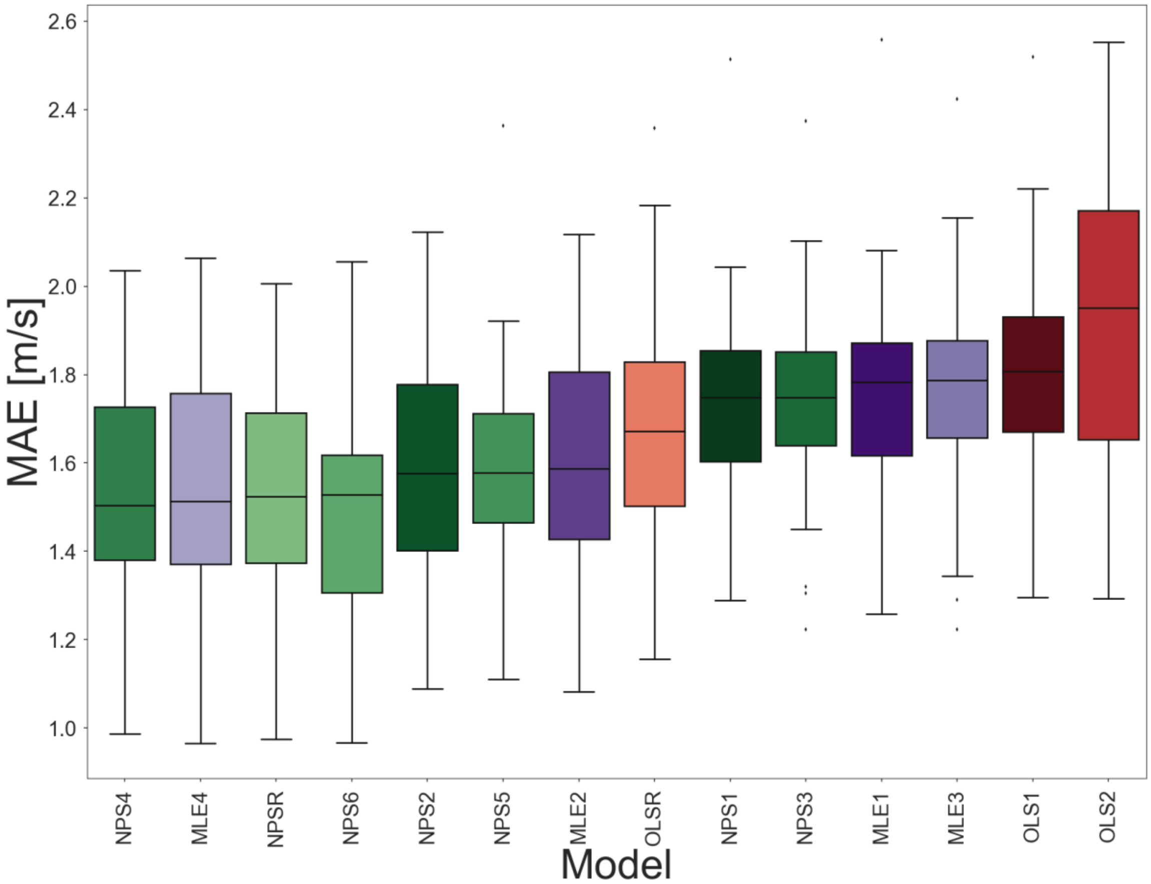


Figure 55. Box plots of inter-city MAE distributions for maximum diurnal surface wind speed predictions. “Recent” and “Recent-Extended” training data for 14 model perturbations over 21 cities during the May 2017 test period are depicted. Models are sorted in ascending order according to median inter-city MAE.

A summary of MAE performance for each model perturbation in Table 2 is reported by the heatmaps in Figures 56 and 57. Each column is formed by averaging the absolute error of each statistical model over all training periods, cities, and days for the May test period. In this way, the heatmaps provide the broadest generalization of MAE performance for each of the three predictands. Figure 56 describes this generalized MAE comparison in the original units of the sensible weather variables—that is, [K] for diurnal surface temperature extrema and $[m\ s^{-1}]$ for diurnal maximum surface wind speeds. Figure 57 reports the standardized MAE performance so that all MAE scores have been centered and scaled according to the mean and variance (i.e., z-scores) of each predictand group.

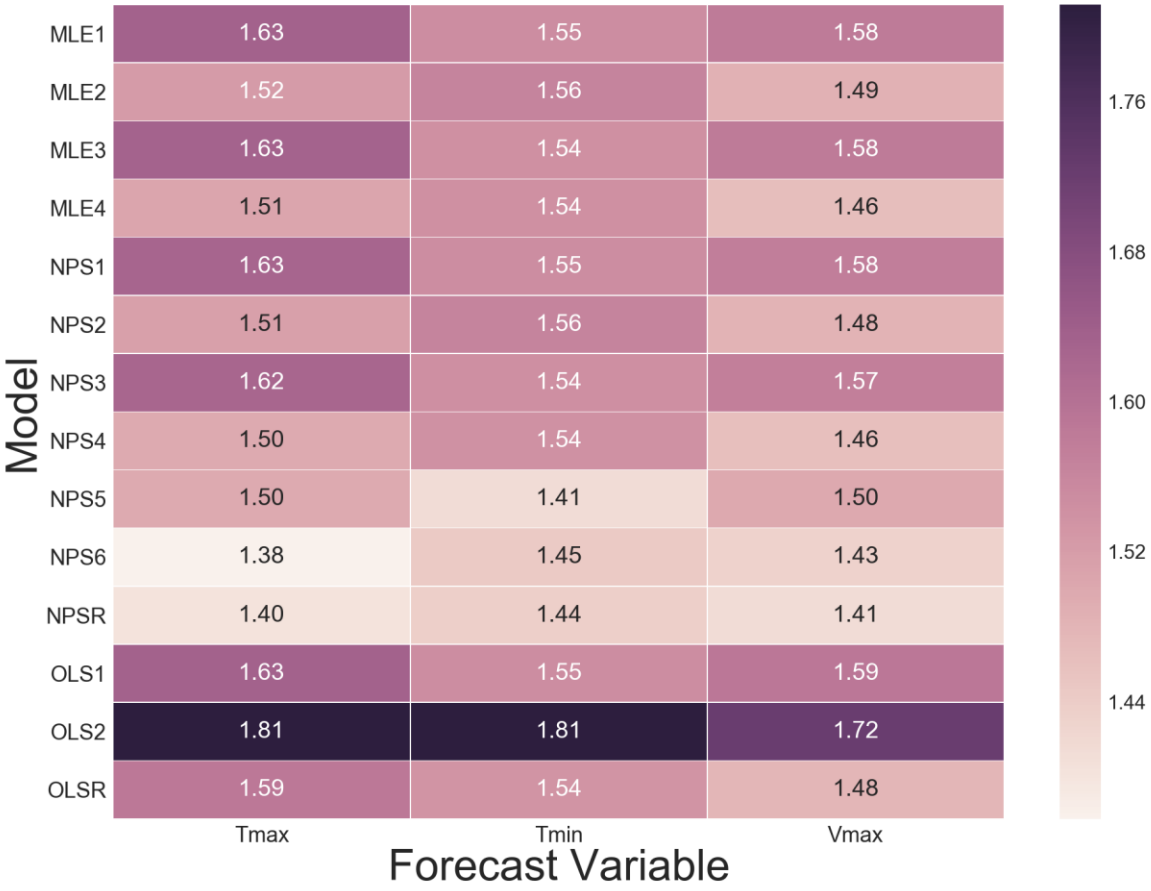


Figure 56. Heatmap of MAE for the statistical models in Tables 2 averaged over all training periods, cities, and days for the May test period.

The sequential color palette applied to Figure 56 is consistent with the MAE reported for each column in the original units of the predictand. In this way, darker colors indicated higher relative MAE (i.e., poorer performance) within each group. Figure 56 confirms that OLS 2 had the poorest MAE performance of all models over all predictands. It also indicates that NPSR, NPS6 and, to a lesser extent, NPS5 demonstrated the best relative performance. Comparisons between MVN and MVT likelihood functions (e.g., NPS1 and NPS3) showed a slight advantage for the latter. The presence of hierarchical prior information for the NPS BEMOS instances a slight impact as well. Figure 57 communicates similar findings. It shows a clear trend for relative OLS performance; with one exception, all three performed at or below group means.

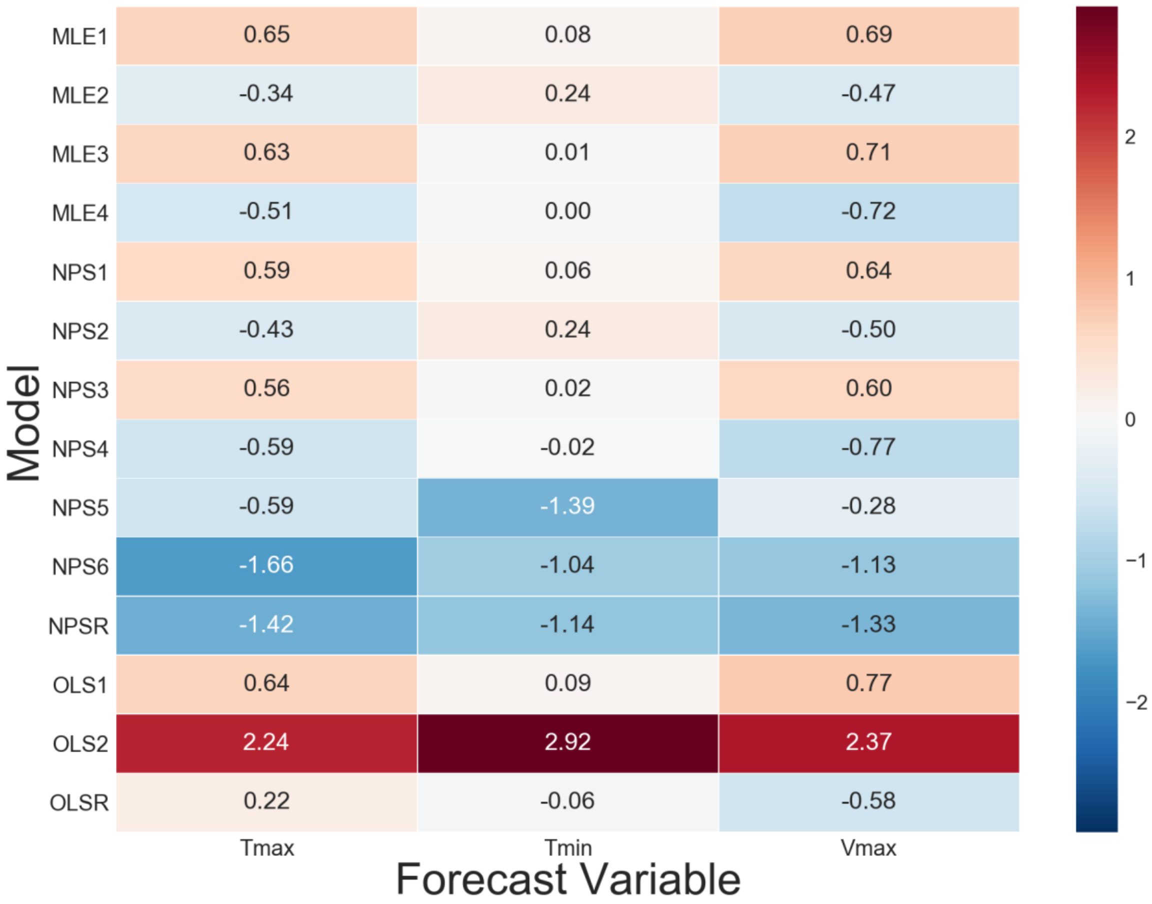


Figure 57. Heatmap of standardized MAE performance for the statistical models in Tables 2. Values are averaged over all training periods, cities, and days for the May test period.

3. Distributions-Oriented Performance

NPS BEMOS forecasts were produced according to Equation 45 by drawing 10,000 random posterior parameter samples from the thinning region of their respective Markov chains—that is, the portion of the Markov chain beyond burn-in vis-à-vis Figure 31. In this way, NPS BEMOS PPDs (e.g., Figure 25) are explicitly probabilistic and compose joint probability distributions in the 3-dimensional state space of the predictands. Nevertheless, the NPS BEMOS estimates evaluated in the previous section were derived from descriptive statistics extracted from the full Bayesian PPDs. Figure 58 recovers the full 3-dimensional probability structure and decomposes the pairwise relationships along each dimension of the multidimensional predictand—that is, a pairplot matrix. Univariate PPDs for each predictand are indicated on diagonal elements. Pairwise scatter plots of the joint posterior predictive samples are depicted in the upper triangular elements. Two-dimensional KDE is represented on the lower triangular elements and describes the probabilistic density associated with each joint scatterplot comparison. Finally, dashed black lines indicate true values for each predictand.

While representing only one PPD for the NPSR_F model instance over the full 61-day test period, Figure 58 describes the probabilistic relationships between each predictive dimension and, indeed, a complete Bayesian probability statement for each variable. Positive skew with the diurnal maximum surface wind speed is also evident, and this structure gives rise to the covariance observed in off-diagonal elements that include this variable. The marginal probability densities produced by two-dimensional KDE techniques are further examined in Figure 59 and correspond to the forecast indicated in Figure 25. In this way, the lower diagonal elements of the pairplot matrix can be seen as the normalized density of the upper triangular scatter plots; higher densities yield brighter colors in Figure 59 and describe forecast regions with more probabilistic mass. The observed values (dashed lines) indicated in Figure 58 and 59 suggest that the joint PPD structure is consistent with the observations; they could be considered plausible IID samples from each marginal PPD in Figure 58. However, more forecast trials are needed to properly assess the performance and calibration of thee distributions.

