
# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# DISSERTATION

**AMBIGUITY IN ENSEMBLE FORECASTING: EVOLUTION, ESTIMATE VALIDATION AND VALUE**

by

Mark S. Allen

September 2009

Dissertation Supervisor: F. Anthony Eckel

**Approved for public release; distribution is unlimited**

| REPORT DOCUMENTATION PAGE | | | Form Approved OMB No. 0704-0188 |
|---|---|---|---|
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503. | | | |
| **1. AGENCY USE ONLY** ***(Leave blank)*** | **2. REPORT DATE** September 2009 | **3. REPORT TYPE AND DATES COVERED** Dissertation | |
| **4. TITLE AND SUBTITLE:** Ambiguity in Ensemble Forecasting: Evolution, Estimate Validation and Value | | **5. FUNDING NUMBERS** | |
| **6. AUTHOR(S)** Mark S. Allen | | | |
| **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 thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. | | | |
| **12a. DISTRIBUTION / AVAILABILITY STATEMENT** Approved for public release; distribution is unlimited | | **12b. DISTRIBUTION CODE** | |


**13. ABSTRACT (maximum 200 words)**

An ensemble prediction system (EPS) generates flow-dependent estimates of uncertainty (i.e., random error due to analysis and model errors) associated with a numerical weather prediction model to provide information critical to optimal decision making. Ambiguity, or uncertainty in the prediction of forecast uncertainty, arises due to EPS deficiencies, including finite sampling and inadequate representation of the sources of forecast uncertainty. An EPS based on a low-order dynamical system was used to investigate the behavior of ambiguity, validate two practical estimation methods against a theoretical (impractical) technique, and apply ambiguity in decision making. Ambiguity generally decreased with increasing lead time and was found to depend strongly on ensemble forecast variance and the variability of ensemble mean error. The practical estimation techniques provided reasonably accurate ambiguity estimates, although they were too low at early lead times. The theoretical ambiguity estimate added significant value when combining ambiguity with forecast uncertainty to provide a single normative decision input. Additionally, value added to secondary user criteria (e.g., minimizing repeat false alarms), was explored using the practical estimations. Repeat false alarms were significantly reduced while maintaining primary value by using ambiguity information to selectively reverse normative decisions to take protective action, which effectively redistributed negative outcomes.


| **14. SUBJECT TERMS** Ensemble Forecast, Ambiguity, Uncertainty, Ensemble-of-Ensemble, Calibrated Error Sampling, Randomly Calibrated Resampling, Optimal Decision Making, Cost-Loss, Uncertainty-Folding, Secondary Criteria, Lorenz '96, Ensemble Prediction Systems | | | **15. NUMBER OF PAGES** 237 |
|---|---|---|---|
| | | | **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 |

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

**AMBIGUITY IN ENSEMBLE FORECASTING: EVOLUTION, ESTIMATE VALIDATION AND VALUE**

Mark S. Allen
Major, United States Air Force
B.S., Florida State University, 1998
M.S., Air Force Institute of Technology, 2003

Submitted in partial fulfillment of the
requirements for the degree of

**DOCTOR OF PHILOSOPHY IN METEOROLOGY**

from the

**NAVAL POSTGRADUATE SCHOOL**
**September 2009**

Author: _____
Mark S. Allen

Approved by:

_____
F. Anthony Eckel
Military Instructor
Dissertation Supervisor/
Dissertation Committee Chair

_____
Wendell Nuss
Professor of Meteorology

_____
Patrick Harr
Professor of Meteorology

_____
Eva Regnier
Associate Professor of
Decision Science

_____
James Hansen
Naval Research Lab, Monterey

Approved by: _____
Philip Durkee, Chair, Department of Meteorology

Approved by: _____
Douglas Moses, Vice-Provost for Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

An ensemble prediction system (EPS) generates flow-dependent estimates of uncertainty (i.e., random error due to analysis and model errors) associated with a numerical weather prediction model to provide information critical to optimal decision making. Ambiguity, or uncertainty in the prediction of forecast uncertainty, arises due to EPS deficiencies, including finite sampling and inadequate representation of the sources of forecast uncertainty. An EPS based on a low-order dynamical system was used to investigate the behavior of ambiguity, validate two practical estimation methods against a theoretical (impractical) technique, and apply ambiguity in decision making. Ambiguity generally decreased with increasing lead time and was found to depend strongly on ensemble forecast variance and the variability of ensemble mean error. The practical estimation techniques provided reasonably accurate ambiguity estimates, although they were too low at early lead times. The theoretical ambiguity estimate added significant value when combining ambiguity with forecast uncertainty to provide a single normative decision input. Additionally, value added to secondary user criteria (e.g., minimizing repeat false alarms), was explored using the practical estimations. Repeat false alarms were significantly reduced while maintaining primary value by using ambiguity information to selectively reverse normative decisions to take protective action, which effectively redistributed negative outcomes.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

# LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| *BS* | Brier Score |
| *BSS* | Brier Skill Score |
| BV | Bred Vector |
| CDF | Cumulative Density Function |
| $CES_G$ | Calibrated Error Sampling – Global |
| $CES_L$ | Calibrated Error Sampling – Local |
| CI | Confidence Interval |
| *C/L* | Cost-Loss Ratio |
| CONUS | Continental United States |
| DA | Data Assimilation |
| ECMWF | European Center for Medium-Range Weather Forecasts |
| EF | Ensemble Forecasting |
| EnKF | Ensemble Kalman Filter |
| EoE | Ensemble-of-Ensemble |
| EPS | Ensemble Prediction System |
| ETKF | Ensemble Transform Kalman Filter |
| GEFS | Global Ensemble Forecast System |
| GFS | Global Forecast System |
| IC | Initial Condition |
| *IOVS* | Integrated Optimal Value Score |
| KF | Kalman Filter |
| L96M | Lorenz 96 Model |
| L96S | Lorenz 96 System |
| LAM | Limited Area Model |
| LBC | Lateral Boundary Condition |
| *ME* | Mean Error |
| *MSE* | Mean Square Error |
| NCEP | National Center for Environmental Prediction |

| | |
|---|---|
| NWP | Numerical Weather Prediction |
| PDF | Probability Density Function |
| *POD* | Probability of Detection |
| POFD | Probability of False Detection |
| *POMD* | Probability of Missed Detection |
| RCR | Randomly Calibrated Resampling |
| *rel* | Reliability Component of *BSS* |
| *res* | Resolution Component of *BSS* |
| RK2, RK4 | $2^{nd}$ or $4^{th}$ Order Runga-Kutta |
| *RMSE* | Root Mean Square Error |
| *ROC* | Relative Operating Characteristic |
| SREF | Short-Range Ensemble Forecasting |
| SST | Sea Surface Temperature |
| SV | Singular Vector |
| TIGGE | THORPEX Interactive Grand Global Ensemble |
| THORPEX | The Observing System Research and Predictability Experiment |
| UKMO | United Kingdom Meteorological Office |
| VOP | Verification Outlier Percentage |
| VRH | Verification Rank Histogram |
| *VS* | Value Score |

# LIST OF SYMBOLS

| | |
|---|---|
| $a$ | Number of hits from a contingency table of forecast-observation pairs |
| $b$ | Number of false alarms from a contingency table of forecast-observation pairs |
| $c$ | Number of misses from a contingency table of forecast-observation pairs |
| $d$ | Number of correct rejections from a contingency table of forecast-observation pairs |
| $e$ | Single ensemble member |
| $\overline{e}$ | Ensemble forecast mean |
| $\tilde{e}$ | Single bias-corrected ensemble member |
| $e^*$ | Single bulk calibrated ensemble member (first- and second-moment) |
| $E(\ )$ | Expected value |
| $\mathbf{H}$ | Linear transformation matrix |
| $\boldsymbol{H}$ | Nonlinear transformation matrix |
| $\hat{\mathrm{K}}$ | Estimated Kalman gain |
| $M$ | Total number of forecast-observation pairs used for verification |
| $\boldsymbol{M}$ | Nonlinear dynamical model |
| $n$ | Number of ensemble members (ensemble size) |
| $N$ | Number of EoE constituents |
| $\overline{o}$ | Sample climatological rate of occurrence |
| $p_a$ | Uncertainty-folding forecast probability combining first- and second-order uncertainty |
| $p_e$ | Forecast probability |
| $p_e^*$ | Calibrated forecast probability |
| $p_g$ | Grand ensemble forecast probability |
| $p_T$ | True forecast probability |

| | |
|---|---|
| $\hat{p}_T$ | Estimate of the true forecast probability – single element of an estimated ambiguity distribution |
| $\hat{\mathbf{P}}^b$ | Estimated background error covariance matrix |
| $r(\delta)$ | Relative frequency of $\hat{p}_T$ values within each bin used during uncertainty-folding – represents the second-order uncertainty |
| $\mathbf{R}$ | Observation error covariance matrix |
| $\mathrm{x}^a$ | Single analysis state vector |
| $\mathrm{x}^b$ | Single background state vector |
| $\mathrm{x}'^b$ | Single background ensemble perturbation from the ensemble mean |
| $\mathrm{X}^b$ | Ensemble of background state vectors |
| $X_k$ | Large-scale, resolved variables in the L96 system |
| y | Single observation |
| $Y_j$ | Small-scale, unresolved variables in the L96 system |
| $\alpha$ | *C/L* ratio |
| $\delta$ | Representative value of bins used during uncertainty-folding – represents the first-order uncertainty |
| $\sigma'$ | Fractional error in ensemble spread |
| $\sigma_C^2$ | Climatological variance |
| $\sigma_e$ | Ensemble forecast spread (standard deviation of ensemble members) |
| $\overline{\sigma_e^2}$ | Average ensemble variance |
| $\tau$ | Tau or forecast lead time |
| $\Theta$ | Event threshold |