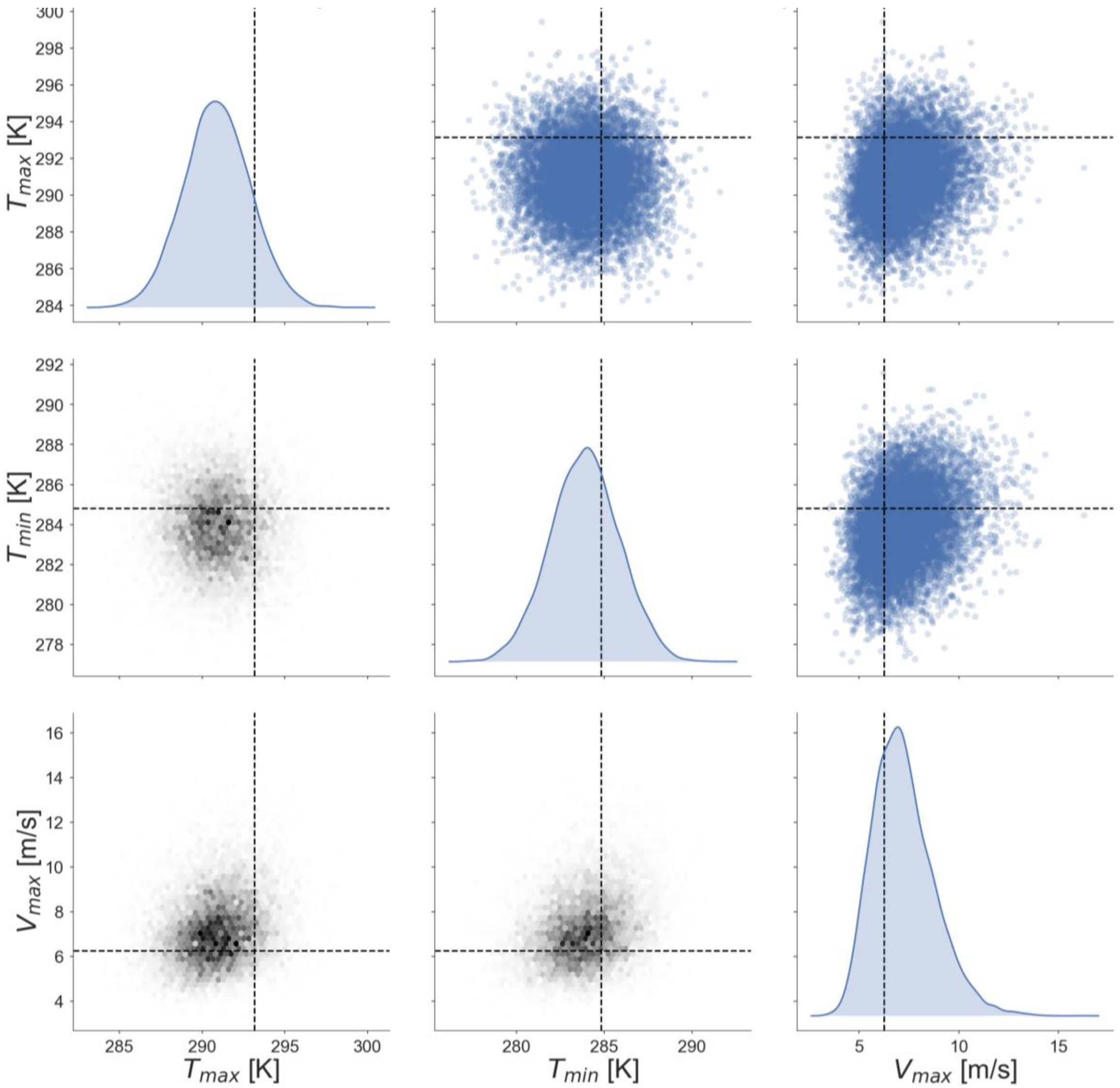


Figure 58. Pairplot matrix for the NPSR model with "Full" training data corresponding to a PPD valid May 31, 2017 for Monterey, CA. The 3-dimensional structure of the PPDs are decomposed along each predictive dimension. Univariate PPDs for each predictand are depicted as marginal distributions on diagonal panels. Joint scatter plots of PPD samples for each pairwise relationship are depicted in the upper triangular panels. Two-dimensional KDE estimates of the joint probability distribution for each pairwise relationship are depicted in lower triangular panels. Truth is indicated with dashed black lines.

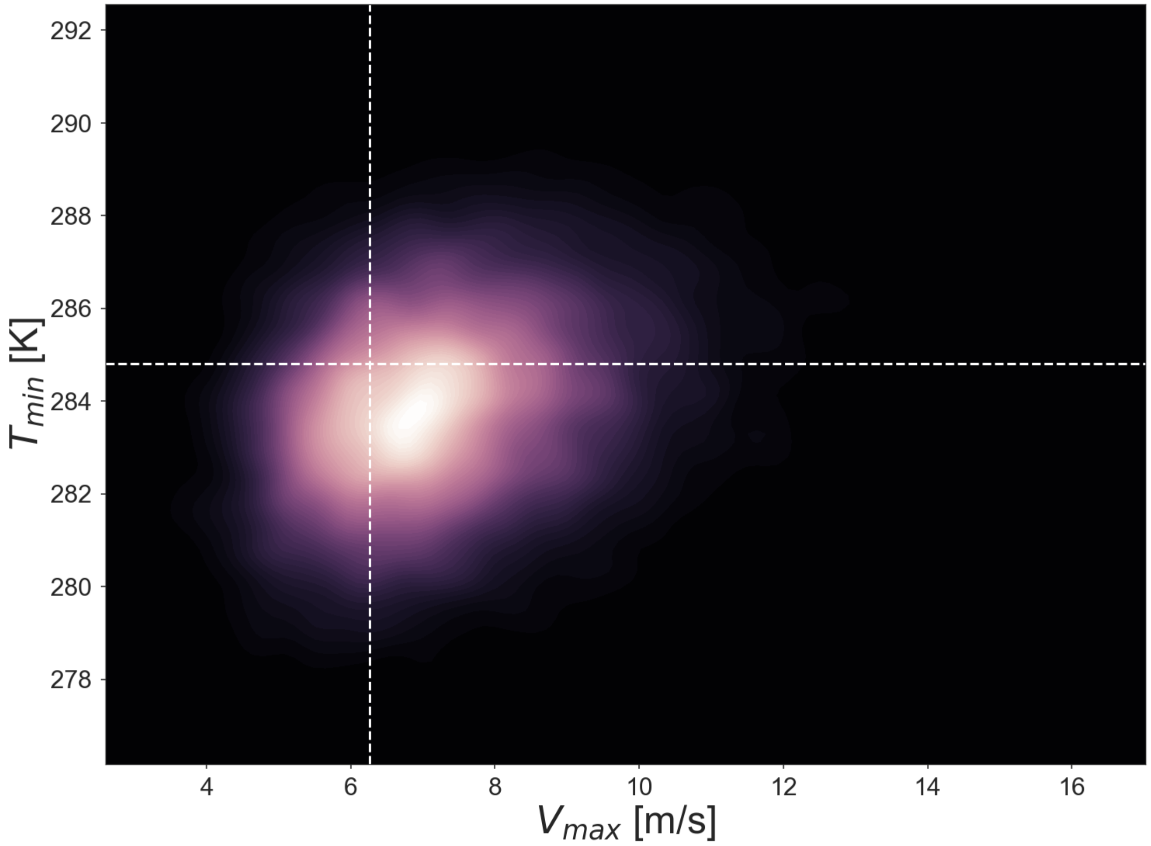


Figure 59. Joint PPD density for minimum diurnal surface temperature and maximum diurnal surface wind speed for the NPSR model with "Full" training data. Forecast corresponds to May 31, 2017 for Monterey, CA. Brighter colors represent higher probability densities. Dashed white lines indicate the true values observed for this forecast trial.

The MAE comparisons in the previous section can be extended for full predictive distributions with the CRPS according to Equation 68. To provide a fair comparison with Bayesian PPDs (i.e., NPS BEMOS instances), OLS forecasts have been dressed according to Equation 52 and 54. More specifically, the present work produced suitable estimates for the mean and variance with the OLS/frequentists methods described in Chapter III. Once these distributional parameters were determined for each forecast trial, 10,000 forecast samples were drawn from a MVN normal distribution according to Equation 56. Raw SREF output was dressed in a similar manner. Each dynamical solver and the full ensemble produced a sparsely populated ensemble of discrete forecast estimates for each forecast trial in the test period. The mean and variance of each distribution (i.e., AREF, MBEF, and SREF) was computed and 10,000 multidimensional samples were subsequently drawn from a corresponding MVN distribution for each test day. These discrete forecast estimates formed equivalent SREF and OLS forecast distributions that can be scored with the same functions as NPS BEMOS PPDs. In simple terms, this approach formed SREF PPDs so that they would retain the original spread of their ensembles; OLS PPDs have been post-processed with frequentist uncertainty information according to traditional methods of regression modeling.

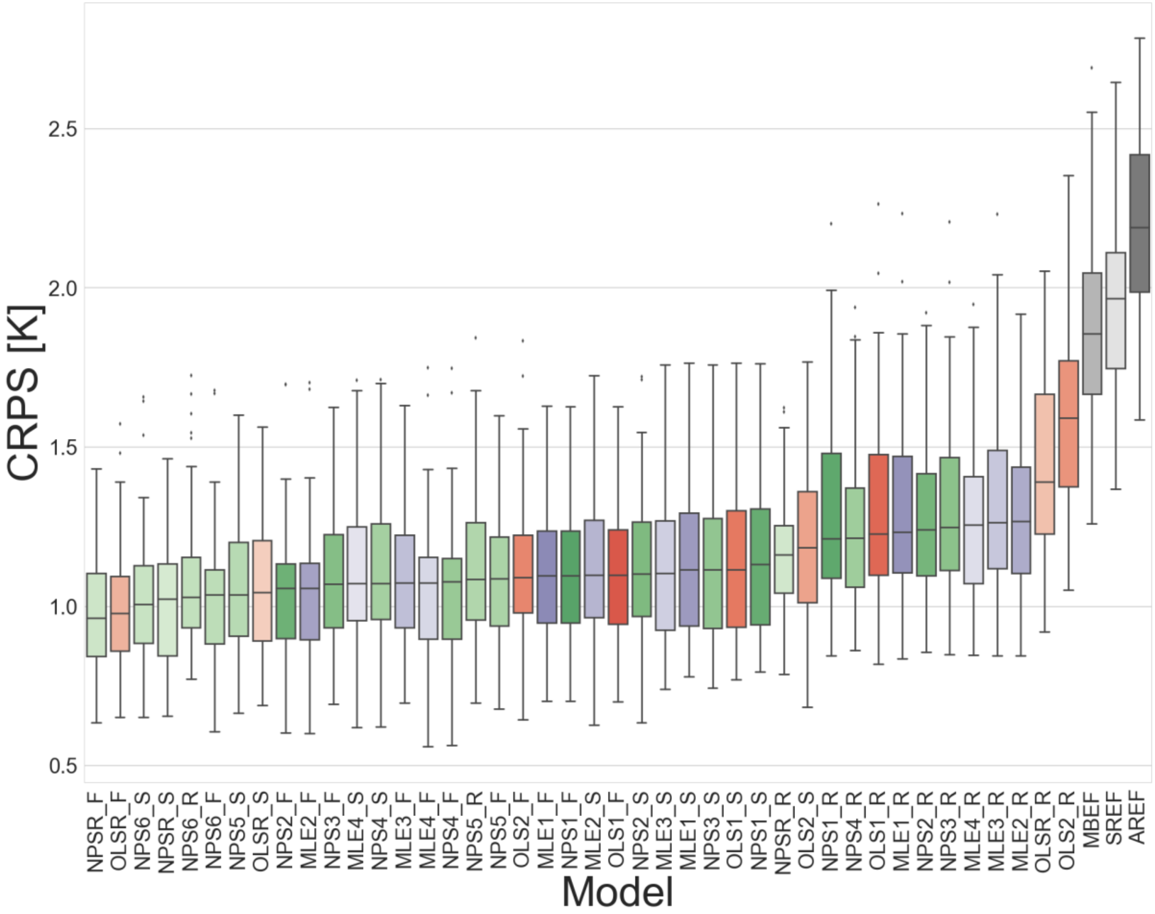


Figure 60. Box plots of inter-city CRPS distributions for diurnal maximum surface temperature predictions. Results correspond to 45 models evaluated over 21 cities and 61 test days. Models are sorted in ascending order according to median inter-city CRPS. Raw dynamical predictions extracted from the SREF EPS are indicated in grey; MLE in purple; NPS in green; and OLS in red.

To this end, Figure 60 shows a full distributions-oriented comparison of diurnal maximum surface temperature forecast performance with the CRPS for the dynamical and statistical models originally considered in Figure 51. The CRPS for each model, city, and day are averaged over the 21 cities to produce a 61-day distribution of inter-city CRPS for each model. Lower CRPS indicates better predictive performance; values are reported in the original units of the predictands—that is, [K] for diurnal temperature extrema and $[m\ s^{-1}]$ for diurnal maximum surface wind speeds. A modest performance gradient is indicated by the median CRPS of each model with "Recent" model perturbations and the raw dynamical output indicating the worst performance over the full 61-day test period, which are located on the right side of Figure 60. Reference models OLS1 and OLS2 performed similarly to the comparisons in Figure 51 for analogous MAE scores. Conversely, the "Full" instances tended to provide the best performance with NSPR_F and OLSR_F leading the group. Both models beat equivalent NPS5 and NPS6 instances, which trained with data from the April 2017 test period. NPS BEMOS models generally outperformed OLS peers trained with similar predictors.

As with the previous MAE comparisons, the dynamical models were found to be outliers for poor CRPS performance with all three predictands. A more granular view of the statistical model performance can be obtained by conditioned the remaining set of 42-models from Table 2 according to training period as indicated by Figure 61. These box plots mirror the form of Figure 52 and show the notable disparities in forecast performance between the training periods for maximum diurnal surface temperature predictions. "Recent" training data (gold) again provided the poorest performance; however, it also expressed more differences between the perturbations and notably provided the largest performance margin for NPS BEMOS models. In this way, the Bayesian approach to parameter estimation appears to provide a greater return on modeling investment with less training data—especially when the training data is less consistent with the test period (e.g., the "Recent" versus "Similar" performance comparison).

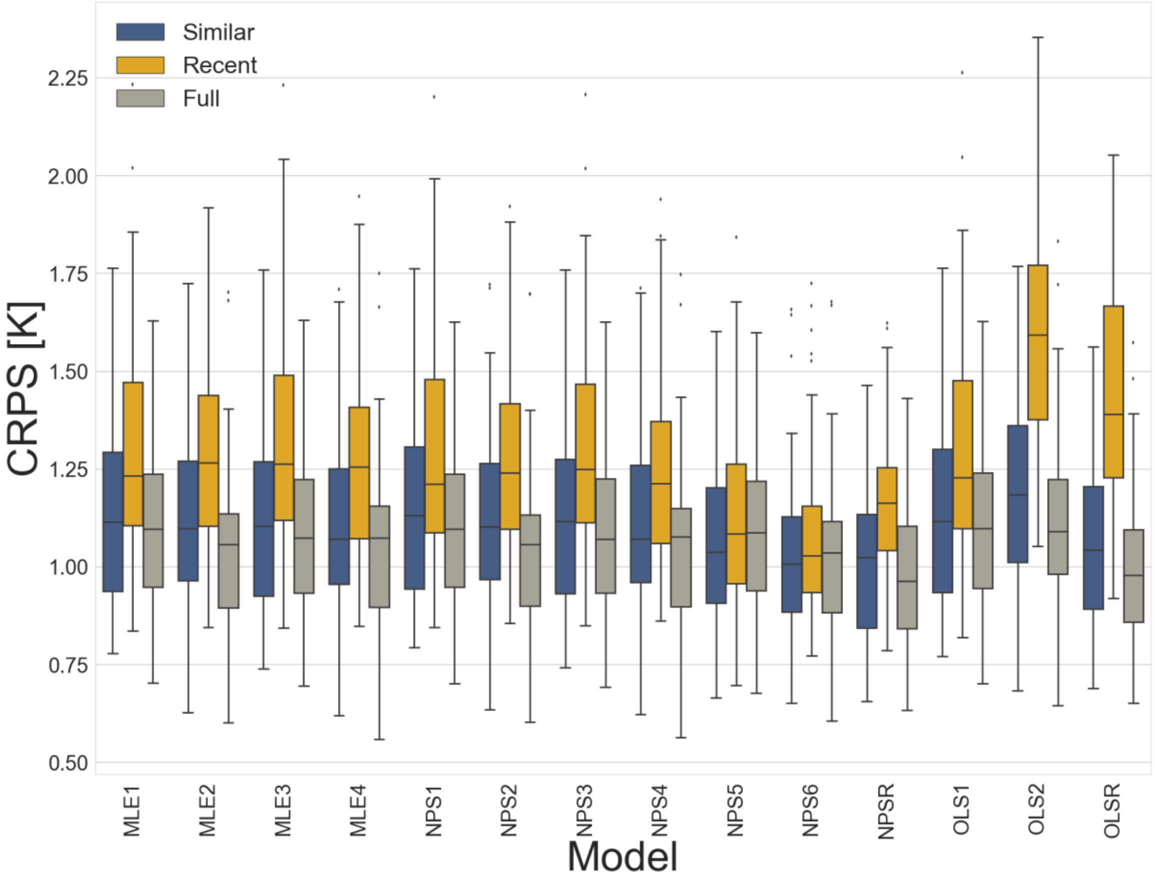


Figure 61. Box plots of inter-city CRPS performance distributions for maximum diurnal surface temperature predictions. Data are conditioned by training period for 42 model perturbations over 21 cities and 61 test days.

It is notable that NPSR performance generally matched or exceeded that of NPS5 and NPR6, both of which have been trained with data from the April 2017 test period. This suggests that the influence of SREF ensemble mean predictors was highly beneficial to performance, for both SREF and OLSR models, and in some cases was more impactful than including test data in the training period. To provide a fair comparison, Figure 62 repeats this analysis for the May 2017 test period and examines CRPS performance for maximum diurnal surface temperature predictions with “Recent” and “Recent-Extended” training data. NPS5 and NPS6 provide the best CRPS performance over this test period, while NPSR provides similar—albeit lower—performance. It should be noted that all three references models provided the poorest CRPS results with these test conditions—even OLSR. Models with MVT likelihoods provided a slight performance advantage.

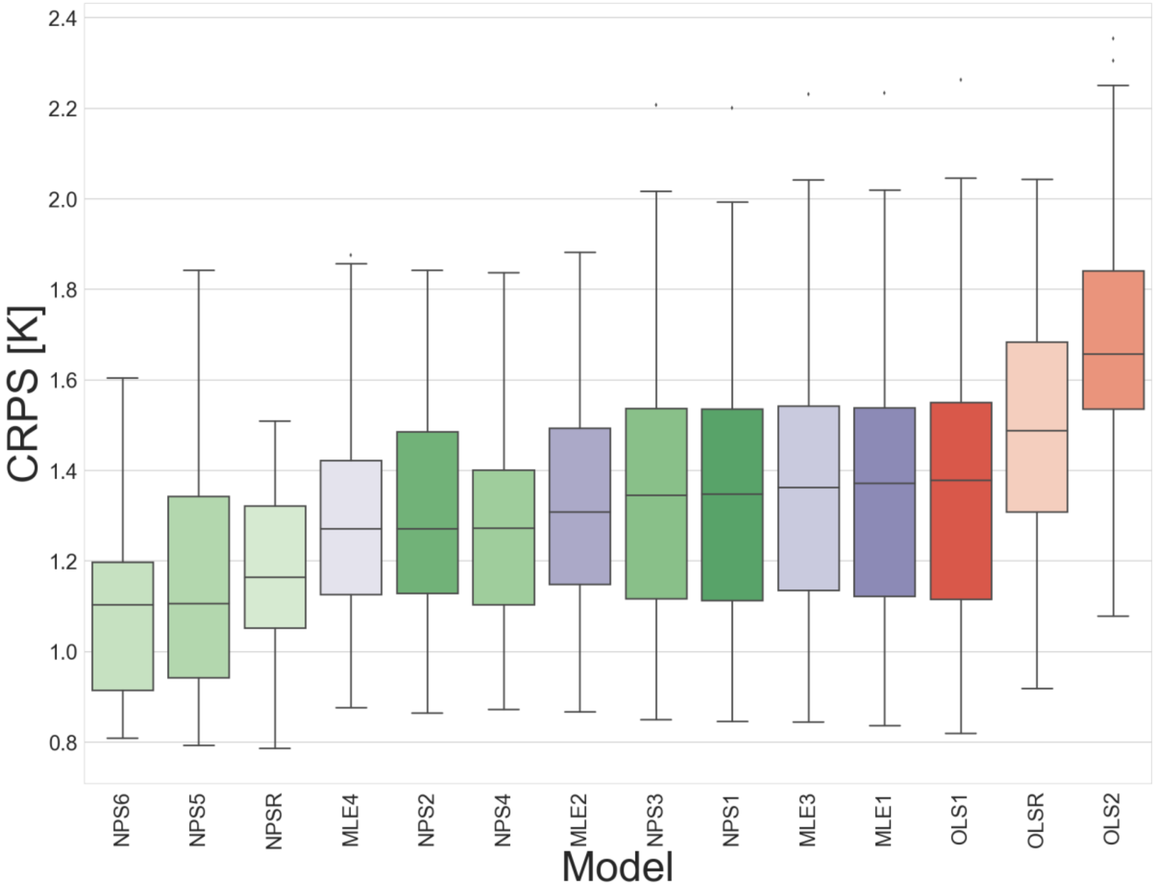


Figure 62. Box plots of inter-city CRPS distributions for maximum diurnal surface temperature predictions with "Recent" and "Recent-Extended" training data. Data correspond to 14 statistical model perturbations over 21 cities during the May 2017 test period. Models are sorted in ascending order according to median inter-city.

A complete summary of CRPS performance is provided by the heatmaps in Figures 63 and 64. As before, the heatmap columns are formed by averaging the absolute error of each statistical model over all training periods, cities, and days for the May test period. Figure 63 describes the generalized CRPS performance of each model in the original units of the sensible weather variables—that is, [K] for diurnal surface temperature extrema and $[m\ s^{-1}]$ for diurnal maximum surface wind speeds. Figure 64 describes standardized CRPS performance with centered and scaled scores according to z-score transformations within each predictand group. This permits comparisons within and between the predictand groups and provides additional contrast in performance.

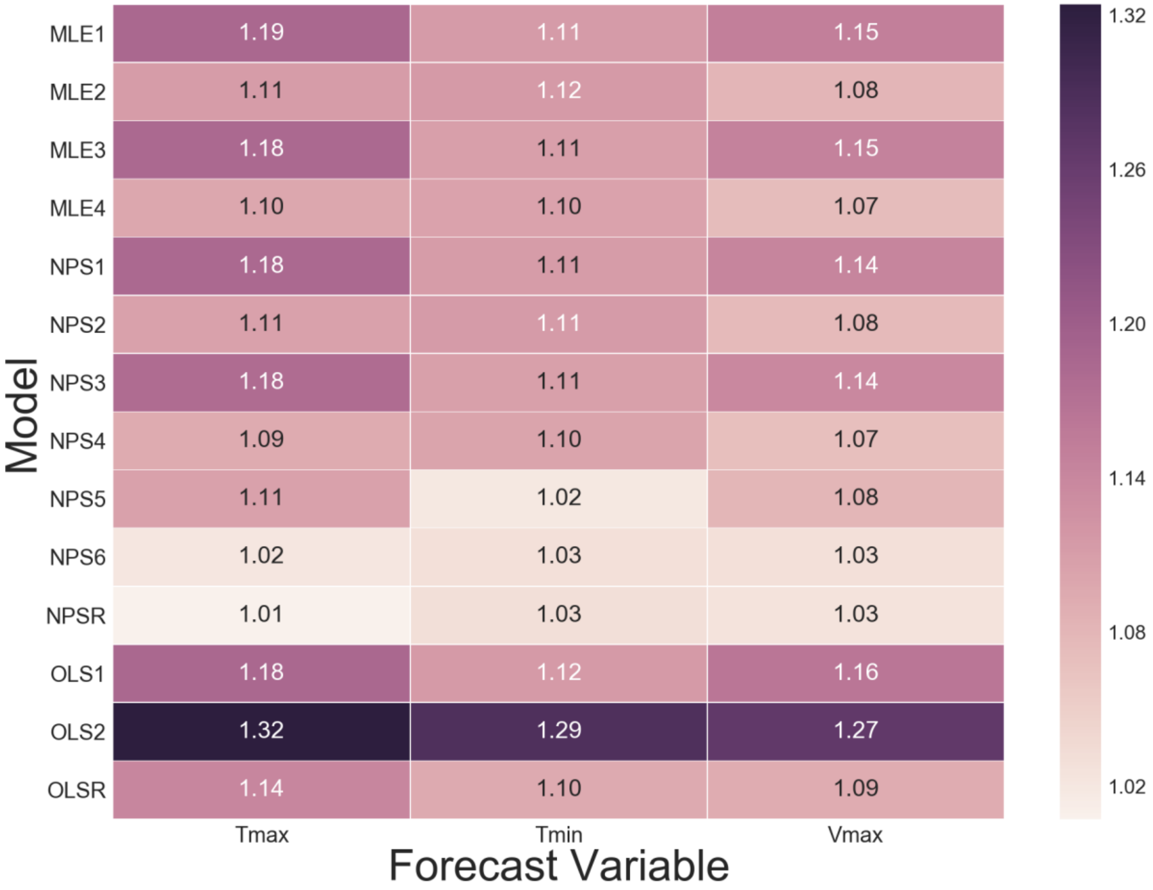


Figure 63. Heatmap of CRPS performance for the statistical models in Tables 2 averaged over all training periods, cities, and days for the May test period.

NPSR, NPS6, and, to a lesser extent, NPS5 provided the best distributions-oriented performance according to the CRPS comparisons in Figure 63. The color palette similar indicates that perturbations trained with control predictors from the NMMB core generally outperformed their peers trained with the ARW control predictors. Figure 64 indicates that models trained from the former provided above-average performance. Moreover, the reference OLS distributions performed poorly overall—especially OLS2.

This result is notable because it represents the reference regression method trained with NMMB predictors—an inversion of the performance pattern observed with the Bayesian schemes. As with the MAE comparisons, no notable differences were observed between model instances with noninformative and hierarchical priors (e.g., MLE1 vs NPS1). However, a slight performance advantage was observed for MVT likelihood functions.

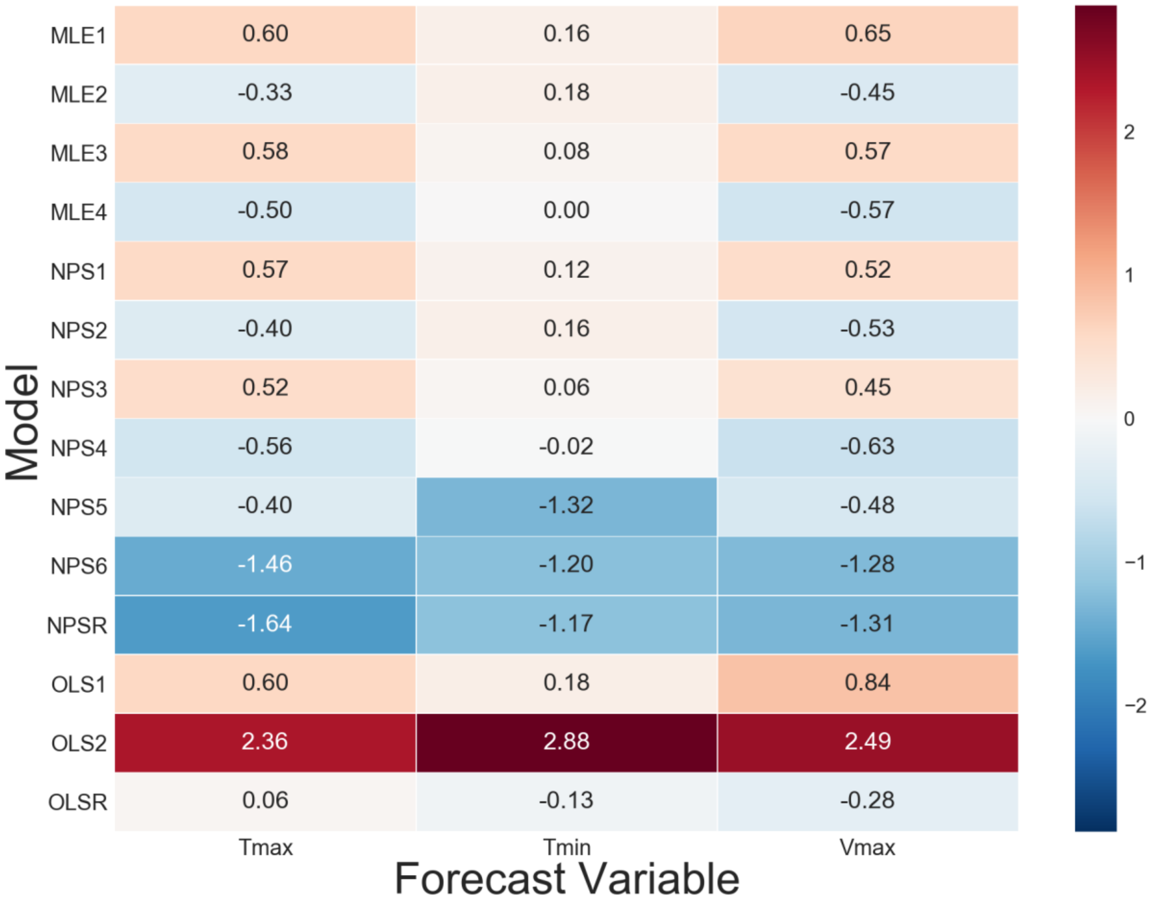


Figure 64. Heatmap of standardized CRPS performance for the statistical models in Tables 2 averaged over all training periods, cities, and days for the May test period. Values represent dimensionless z-score comparisons of CRPS.

Mirroring the original analysis of Gneiting et al. (2005), the present work found similar patterns in ignorance (logarithmic) scores with the aforementioned test conditions. As a strictly proper scoring rule that examines the *a posteriori* probability density of observed truth with the relevant predictive distributions vis-à-vis Equation 58, it measures performance without regard for the desperate units of the predictands. In this way, model skill across predictands can be validly compared according to the idealized point mass of an observation in the associated probability distribution. A convenient reference is provided by the dashed lines in Figure 58 and 59. Ignorance scores measure the agreement between PPDs and observations according to the probability density each forecast distribution assigned to the true value; lower ignorance scores are better.

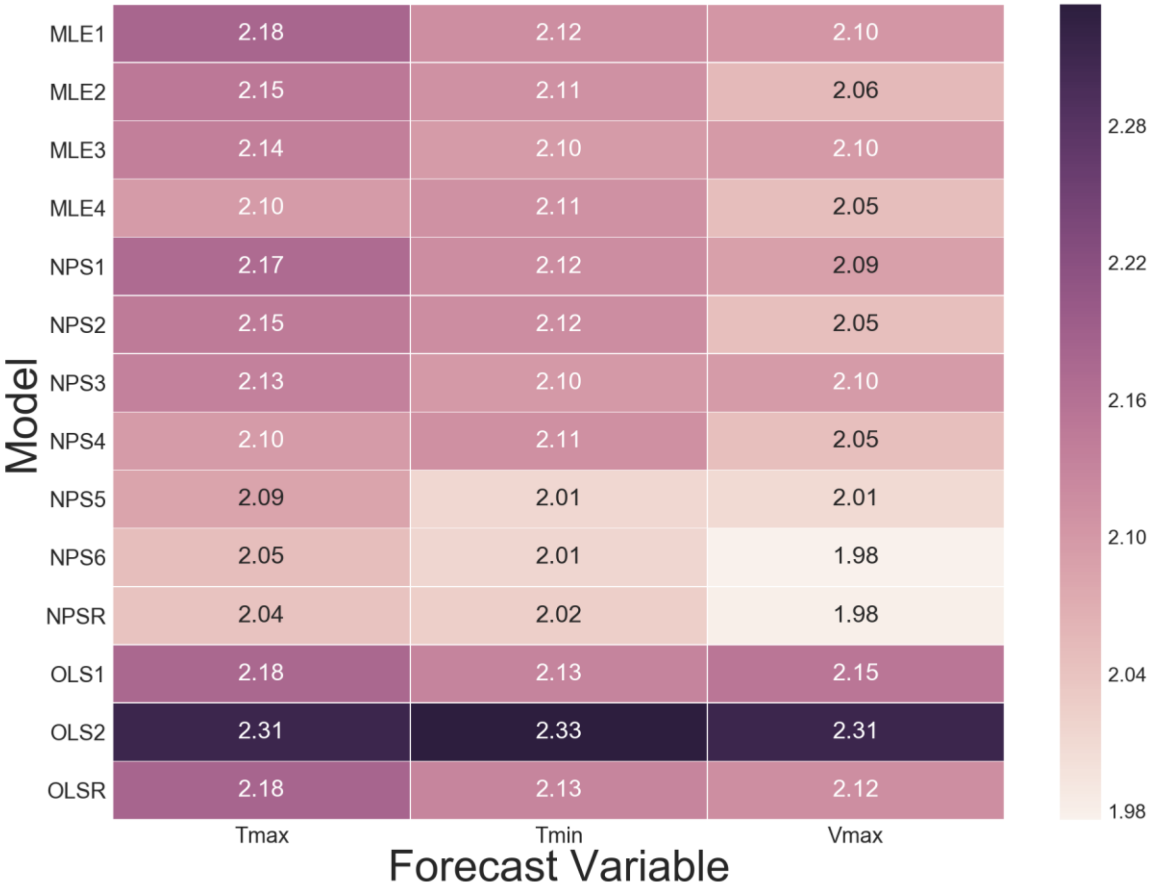


Figure 65. Heatmap of ignorance scores for the statistical models in Tables 2 averaged over all training periods, cities, and days for the May test period. Lower ignorance scores (lighter shading) indicate better performance.

To this end, and for the sake of brevity, the present work will examine ignorance score comparisons beginning with the full heatmaps previously examined. Figure 65 provides such a comparison averaged over all cities, training periods, and days for the full 61-test period covering April and May 2017. JPMs with noninformative and hierarchical prior information show no meaningful distinctions (e.g., MLE2 vs. NPS2). Moreover, the reference OLS distributions produced some of the highest ignorance scores in group—especially OLS2—and were always observed above group means. These ignorance scores also reveal the first valid comparisons—that is, without additional transformations—between predictive performance across the predictands. In the present work, they suggest that the best skill (i.e., the lightest shading) was obtained for minimum diurnal surface temperature and maximum diurnal wind speed predictions. However, differences between predictands were relatively small.

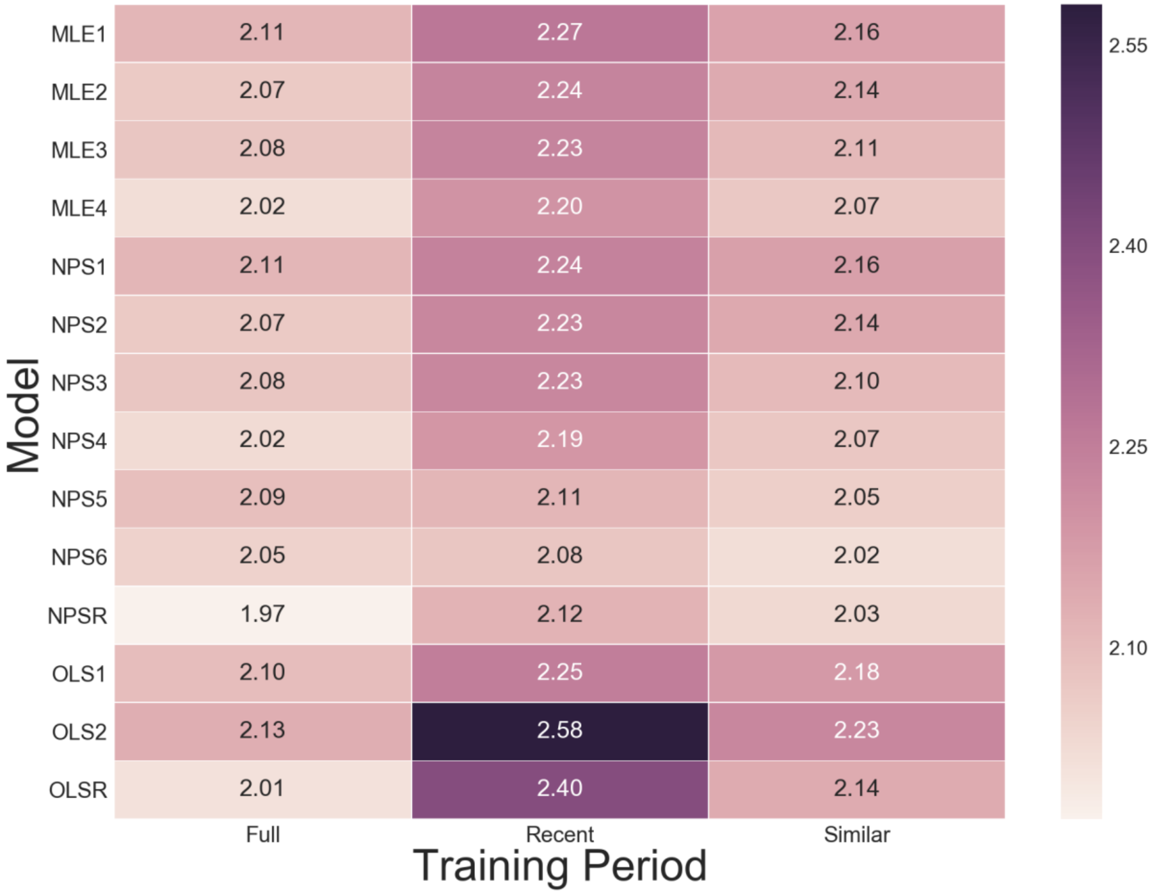


Figure 66. Heatmap of ignorance scores for diurnal maximum surface temperature predictions with the statistical models in Tables 2. Results are averaged over all cities and days for the full 61-day test period. Lower ignorance scores (lighter shading) indicate better performance.

Nevertheless, an interesting result emerged in Figure 65 for the likelihood function comparisons. MVT likelihood functions appeared to affect more noticeable performance benefits vis-à-vis robust regression. In particular, we note that NPS3 provided a modest yet non-trivial advantage over NPS1; this finding was repeated with the other dynamical solver (i.e., NMMB) for NPS2 and NPS4. However, this effect was only observed with diurnal maximum surface temperature predictions. Regardless, the parallel performance gains between both SREF cores are compelling. To examine this finding in more detail, we can condition the information in Figure 65 by training period in the same heatmap format. Figure 66 depicts the ignorance score results for maximum diurnal surface temperature predictions as a function of training period when averaged over all cities and days in the full 61-day test period. Similar to the findings with MAE and CRPS comparisons, ignorance scores revealed the best performance with “Full” and “Similar” training data. However, Figure 66 indicates that MVT likelihood functions expressed performance gains during all three training periods (e.g., MLE1 vs MLE3). As previously indicated, this finding was not duplicated with the remaining two predictands.

Heatmaps conditioned by other variables can be similarly explored. Figure 67 considers standardized ignorance scores associated with maximum diurnal surface temperature predictions for each forecast city. It also adds ignorance score information for the raw dynamical distributions for a full 17-model comparison. The poor performance of the raw ensemble distributions is evident with the horizontal strip of warm shading in the first three rows; this indicates ignorance scores that are well above the average the model group within each city. AREF was notable for producing high ignorance scores with every city; however, the MBEF ensemble—which is based on the 13-memebr distribution from the NMMB dynamical solver—did perform well during the 61-day test period in certain cities. Dallas/Fort Forth (i.e., DFW) was notable for providing an inversion of distributions-oriented forecast performance for the OLSR model. Moreover, MVT distributions in MLE3 and NPS3 mitigated the poor showing by other ARW models (i.e., AREF, MLE1, NPS1, and OLS1) for this city. Nevertheless, both OLS and NPS BEMOS perturbations consistently outperformed the raw ensembles from which their predictors were drawn. The latter was notable for outperforming their dynamical counterparts in almost every city for maximum diurnal surface temperature and wind speed predictions. Figure 68 provides the same basic ignorance score information—now without the z-score standardization and conditioned by training day. Moreover, it describes raw ignorance score results for minimum diurnal surface temperature predictions. In this way, each column represents ignorance scores averaged over all 21 cities and test periods for each forecast trial in the full 61-day test period.

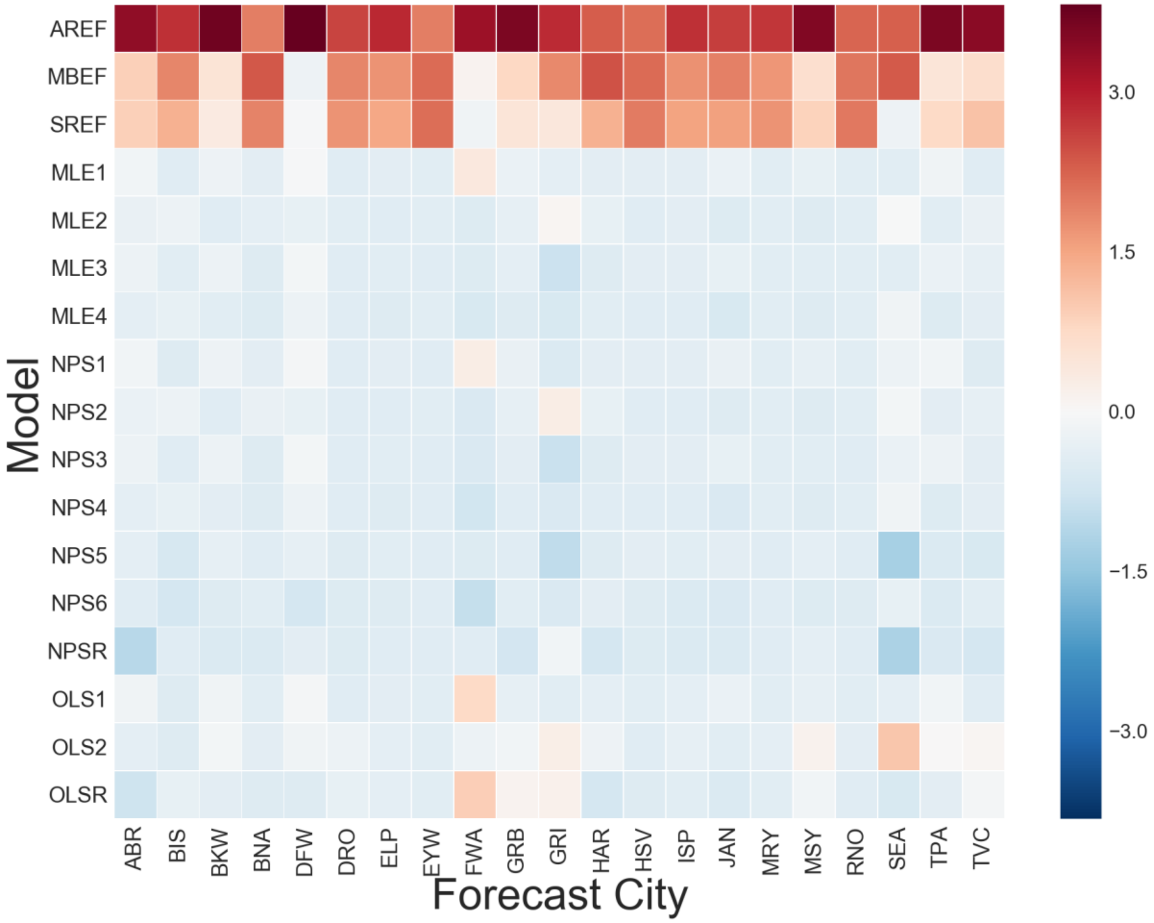


Figure 67. Heatmap of standardized ignorance scores for all 17 primary models. Scores are averaged over all training periods for diurnal maximum surface temperature predictions during the full 61-day test period. Shaded values represent dimensionless z-score comparisons within each city (column). Lower (cooler) values indicate better performance.

Figure 68 similarly indicates that the raw dynamical distributions consistently provided the poorest distributions-oriented performance. This is evident with the darker horizontal strip of ignorance scores in the first three rows. An inversion of the performance comparison between AREF and MBEF is apparent as well; the former provided the worst performance when ignorance results were conditioned by city. However, this result is localized to the predictand; the reverse was found with maximum diurnal surface temperatures—as before. Finally, the conditioning in Figure 68 indicates that the statistical forecasts did not experience any "catastrophic" forecast days; this is, forecast trials where the predictive distribution was notably inconsistent with truth.

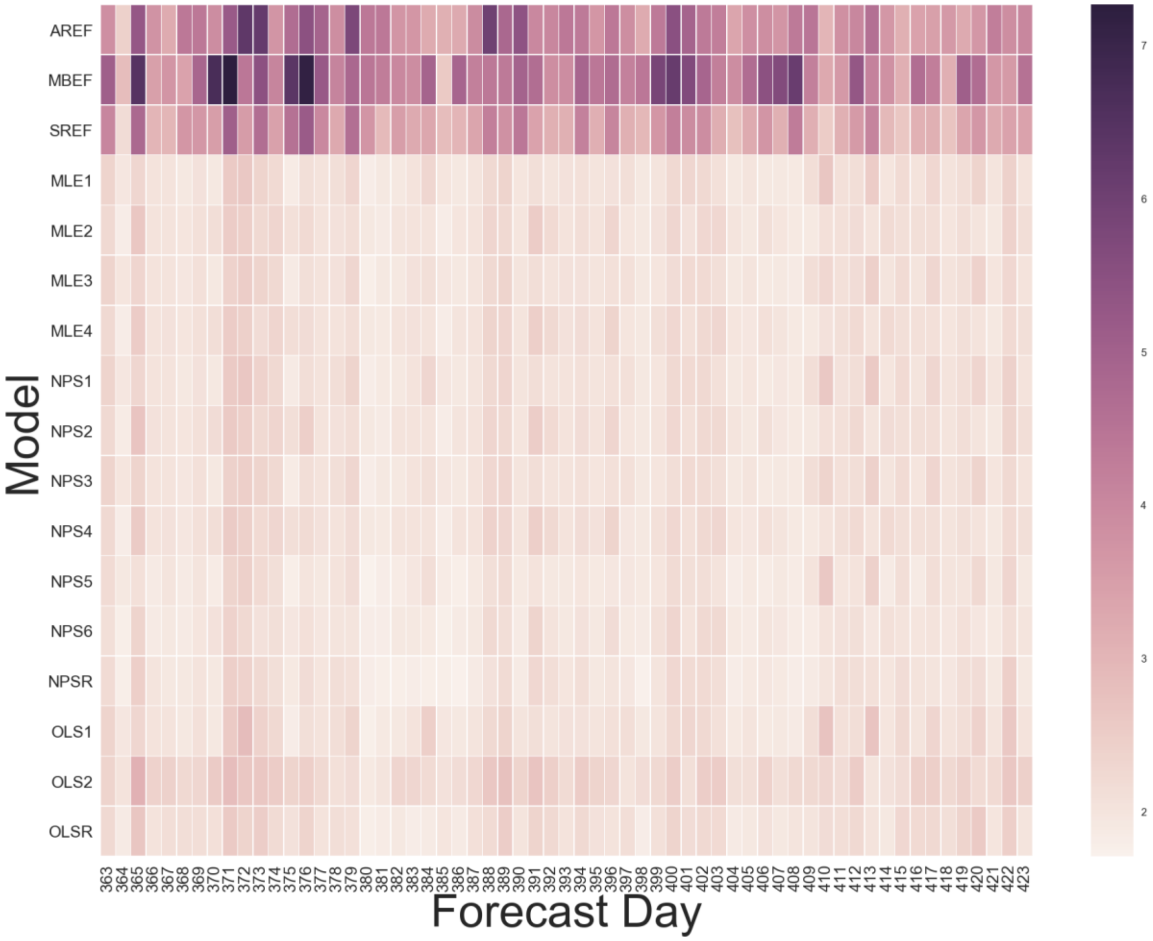


Figure 68. Heatmap of ignorance scores for diurnal minimum surface temperature predictions. Results are averaged over all cities and training periods for the full 61-day test period. Lighter shading indicates better performance.

To summarize these dimensionless ignorance scores with additional clarity, an aggregated performance measure that combines information from all predictands has been formed according to a sum of squares so that

$$IGN_C = \sqrt{IGN_1^2 + IGN_2^2 + IGN_3^2}, \tag{70}$$

where IGN_i describes the ignorance score obtained from the i^{th} predictive dimension (e.g., IGN_1 corresponding to ignorance scores for maximum diurnal surface temperatures) and IGN_C describes the combined performance along all three predictands. Using this format, dotplots for the combined performance of the statistical models in Table 2 are reported in Figure 69. These results combine ignorance scores from all predictive dimensions averaged over all cities, training periods, and test days. Models have been sorted in ascending order according to the combined ignorance scores so that relative performance is indicated by model order; better performance is indicated on the left of each panel. Figure 69a considers the full 61-day test period (i.e., April and May 2017) from Table 3; Figure 69b fairly considers the influence of the April training update with NPS5 and NPS6 by showing combined ignorance scores during May 20017 only. Raw dynamical distributions have been omitted from Figure 69 due to poor performance and scale considerations. In this way, Figure 69 arguably represents the most complete single measure of performance appropriate for the full forecast distributions produced by the statistical models in this dissertation.

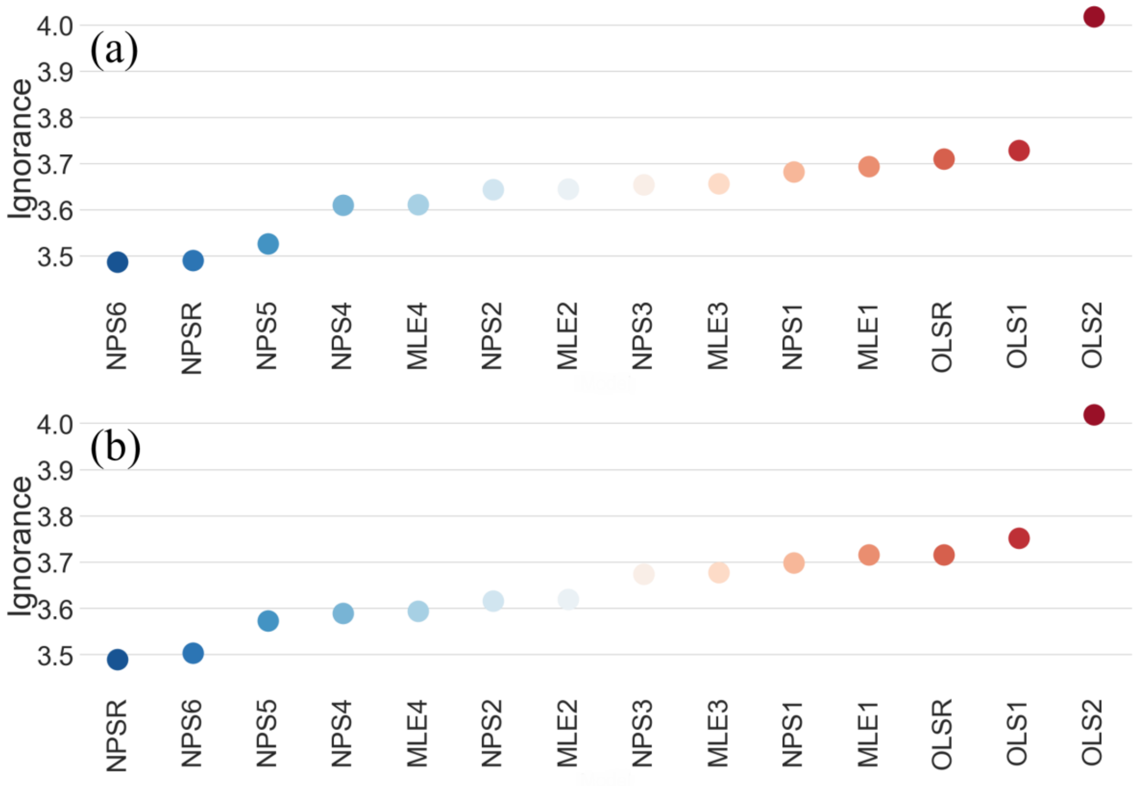


Figure 69. Summary of statistical model performance with combined ignorance scores averaged over all cities, training periods, and forecast days. The top (bottom) panel considers the full (May) test period.

Figure 69 suggests that hierarchical prior information had a very small effect on measures-oriented performance when compared with noninformative JPMs. While both NPS1, NPS2, NPS3, and NPS4 outperformed their noninformative counterparts (i.e., MLE1, MLE2, MLE3, and MLE4)—a consistent trend that is notable when averaged over all design variables—the margins were nevertheless small. Comparisons between MVT and MVN likelihood functions yielded larger and more measurable performance margins. This result was also repeated over both dynamical solvers and, as a result, is likely non-trivial. However, these margins were smaller than the impact provided by predictor variables and training period. The former is demonstrated by SREF ensemble mean predictors consistently outperforming control predictor from each core. Moreover, NMMB predictor variables produced statistical forecasts that outperformed their ARW counterparts. Figure 69b similarly suggests that updated training information provided a notable improvement in predictive performance relative to peer-models (i.e., NPS5 vs. NPS1 and NPS6 vs. NPS2). This indicates that predictor source and the length/character of the training data had a larger impact on measures-oriented performance than JPM perturbations. Finally, it should be noted that all NPS BEMOS instances added value over the reference OLS methods—even the OLSR instance trained with ensemble mean predictors.

4. Calibration

Ensemble calibration can be partially assessed with the distributions-oriented scoring rules considered by the previous section. In this way, lower CRPS and ignorance scores suggest that a model is producing forecast distributions that are relatively more calibrated. Nevertheless, the reliability and resolution of the probabilistic forecasts is frequently evaluated with other visual methods of inspection. Gneiting et al. (2005) and Richter (2012) notably use the PIT histograms described in Chapter III to assess the sharpness and calibration of their forecast distributions. In this way, the present work has aggregated PIT values from every training period considered by the statistical forecast distributions in Table 2. Much like the Taylor diagrams, the distribution of PIT values communicates a great deal about the characteristics of the forecast distributions and how well they correspond to the observations. Uniform PIT histograms indicate a calibrated distribution; a non-uniform distribution suggests the ensemble is biased or contains anomalous dispersion.

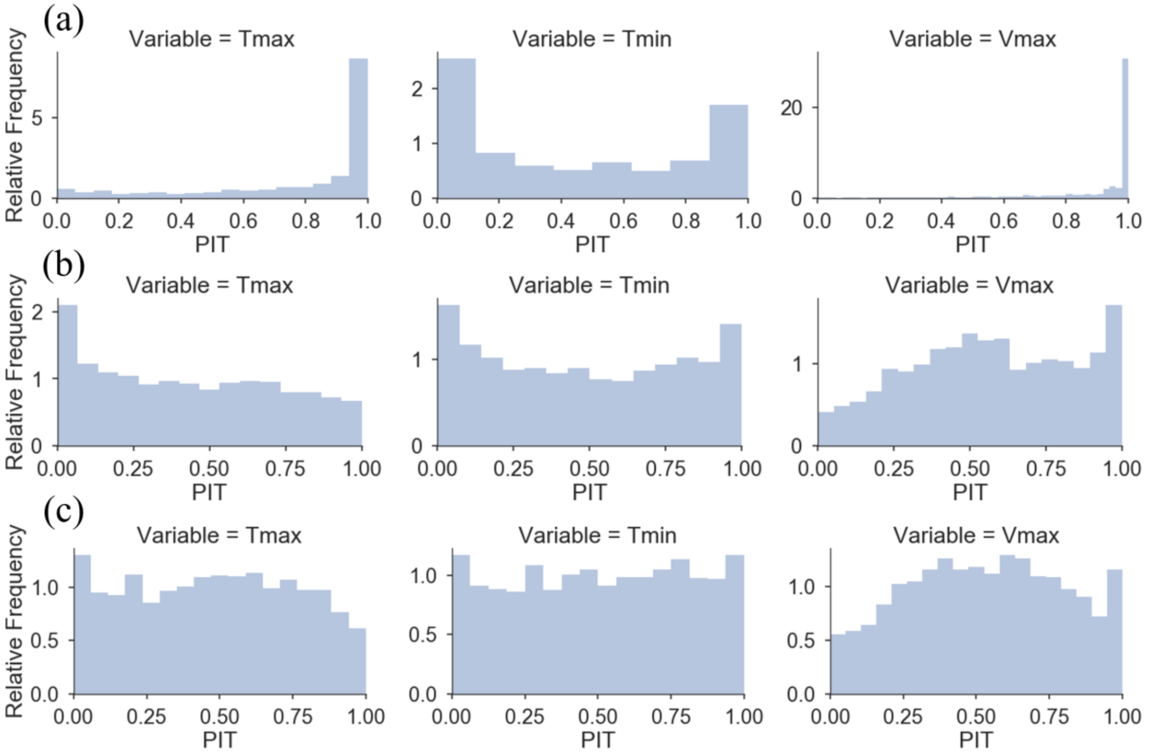


Figure 70. PIT histograms for SREF (a), OLSR (b), and NPSR (c) forecast distributions over all cities, training periods, and test days.

To this end, Figure 70 shows a PIT histogram comparison for all models associated with the full SREF ensemble. The top panel (a) corresponds to the PIT values for each predictand extracted from the raw EPS for every city and test day; the middle panel (b) describes the PIT values obtained from the OLS reference method trained with ensemble mean predictor variables for every city, training period, and test day; the bottom panel depicts PIT values collected from the NPSR model for every city, training period, and test day. In this way, each panel in Figure 70 examines forecast distributions that are based on the same dynamical information. Raw EPS output (top panel) produced notably uncalibrated forecast distributions. Predictive distributions for maximum diurnal surface temperature and wind speed show significant bias; the left skew in both subplots indicates a cold bias for the former and a calm bias for the latter. Minimum diurnal surface temperature predictions appear to be relatively unbiased from the negligible skew in the center column of the top panel (a); however, the "u-shaped" curvature of this PIT distribution indicates that the raw ensemble is underdispersive—that is, it is overconfident with minimum diurnal surface temperatures.

OLSR predictive distribution calibration is indicated in Figure 70b. These results are notably more calibrated—especially for maximum diurnal surface temperature and wind speed predictive distributions. The first column of the middle panel (b) is relatively uniform in appearance—save for the large number of observations found below the predictive distributions. This suggests that OLSR has a slight warm bias but is otherwise calibrated. The middle column is also unbiased; however, it is also underdispersive and provided no meaningful improvement to calibration relative to the original dynamical EPS. The third and final column in the middle panel (b) is marginally overdispersive and contains a slight bias toward calmer winds relative to the observations. This is certainly an improvement over the raw EPS output. In this way, OLSR forecast distributions were adequately calibrated relative to the original SREF ensemble.

The Bayesian estimation scheme associated with NPSR is considered in Figure 70c. Forecast distributions for the diurnal surface temperature extrema have relatively uniform PIT distributions—especially for diurnal minimum surface temperature. The latter suggests that the NPS BEMOS model has properly calibrated the original EPS output for this predictand. Based on the change in scale for the first column of Figure 70c, maximum diurnal surface temperatures are also well calibrated. A slight warm bias is indicated, but it is smaller in magnitude than the raw EPS and OLS solutions. Finally, the NPSR perturbation is overdispersive relative to wind speed observations. This suggests that the associated forecast distributions have too much spread—that is, they are underconfident. There is a slight bias toward calmer surface winds too. Nevertheless, the magnitudes of these defects are smaller than with other methods in Figure 70. In this way, the NPSR model produced forecast distributions with the best calibration across all predictands relative to peer models using similar dynamical information.

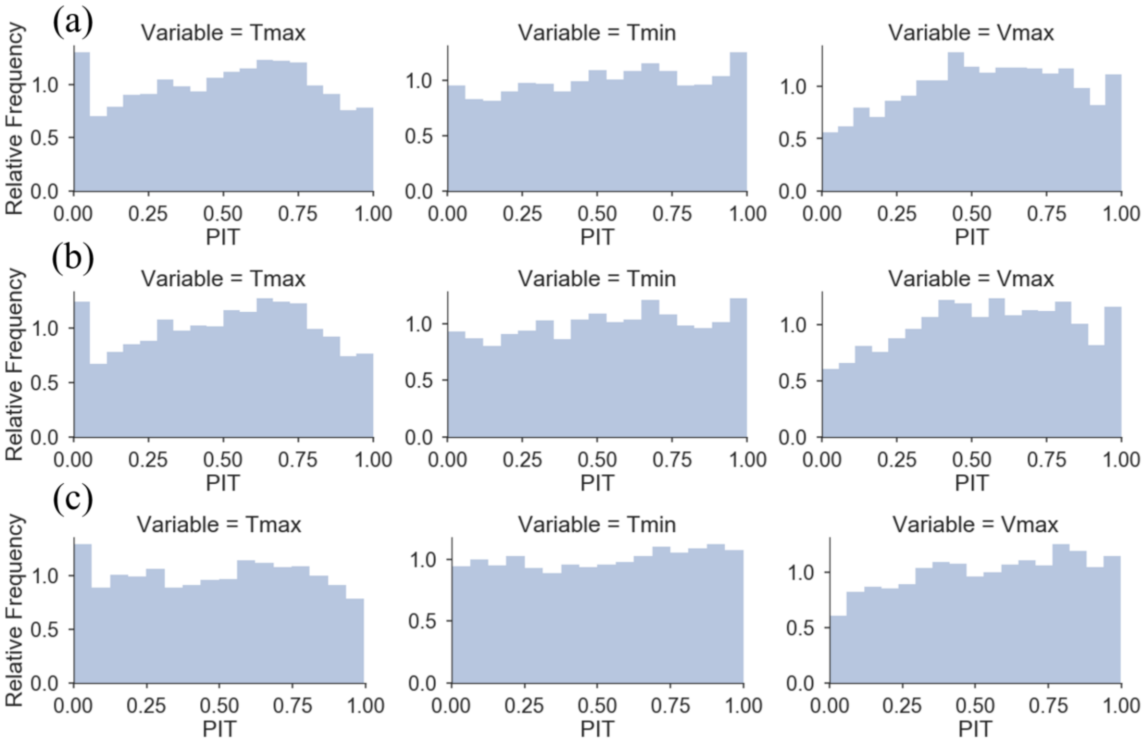


Figure 71. PIT histograms for MLE1 (a), NPS1 (b), and NPS3 (c) forecast distributions. The PIT values represented in each distribution are aggregated from all cities, training periods, and test days.

Additional calibration comparisons are available in Figure 71. Each panel depicts PIT values from models trained with control predictors from the ARW dynamical solver over every city, training period, and test day. The top panel (a) corresponds to MLE1; the middle panel (b) describes NPS1; the bottom panel (c) depicts NPS3. In this way, Figure 71 explores two of the research questions posed for this dissertation: it considers the impact of JPM structure on calibration vis-à-vis likelihood functions and Bayesian priors. Figures 71a and 71b have identical likelihoods but different priors. Nevertheless, they are virtually indistinguishable. This suggests that the hierarchical Bayesian prior information in NPS1 had little impact on calibration. However, Figure 71c provides modest but noticeable improvements to calibration for all three predictands; the center column is particularly uniform. In this way, all three predictive dimensions displayed more uniform PIT histograms for NPS3. This result was replicated with NPS4, which suggests MVT likelihood functions were better calibrated than peer models with MVN likelihoods.

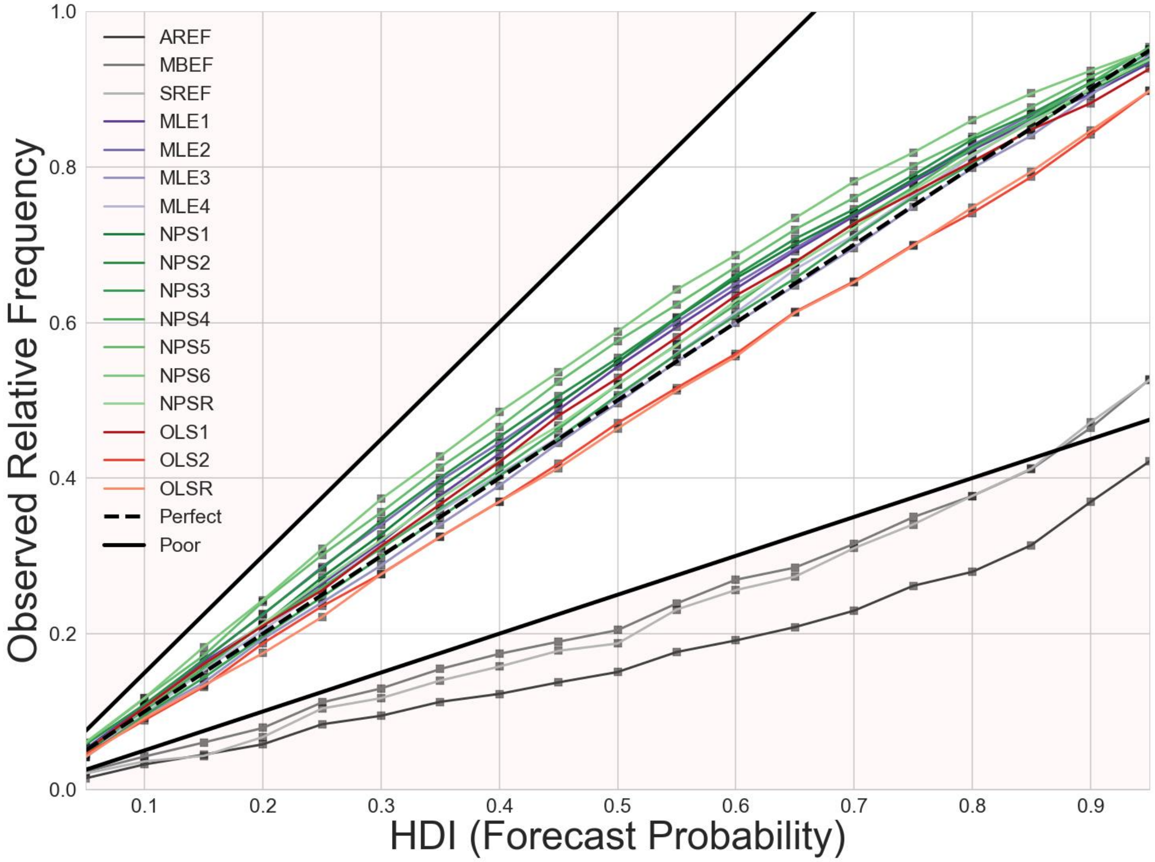


Figure 72. Reliability diagram comparing the relative frequency of diurnal maximum surface temperature observations in various predictive HDIs. Raw dynamical PPDs are indicated in grey; OLS PPDs in red; NPS PPDs in green; and MLE PPDs in purple. Perfect reliability is indicated with a dashed black line; solid black lines border a red shaded region of “poor” reliability.

Figure 72 considers another calibration metric—that is, reliability—for the models compared heretofore. It depicts the relative reliability performance of diurnal maximum surface temperature PPDs produced by raw dynamical (grey), OLS (red), NPS (green), and MLE (purple) methods. A dashed black line of perfect reliability indicates ideal performance. These results have been averaged over all cities, training periods, and days in the full 61-day test period. Figure 72 indicates that the statistical models are notably more reliable than the raw ensemble distributions; the latter are typically observed in a region of “poor” reliability. However, there are no meaningful distinctions between statistical methods were readily apparent; NPS BEMOS models are generally observed above the line of perfect reliability, while OLS models (blue) are generally below this line. In this way, the former would be properly described as underconfident (i.e., overdispersive HDIs) while the latter is overconfident (underdispersive HDIs). Nevertheless, the magnitudes of theses reliability anomalies relative to the “prefect” line are notably smaller than the reliability anomalies associated with the dynamical models.

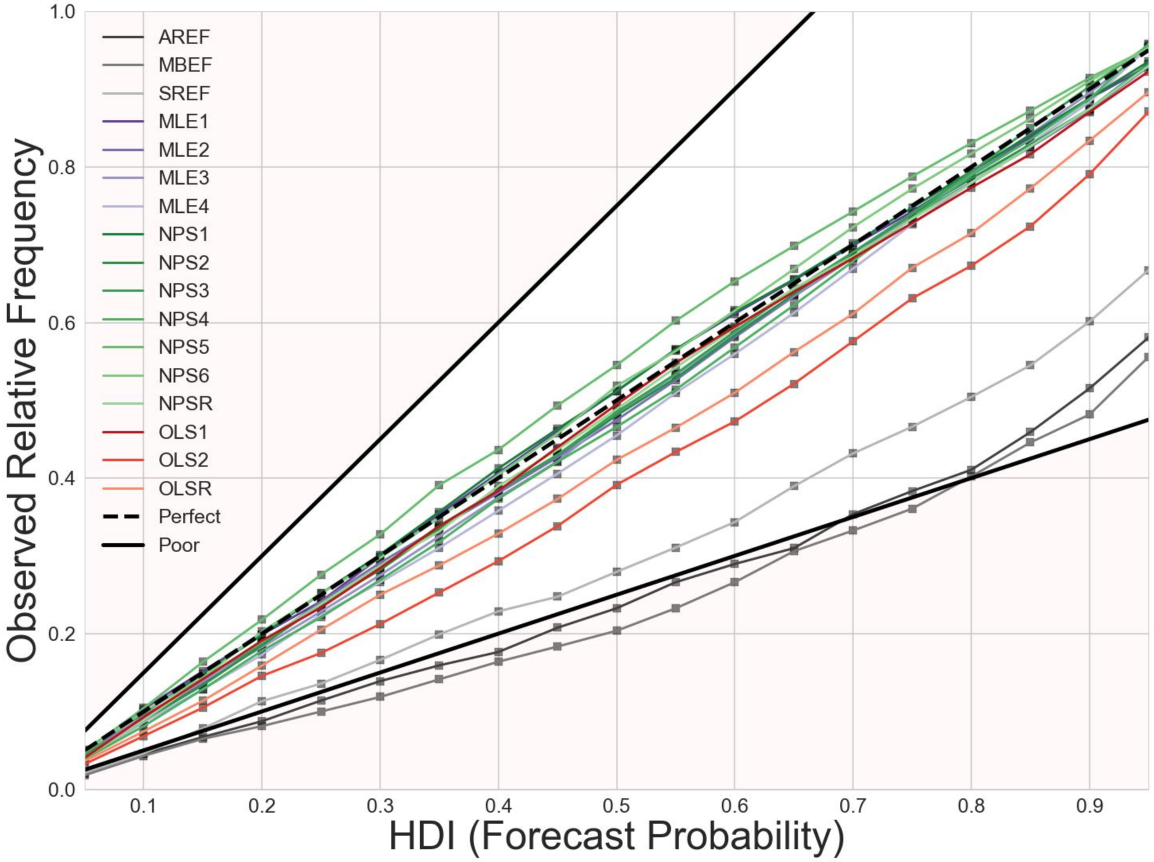


Figure 73. Reliability diagram comparing the relative frequency of diurnal minimum surface temperature observations in various predictive HDIs. Raw dynamical PPDs are indicated in grey; OLS PPDs in red; NPS PPDs in green; and MLE PPDs in purple. Perfect reliability is indicated with a dashed black line; solid black lines border a red shaded region of “poor” reliability.

Results for diurnal maximum surface wind speed PPDs (not shown) were consistent with Figure 72. The raw ensemble PPDs produced the poorest reliability performance while the statistical models were tightly clustered near the line of perfect reliability. NPS BEMOS models again show a slight tendency toward underconfident HDIs relative to OLS methods. However, Figure 73 shows an advantage for NPS BEMOS models (green and purple) with HDIs for diurnal minimum surface temperature. In particular, OLS models (red) were underdispersive compared with their Bayesian counterparts. Raw ensembles (grey) produced their best relative reliability performance with this predictand—especially the full SREF ensemble.

A combined summary of reliability performance is provided in Figure 74. It considers the absolute deviation of each individual reliability curve from perfect reliability (i.e., the dashed black reference line) to form an absolute anomaly for each model. These values were then averaged over all HDI mass fractions and predictands for each of the 17 primary models considered by this dissertation. Models have been sorted in ascending order according to their mean reliability anomaly rank; better performance is indicated on the left. Figure 74a depicts the combined anomaly performance for the full 61-day test period; Figure 74b describes the same but for the May 2017 test period only. Both panels indicate a poor reliability performance for the raw dynamical distributions. This was especially true for the MBEF core, which produced mean absolute reliability anomalies above 0.4 for both test periods. Calibration via OLS methods fared better; however, only one of the OLS models was able to generate reliability performance above the group mean. In this way, NPS BEMOS models consistently outperformed their dynamical and OLS counterparts—albeit with small margins for the latter. Bayesian models with hierarchical priors outperformed peer models with noninformative priors. Moreover, Bayesian models with MVT likelihood functions lead the group and produced small yet consistent margins over similar models with MVN likelihood functions. This finding was duplicated with combined ignorance scores and with PIT histograms and is, therefore, likely meaningful. Nevertheless, the performance margins between statistical methods was comparatively small. As before, the primary performance distinctions are expressed between statistical and dynamical methods—not within the former.

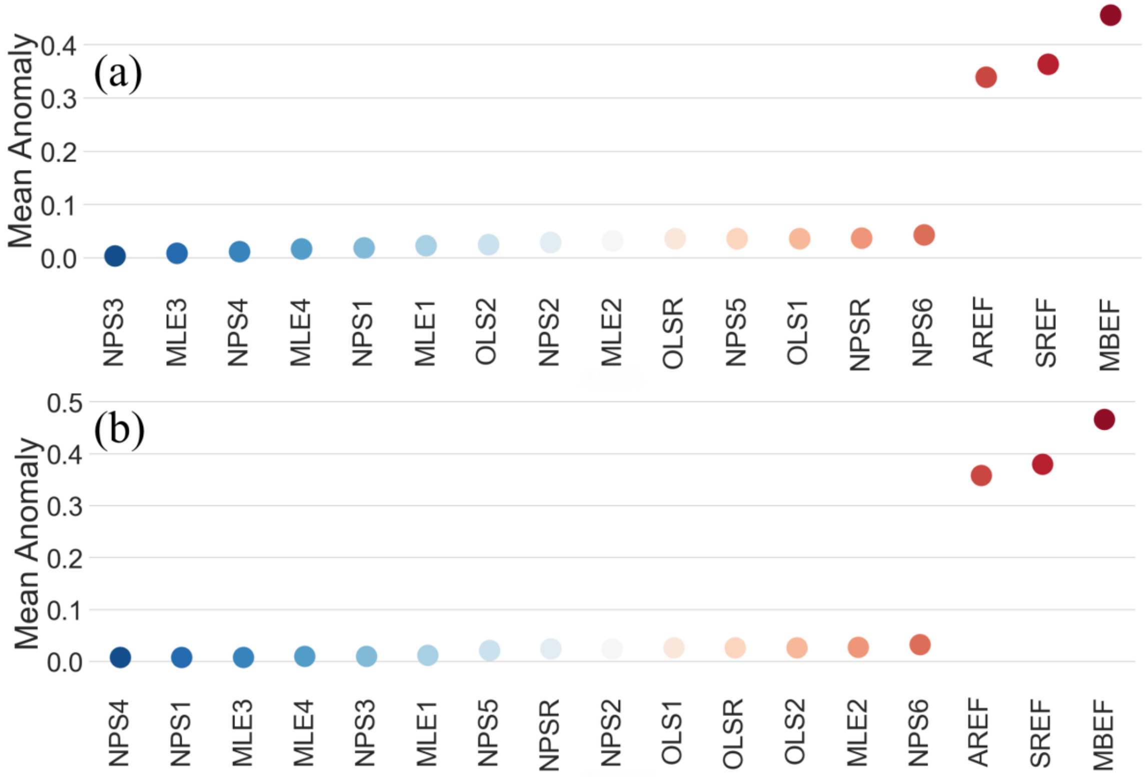


Figure 74. Mean absolute reliability anomalies averaged over all HDI mass fractions, predictands, cities, training periods, and test days. The top (bottom) panel considers the full (May) test period.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY

A. DISCUSSION

Previous research in statistical post-processing methods has found systematic deficiencies in deterministic forecast guidance. Gneiting et al. (2005), Raftery et al. (2005), and Hodyss et al. (2016) all note that ensemble predictions are frequently observed with biased central tendencies and anomalous dispersion. In this way, regression techniques that statistically characterize the relationship between objective guidance and its predictive errors have been developed to post-process raw NWP output and affect better forecast performance. To this end, Gneiting et al. (2005) introduced the NR/EMOS technique as a parametric extension of MOS (Glahn and Lowry 1972) to generate probabilistic forecasts for continuous weather variables in the form of predictive distributions (Gneiting et al. 2005). Richter (2012) then adapted the NR/EMOS approach for Bayesian inference to adopt a more rigorous framework of parameter estimation. This dissertation independently developed a Bayesian approach to EMOS that is consistent with the BEMOS technique originally introduced by Richter (2012). It was conceived in the search for a programmatic scheme that transforms deterministic predictors into reliable forecast distributions that outperform raw objective guidance and other common forms of statistical post-processing.

This dissertation developed a multivariate multiple linear regression model that uses Bayesian parameter estimation and MCMC sampling techniques to heuristically “learn” optimal posterior inferences for model parameters conditioned on the structure of the joint probability model and available training data. Posterior beliefs in model parameters contain probability statements in the form of PDFs and HDIs that explicitly and reliably communicate the joint uncertainty of the regression coefficients. Bayesian posteriors were formed by optimizing a cost function associated with a parametric predictive distribution (i.e., likelihood function) and hierarchal priors shaped by model parameter structure and training data. Posterior predictive distributions are then formed by drawing representative samples from the parametric predictive distribution weighted by Bayesian posterior beliefs in our model parameters. These Bayesian PPDs were then compared with raw EPS guidance from the SREF system and a peer multivariate multiple regression model completed with traditional OLS methods. Predictive distributions for the dynamical models was formed by dressing the daily central predictive tendency of SREF dynamical solvers with the variance of their members; OLS distributions were created by fitting frequentist variance estimates to the associated single-valued parameter estimates.

In this way, the present work represents an extension of the original BEMOS concept. It permits full multivariate predictions with coupled covariance between vector predictands. Moreover, it introduces a hierarchical parameter structure that treats some hyperparameters as random variables that must be inferred from training data. It also uses a customized adaptive Metropolis sampler with block-wise multiparameter updates to complete non-conjugate Bayesian inferences. It engaged robust regression through the novel use of a multivariate t location-scale distributions as a parametric predictive distribution. Finally, it uses logarithmic data transformations to produce asymmetric predictive distributions for sensible weather variables with intrinsic skew. Unlike the NR/EMOS/BEMOS techniques of Gneiting et al. (2005) and Richter (2012), the present work primarily used information available outside of a full EPS. In this way, the NPS BEMOS model can be directly compared with raw dynamical EPS guidance to assess the relative forecast performance of both methods vis-à-vis their generalized computational costs.

This dissertation introduced four primary research questions associated with the impact of various model components on forecast performance. More specifically, it investigated the influence of training period length and character, the type of ensemble predictor(s) used to train the model, the distributional form of the parametric likelihood function selected, and the presence of Bayesian prior information. Based on the theory and methods in Chapters II and III, it was assumed that more training data affects better performance; additionally, training data with more meteorological similarity/relevance to the test period was believed to be more ideal. The present work also assumed that statistical models using control predictors (i.e., estimates available outside of an EPS) can meet or exceed the performance of raw ensembles. Moreover, these control predictors should be able to match or beat the performance of predictors formed from ensemble statistics (e.g., the ensemble mean). A MVT likelihood function was assumed to provide better forecast performance vis-à-vis robust regression. Finally, hierarchical prior information was assumed to be superior to JPMs with noninformative priors. These conjectures were formalized and explored by the statistical model perturbations summarized in Table 2.

Predictive trials for these models were evaluated over three predictands, 21 cities, and 61 test days during 2016 and 2017. The test period comprised April and May of 2017. This represents a modification to the WxC forecast competition format so that the present work considers locations in parallel over a single, comparatively large continuous test period. Three disparate statistical model training periods were formed according to their relationships to the test period. "Similar" training data was collected from a 60-day period with identical seasonal synoptic forcing (i.e., April and May 2016). "Recent" training data was formed by the most recent 60 days to the training period (i.e., February and March 2017). The "Full" test period considered the previous calendar year prior to the test period and contained data from both of the shorter periods. Forecast performance was evaluated by measures-oriented (e.g., MAE) and distributions-oriented (e.g., CRPS) scoring rules to assess the agreement between the predictive distributions and the observations. PPD calibration—that is, sharp forecast distributions subject to observational consistency—was also assessed according to the methods of Gneiting et al. (2005), Gneiting et al. (2007), and Richter (2012).

Results from the analysis and forecast evaluation suggest that the local approach to parameter estimation was appropriate. Posterior parameter beliefs showed meaningful variability across cities and, in some cases, training periods. Dynamical model perturbations (i.e., SREF ensemble members) were found to be highly correlated. Statistical forecasts produced by OLS and Bayesian methods exhibited similar levels of group covariance; however, the statistical models showed notably reduced linear correlation with the central tendencies of each dynamical solver (i.e., SREF core) and, indeed, the entire 26-member ensemble. The comparatively low covariance between statistical and dynamical forecasts suggests a unique predictive character for the former.

In simple terms, the statistical perturbations were not found to be superficial bias corrections that mirrored the variability of the dynamical estimates. Taylor diagrams were then introduced to examine the variability of the models relative to the observations and determine if the low relative covariance was beneficial vis-à-vis measures-oriented forecast performance. These results were averaged over all cities, training periods, and test days and indicate that the statistical models—both Bayesian and OLS—had forecast variability that better matched the observations compared with dynamical forecast variability. However, NPS BEMOS model perturbations were found to consistently outperform OLS models in linear correlation with the observations and centered RMS error—especially the NSPR model given "Full" and "Similar" training data. The Bayesian models also demonstrated notable performance margins in relative forecast variability with diurnal maximum surface temperature predictions. The "Recent" training period accentuated the aforementioned performance comparisons for all predictands. Examining the May 2017 subset of test data revealed similar findings. However, NPS BEMOS models given access to extended training data from April 2017 (i.e., NPS5 and NPS6) showed notable performance gains during this test period. This suggests that training updates can have a beneficial impact on forecast performance.

Measures-oriented forecast performance was then examined through MAE scores. A rank-sorting of the models according to median inter-city MAE indicated that "Full" and "Similar" models generally provided the best results; NPSR_F lead all models. Instances trained with SREF ensemble mean predictors (i.e., all variants of NPSR and OLSR) also did well within their respective training groups. Dynamical model performance was notably poorer in the MAE comparison and fell short of the statistical models in each comparison. However, the performance gradient within statistical models was small. While Bayesian models typically outperformed their OLS counterparts within each training period group, the margins were generally negligible. As with the Taylor diagram comparisons, the "Recent" period produced greater contrasts in forecast performance. In particular, the OLS models fared poorly relative to NPS BEMOS instances for all three predictands in this training period. Moreover, important patterns in relative performance was expressed between models with MVT and MVN likelihood functions and, separately, hierarchical and noninformative priors. MAE results in the "Recent" test period suggest that MVT models may have a small performance advantage; however, no meaningful differences between noninformative and hierarchical JPMs were observed.

The theoretical similarities between MAE and CRPS compelled the first distributions-oriented performance considered by this dissertation. The latter describes the dispersion of ensemble members and extends the performance comparison initially pursued with MAE to full predictive distributions. To this end, the CRPS results mirrored the trends observed with MAE scores. In particular, models trained with SREF ensemble mean predictors performed well. Bayesian models generally outperformed their OLS counterparts—especially during the "Recent" training period. Raw EPS distributions fared poorly and consistently provided the worst CRPS values. Standardized heatmaps for CRPS performance indicate larger relative improvements for MVT models—a small yet consistent finding. However, hierarchical prior information had a negligible impact on performance. Dimensionless ignorance scores were then combined in a sum of squares formulation to considered forecast performance coupled along all three predictands in Figure 69. These results were averaged over all cities, training periods, and forecast variables for both test periods. In this way, Figure 69 provided the most generalized distributions-oriented performance measure considered by this dissertation. Diurnal maximum surface temperature predictions were found to be the most challenging for the models. NPS BEMOS models showed consistent performance advantages; every single Bayesian model outperformed their OLS and dynamical counterparts. The combined ignorance score findings also provided clearer evidence for the impact of the MVT likelihood functions. This result was replicated over both dynamical solvers and JPMs. The benefit of hierarchical priors was evident as well—albeit to a smaller degree. However, the largest performance margins were associated with models trained with SREF ensemble mean predictors (e.g., NPSR) and extended data from April 2017 (i.e., NPS5 and NPS6).

Finally, forecast calibration was visually inspected with PIT histograms and a modified form of reliability diagram appropriate for continuous forecast variables and

Bayesian HDIs. PIT histograms indicated that raw ensemble distributions were poorly calibrated and generally biased. When models based on the same dynamical information were directly compared, Bayesian PPDs were found to have less bias and anomalous dispersion (i.e., better calibrated) than the reference OLS models. Moreover, models with MVT likelihood functions were observed to have more uniform PIT distributions. This result was replicated with both SREF cores and all JPMs and reinforces a similar finding from the combined ignorance scores. However, Bayesian models with hierarchical priors showed no meaningful distinctions in PIT distributions when compared with models using noninformative priors. Results from the reliability diagrams indicated that NPS BEMOS model instances were generally underconfident (overdispersive) while the OLS distributions were slightly underdispersive (overconfident). The dynamical models produced the lowest forecast reliability and were generally found in a region of "poor" performance. Reliability scores were then combined in a manner analogous to ignorance scores in Figure 74. These results suggest that Bayesian models consistently produced the lowest mean absolute reliability anomalies when averaged over all cities, training periods, test days, and HDIs. While the performance margins were small between statistical models, the results were consistent with previous findings and indicate OLS methods lagged behind their Bayesian counterparts. Moreover, NPS BEMOS instances with MVT likelihood functions produced the best reliability performance and consistently outperformed their MVN peers. Hierarchical priors also outperformed their noninformative peers; however, the margins were even smaller. As previously stated, distinctions within statistical methods were notably smaller than the differences between dynamical and statistical models. Regardless, the reliability results indicate that Bayesian models were better calibrated than their OLS and dynamical peers.

B. CONCLUSIONS

A primary finding of this research concerns the efficacy of the ensemble approach to meteorological forecasting. While Gneiting et al. (2005) notes that many forms of statistical post-processing correct biased dynamical guidance, there are fewer schemes that engage full ensemble calibration. In this way, the NR/EMOS technique is special. It "plays dice" with the atmosphere to directly parameterize meteorological phenomena with probability distributions that describe the intrinsic structure of the observable data. These parametric distributions should be constructed so that future observations could be mistaken for IID samples from the forecast distributions. Calibration vis-à-vis Gneiting et al. (2007) describes this relationship perfectly, and it provides the motivation for the Bayesian methods explored by the present work.

However, this research considered an important modification to the NR/EMOS technique: it used predictor variables outside of the parent ensemble. In this way, 10 of the 11 primary Bayesian models used dynamical controls instead of information sampled from the EPS perturbations. All of these NPS BEMOS models consistently outperformed forecast distributions formed from the raw ensemble distributions. A similar—albeit smaller—result was obtained when the SREF EPS was calibrated with classical OLS methods using the same multivariate multiple linear regression framework given to the Bayesian models. Bayesian PPDs provided the most accurate and reliable forecast distributions considered by this dissertation—and they did so at a fraction of the computational cost of the parent EPS. While OLS methods are even more efficient, since they obviate the need for sophisticated MCMC sampling techniques, the difference on a single forecast trial is trivial (i.e., on the order of minutes). Dynamical ensembles, for comparison, require many additional hours over their deterministic control members. The NPS BEMOS model can learn from large sets of multivariate training data and produce a calibrated multivariate forecast distribution in roughly 10 minutes on a personal computer. Perhaps most importantly, the central tendency and spread of these Bayesian PPDs is rigorously conditioned to provide consistent agreement with the observations—a result that was replicated with nearly every forecast metric considered by this dissertation.

Moreover, the results engaged key aspects of the primary research questions. As previously stated, Bayesian models trained with ensemble control predictors were found to significantly outperform the raw ensembles and, to a lesser extent, equivalent OLS regression methods. It should be noted that performance margins between NPS BEMOS and OLS models were generally small and might be regarded as negligible in some forecast applications. In this way, classical OLS methods are entirely appropriate when a "good-enough" solution is desired or, perhaps, when measures-oriented metrics are the primary concern. OLS methods may also be ideal for model prototyping; that is, one can use OLS methods to quickly find skilled predictor variables and then use a full Bayesian framework for operational applications. In this way, Bayesian models consistently offered the best predictive performance—especially with scoring rules appropriate for full predictive distributions. The single best performer was also found in a Bayesian model; however, it was trained on SREF ensemble mean predictors (i.e., NPSR). This suggests that ensemble information can be useful in statistical post-processing and that additional predictors can potentially affect better performance. Nevertheless, the performance observed with control predictors was still quite good and, in some cases (e.g., mean absolute reliability anomalies), was the best in the comparison. This reinforces the primary result: full dynamical ensembles are not required to produce calibrated forecast distributions.

The impact of training data length and character was well described by the consistent performance advantages associated with the "Full" training period. The "Full" training set contained both of the other training periods and was notably longer. It also produced the best models in the majority of the performance comparisons. This result was duplicated with multiple performance metrics for both measures-oriented and distributions-oriented scoring rules. In this way, the research conjecture that more data are better was validated. Nevertheless, the largest performance margins between Bayesian and OLS methods were observed with training data that was comparatively short and meteorologically dissimilar (i.e., the "Recent" train period) from the test period. This suggests that BEMOS techniques likely offer the best utility in data-limited environments where time, resources, and operational constraints may preclude thorough model training. Performance with "Similar" test data suggests shorter data sets are feasible, and sometimes desirable vis-à-vis computational resources, when the training data are meteorologically consistent test data. In this way, generalized performance with "Similar" training data was remarkably consistent with the "Full" period. This aspect of the research conjecture was also validated. Regardless of the data available for training, Bayesian models appear to offer better predictive performance. It's merely a question of how large that margin is likely to be and whether the additional cost of MCMC sampling is appropriate for the forecast application.

The results appeared to contradicted the research conjecture that hierarchical Bayesian priors—at least as formulated by the present work—make a meaningful contribution to the MLE/Bayesian cost function. While combined ignorance scores and mean absolute reliability anomalies indicated a consistent advantage for hierarchical JPMs, the margins were so small that the effect would likely go unnoticed in the majority of forecast applications. PIT histograms for noninformative and hierarchical JPMs were also indistinguishable, so the impact on calibration is likely minimal. Large training sets are known to minimize the influence of prior information in Bayesian inference (Gelman et al. 2013), so it is possible that hierarchical model structure could add value in data-sparse environments. Different forms of hierarchical modeling—especially an approach that formed local and regional parameters according to clustering in the data—might offer measurable performance gains. The findings in the present work merely indicate that beta regression shrinkage was ineffective for this application with the number of model parameters considered. Alternatively, beta regression shrinkage performed with t-distribution priors vis-à-vis Kruschke (2014) might make a bigger impact on forecast performance.

Finally, robust regression appeared to provide a more meaningful contribution to forecast performance. Bayesian model perturbations with MVT likelihood functions collected small margins in MAE and CRPS results; however, the performance advantage become more apparent with raw and combined ignorance scores, combined reliability diagrams, and, notably, PIT histograms. While the performance margins were still small with the former metrics, the PIT results showed a measurable contribution to PIT distribution uniformity. When one considers that these findings were repeated, to varying degrees, in almost every performance evaluation—across both dynamical solvers and, indeed, noninformative and hierarchical priors alike—it seems likely that MVT distributions can meaningfully impact forecast performance. In this way, the unique shape of the t-distribution is believed to make parametric regression models more resistant to outliers. This could have important relevance to future work that seeks predictions for extreme, so-called “black swan” events. Nevertheless, the performance impact of t-distribution likelihoods should be balanced against their computational cost. NPS BEMOS models with MVT likelihoods generally required 50% more training time.

C. RECOMMENDATIONS

The research questions posed by this dissertation were structured to permit the associated investigation within suitable time constraints. With more time, the present work would focus on four primary areas for additional development: upgrades to the adaptive Metropolis sampler using Hamiltonian Monte Carlo methods; more sophisticated hierarchical prior structures that cluster model parameters according to important structure in the data; full three-parameter t-distribution likelihood functions that permit all distributional parameters as random variables; and, finally, a full evaluation of training data updates that properly explores the relationship between training period length and predictive performance. It would also seek a wider application of the hierarchical multivariate BEMOS approach in other forecast applications (e.g., tropical cyclone forecasting). Future investigators should note that free, open-source MCMC samplers are widely available within the data science community. The present work considers the construction of the customized Metropolis sampler as a necessary pedagogical exercise in Bayesian data analysis appropriate for a doctoral dissertation. However, standard investigations are encouraged to use existing resources (e.g., the sampler developed for this work or, perhaps, the professional PyMC3 probabilistic programming framework).

LIST OF REFERENCES

Abbe, C., 1901: The physical basis of long-range weather forecasts. *Monthly Weather Review*, **29**, 551–61.

Aldrich, J., 1997: R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, **12**, 162–76.

Baran, S., and A. Möller, 2017: Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorology and Atmospheric Physics*, **192**, 99–112.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2008: Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–93.

Bishop, C. H. and K. T. Shanley, 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Monthly Weather Review*, **136**, 4641–52.

Bjerknes, V., 1904: Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik (The problem of weather forecasting as a problem in mechanics and physics). *Meteor. Z.*, **21**, 1–7. English translation by Y. Mintz, 1954, reproduced in *The Life Cycles of Extratropical Cyclones*, 1999, American Meteorological Society.

Bröcker, J., 2012: Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1611–17.

Bröcker, J. and L.A. Smith, 2007: Scoring probabilistic forecasts: on the importance of being proper. *Weather and Forecastin*g, **22**, 382–8.

Brooks, S. P. and A. Gelman, 1998: General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–55.

Casella, G., 2008: Bayesians and frequentists: models, assumptions, and inference. Accessed 1 August 2017, http://www.stat.ufl.edu/archived/casella/Talks/BayesRefresher.pdf.

CAWCR, 2017: Verification project. Accessed 01 August 2017, http://www.cawcr.gov.au/projects/verification/.

Charney, J. G., R. Fjørtoft, and J. von Neuman, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237–54.

Charney, J. G., 1951: Dynamical forecasting by numerical process. *Compendium of Meteorology*, T. F. Malone, Ed., *American Meteorological Society*, 470–82.

Danforth, C. M., 2013: Chaos in an atmosphere hanging on a wall. Accessed 21 December 2016, http://mpe2013.org/2013/03/17/chaos-in-an-atmosphere-hanging-on-a-wall/.

Dawid, A. P., 1984: Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Association*, **147**, 278–92.

Di Narzo, A. F. and D. Cocchi, 2010: A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society*, **59**, 405–22.

Diebold, F. X., T. A. Gunther, and A. S. Tay, 1998: Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39**, 863–83.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier and B. Yang, 2015: EMC implementation briefing of SREF.v7.0 (Q4FY15). Accessed 01 December 2016, http://www.emc.ncep.noaa.gov/mmb/SREF/SREFv7_implementationBriefing.pdf

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Weather Forecasting*, **13**, 1132–1147.

Epstein, E. S., 1969: Stochastic-dynamic prediction. *Tellus*, **21**, 739–59.

Feigelson, E., 2015: Fundamentals of classical & Bayesian inference. Accessed 1 August 2017, http://www2.astro.psu.edu/~edf/ESTEC_2015/4_Inference_regression.pdf.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309–68.

Gelfand, A. E. and A. F. M. Smith, 1990: Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. and D. Rubin, 1992: Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, 2013: *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Press, 675 pp.

Geman, S. and D. Geman, 1984: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, 1996: *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 486 pp.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11**, 1203–11.

Glickman, M. E. and D. A. van Dyk, 2007: Basic Bayesian methods. *Methods Molecular Biology*, **404**, 319–38.

Gneiting, T., A. E. Raftery, A. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–118.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, **69**, 243–68.

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–78.

Gneiting, T., 2014: Calibration of medium-range weather forecasts, ECMWF technical memorandum 719. European Centre for Medium-Range Weather Forecasts: Reading, UK. Accessed 01 April 2017, https://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf.

Good, I. J., 1952: Rational decisions. *Journal of the Royal Statistical Society, Series B*, **14**, 107–14.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, **132**, 1–17.

Gu, L., 2008: Multivariate Gaussian distribution. Accessed 15 December 2016, https://www.cs.cmu.edu/~epxing/Class/10701-08s/recitation/gaussian.pdf.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–60.

Hamill, T. M., 2007: Comments on ‘‘Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging.’’ *Monthly Weather Review*, **135**, 4226–36.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Monthly Weather Review*, **143**, 3300–09.

Hastings, W., 1970: Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97–109.

Hawking, S. W., 1999: Does God play dice? Accessed 30 June 2016, http://www.hawking.org.uk/does-god-play-dice.html.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–70.

Hodyss, D., E. Satterfield, J. McLay, T. M. Hamill, and M. Scheuerer, 2016: Inaccuracies with multi-model post-processing methods involving weighted, regression-corrected forecasts. *Monthly Weather Review*, **144**, 1649–68.

Hopson, T. M., 2014: Assessing the ensemble spread-error relationship. *Monthly Weather Review*, **142**, 1125–42.

Juban, J., L. Fugon, and G. Kariniotakis, 2007: Probabilistic short-term wind power forecasting based on kernel density estimators. *Proceedings of the European Wind Energy Conference*, Milan, Italy, 2007.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press, 342 pp.

Kim, M. J., J. S. Bertino Jr., T. A. Erb, P. L. Jenkins, A. N. Nafziger, 2004: Application of Bayes theorem to aminoglycoside-associated nephrotoxicity: Comparison of Extended-Interval Dosing, Individualized Pharmacokinetic Monitoring, and Multiple-Daily Dosing. *Journal of Clinical Pharmacology*, **44**, 696–707.

Klein, W. H. and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bulletin of the American Meteorological Society*, **55**, 1217–27.

Krishnamoorthy, A., and D. Menon 2013: Matrix inversion using Cholesky decomposition. Accessed 01 August 2017, https://arxiv.org/ftp/arxiv/papers/1111/1111.4144.pdf.

Kruschke, J. K., 2011: Thinning to reduce autocorrelation. Accessed 01 June 2016, http://doingbayesiandataanalysis.blogspot.com/2011/11/thinning-to-reduce-autocorrelation.html.

Kruschke, J. K., 2014: *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 776 pp.

Krzysztofowicz, R. and W. B. Evans, 2008: Probabilistic forecasts from the national digital forecast database. *Weather Forecasting*, **23**, 270–89.

Laplace, P. S., 1902: *A Philosophical Essay on Probabilities.* (F.W. Truscott and F.L. Emory, translated from the Sixth French Edition, First French Edition published in 1814). New York: J. Wiley, 196 pp.

Leigher, W. F., 2013: *U.S. Navy information dominance roadmap: 2013–2028*, Washington, D.C., Department of the Navy.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, **102**, 409–18.

Lin, P., 1972: Some characterizations of the multivariate t distribution, Journal of Multivariate Analysis, **2**, 339–44.

Liu, X. F. and M. J. Daniels, 2006: A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterizatio*n. Journal of Computational and Graphical Statistics*, **15**, 897–914.

Lorenz, E. N., 1962: The statistical prediction of solutions of dynamic equations. *Proceedings of the International Symposium on Numerical Weather Prediction*, Tokyo, 629–35.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20**, 130–41.

Luo, L., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *Journal of Geophysical Research*, **112**, D10102.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, **227**, 3431–44.

Manikandan, S., 2010. Data transformation. *Journal Pharmacology and Pharmacotherapeutics*, **1**, 126–7.

Matheson, J. E. and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–96.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, 1953: Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–92.

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir, 2013: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, **39**, 982–91.

Moore, G. E., 1965: Cramming more components onto integrated circuits. *Electronics*, **38**, 114–7.

Murphy, A. H., and R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bulletin of the American Meteorological Society*, **60**, 12–9.

Murphy, A. H. and R. L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–8.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–93.

Natarajan, V., 2008: What Einstein meant when he said god does not play dice. Resonance, *Journal of Science Education*, **13**, 651–5.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasseman, 1996: *Applied Linear Statistical Models*, Fourth Edition. New York: McGraw-Hill, 1408 pp.

NWS, 2015: Technical implementation notice 15–32. Accessed 15 August 2017, http://www.nws.noaa.gov/os/notification/tin15-32srefaae.htm.

NWS, 2017: CF6 product description. Accessed 15 August 2017, http://w2.weather.gov/climate/index.php?wfo=ilx.

Parzen, E., 1962: On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–76.

Pearl, J., 2000: The logic of counterfactuals in causal inference. *Journal of American Statistical Association*, **95(450)**, 428–35.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–74.

Rajagopalan, B., U. Lall, and S.E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Monthly Weather Review*, **130**, 1792–811.

Richardson, L. F., 1922: *Weather prediction by numerical process*. Cambridge University Press, 262 pp.

Richter, D., 2012: Bayesian ensemble model output statistics for temperature. Diploma thesis, Heidelberg University, Germany.

Robert, C. P. and G. Casella, 2011: A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, **26(1)**, 102–15.

Roberts, G. O., A. Gelman, and W. R. Gilks, 1997: Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–20.

Ron, A., 2010: Review of least squares solutions to overdetermined systems. Accessed 01 August 2017, http://pages.cs.wisc.edu/~amos/412/lecture-notes/lecture17.pdf.

Rosenblatt, M., 1956: Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*. **27(3)**, 832–7.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Weather Forecasting*, **19**, 552–65.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Monthly Weather Review*, **140**, 3204–19.

Sinay, M.S. and J. S. Hsu, 2014: Bayesian inference of a multivariate regression model. *Journal of Probability and Statistics*, **2014**, 1–13.

SREF, 2016: NCEP SREF plumes. Accessed 16 March 2016, http://www.spc.noaa.gov/exper/sref/srefplumes/.

Stansbury, D., cited 2012a: MCMC: the Metropolis sampler. Accessed 15 June 2016, https://theclevermachine.wordpress.com/2012/10/05/mcmc-the-metropolis-sampler/.

Stansbury, D., cited 2012b: MCMC: Multivariate distributions, block-wise, & component-wise updates. Accessed 01 December 2016, https://theclevermachine.wordpress.com/2012/11/04/mcmc-multivariate-distributions-block-wise-component-wise-updates/.

Stensrud, D. J., and N. Yussouf, 2003: Short-range predictions of 2-m temperature and dewpoint temperature over New England. *Monthly Weather Review*, **131**, 2510–24.

Stigler, S. M., 1990: *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, 432 pp.

Sun-tzu and S. B. Griffith, 1964: *The Art of War*. Oxford, Clarendon Press, 197 pp.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106**, 7183–92.

Taylor, K. E., 2005: Taylor diagram primer. Accessed 01 April 2017, http://www-pcmdi.llnl.gov/about/staff/Taylor/CV/Taylor_diagram_primer.pdf.

Thorarinsdottir, T. L. and T. Gneiting, 2010: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A*, **173**, 371–88.

Thorarinsdottir, T. L. and M. S. Johnson, 2011: Probabilistic wind gust forecasting using non-homogeneous Gaussian regression. *Monthly Weather Review*, **140**, 889–97.

Unger, D. A., H. Van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Monthly Weather Review,* **137**, 2365–79

Veenhuis, B., 2013: Spread calibration of ensemble MOS forecasts. *Monthly Weather Review*, **141**, 2468–82.

Vehtari, A., 2017: Prior choice recommendations. Accessed 01 August 2017, https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Vrugt, J., C. J. H. Diks, and M. P. Clark, 2008: Ensemble Bayesian model averaging using Markov-chain Monte Carlo sampling. *Environmental Fluid Mechanics*, **8**, 579–95.

Wang, X. and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society,* **131**, 965–986.

Waskom, M., 2015: Seaborn: Statistical data visualization. Accessed 21 December 2016, http://seaborn.pydata.org/tutorial/distributions.html#plotting-univariate-distributions.

Whitaker, J. S. and A.F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review,* **126**, 3292–302.

Wiecki, T., 2013: This world is far from normal(ly distributed): Bayesian robust regression in PyMC3. Accessed 01 August 2017, http://twiecki.github.io/blog/2013/08/27/bayesian-glms-2/.

Wiecki, T., 2015: While my MCMC gently samples: MCMC sampling for dummies. Accessed 01 June 2016, http://twiecki.github.io/blog/2015/11/10/mcmc-sampling/.

Wikipedia, cited 2017a: Multivariate normal distribution. Accessed 01 August 2017, https://en.wikipedia.org/wiki/Multivariate_normal_distribution.

Wikipedia, cited 2017b: Normal distribution. Accessed 01 August 2017, https://en.wikipedia.org/wiki/Normal_distribution.

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13**, 243–56.

Williams, R. M., C. A. T. Ferro, and F. Kwasniok, 2014: A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1112–20.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, **127**, 956–70.

Winters, H. A., G. E. Galloway Jr., W. J. Reynolds, and D. W. Rhyne, 1998: *Battling the Elements: Weather and Terrain in the Conduct of War*. John Hopkins University Press, 336 pp.

WxC, 2017: WxChallenge contest. Accessed 01 August 2017, http://wxchallenge.com/.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California