# ACKNOWLEDGMENTS

There are several people I need to thank for their help and support while I labored through my research and writing this dissertation. First, I could not have accomplished this daunting achievement without the love and support of my family, especially my parents whose prayers and unwavering confidence in my ability to finish kept me going. I am deeply indebted to my advisor Major Tony Eckel, who acted as a teacher, mentor and friend while constantly pushing me to produce the best research. Additionally, I am thankful for the overwhelming technical and motivational support from the rest of my committee, Eva Regnier, Wendell Nuss, Patrick Harr and Jim Hansen. I would be remiss to fail to thank Bob Creasey for his help with computer systems. Finally, I must thank my officemate and fellow PhD student Major Lou Lussier and his family for helping me maintain my sanity. Their house was a home away from home for many Sunday dinners, and Lou and I spent many hours discussing research ideas and blowing off steam.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

The primary tool for weather forecasters today is the Numerical Weather Prediction (NWP) model, and ensembles are rapidly gaining momentum as the preferred application. Ensemble forecasts provide an estimation of the uncertainty associated with NWP forecasts, but at this time the forecast is typically employed without consideration of the uncertainty associated with the ensemble's prediction of uncertainty. This research is focused on exploring methods to objectively quantify the uncertainty in an ensemble forecast and determine the value of knowing that information.

Over many years, the mold has been cast for using NWP models for deterministic forecasting, i.e., using a single model forecast to convey the future state of the atmosphere. Although great improvements have been made since the birth of NWP (e.g., increased computing power, better model physics, finer grid scales and improved numerical methods), the deterministic application of NWP still produces forecasts with a great deal of uncertainty (Leutbecher and Palmer 2007). We can't get around the fact that even small errors in the initial conditions grow to produce large forecast errors (Lorenz 1969). Thus, deterministic NWP may not be the most effective approach. Improvement of the NWP model can provide only finite improvement in forecast quality (Brooks and Doswell 1993; Lorenz 1993). Ensemble forecasting was introduced as a means of objectively characterizing the uncertainty in NWP forecasts. It involves running multiple, parallel models (*members*), where each member has perturbations to the initial conditions and the model. An ideal ensemble prediction system (EPS) includes perturbations in the initial conditions that capture all possible errors in the analysis, as well as model perturbations representing all possible model errors, which requires an infinite number of members.

Ensemble forecast information has several applications, including predicting deterministic forecast skill (via the ensemble spread) and improving deterministic forecast skill (via the ensemble mean) (Eckel 2008). The definitive application of ensemble forecasts is production of a forecast probability of occurrence for a specific

event (e.g. temperature < 0°C), which can have high value in the decision making process (Eckel 2008). Numerous studies have shown the value of using probabilistic decision inputs over using deterministic or climatological information (e.g., Katz and Murphy 1997; Richardson 2000; Palmer 2002; Zhu et al. 2002). The problem that is largely overlooked at this time is the uncertainty associated with the ensemble forecast itself (Eckel and Allen 2009). Uncertainty in the ensemble forecast is due to design and computational restrictions that preclude running an ideal EPS. Today's EPSs use finite number of ensemble members and inadequate representation of the uncertainty associated with the initial conditions and model design. Thus, there is uncertainty in the estimation of forecast uncertainty, a phenomenon termed ambiguity. Ambiguity has been considered in formal decision science for many years and is generally studied in the vein of understanding people's attitudes towards ambiguity in the decision, or ambiguity aversion (Ellsberg 1961; Camerer and Weber 1992). In these studies, the decision-maker's estimate of the uncertainty is typically subjective (Camerer and Weber 1992; Wallsten 1990). Application of objectively estimated second-order uncertainty to optimize decisions was not attempted.

The main objectives of this research are to: (1) understand the mechanisms behind the evolution of ambiguity associated with an ensemble forecast, (2) validate objective estimates of ambiguity associated with an EPS, and (3) explore methods of applying the ambiguity information in order to add value in decision making.

This dissertation is organized into five chapters, including this Introduction. The Background chapter (Chapter II) provides an overview of basic ensemble forecasting theory with a more in-depth look at sources of error in the EPS directly relating to ambiguity. In addition, Chapter II reviews the methods used during this research to determine the value of the ambiguity information in decision making. Chapter III provides the Methodology used to accomplish the three research objectives, including the NWP model and EPS design. Results of the behavior, validation, and value studies are presented in Chapter IV. Finally, conclusions and future research are addressed in Chapter V.

# II. BACKGROUND

## A. ENSEMBLE FORECASTING

Since the advent of Numerical Weather Prediction (NWP) with the first successful 24-hour forecasts by Charney and his group in 1949, the primary role of NWP has been to produce deterministic prognoses of the future state of the atmosphere (Lewis 2005). Over the following decades, the chaotic nature of the atmosphere has come to be understood by meteorologists, ushering in a new paradigm for atmospheric prediction. First conceived by Poincare in 1914 and later proven by Lorenz in his seminal paper in 1963 (Eckel 2008; Lorenz 1963), chaos describes the behavior of nonlinear dynamical systems. What appears to be randomness in the evolution of the deterministic system is actually the result of sensitive dependence to initial conditions (ICs). Small errors in the ICs are evolved according to the system's (deterministic) rules, and these errors grow nonlinearly with increasing forecast lead time. Ultimately, the error grows so large that the forecast is no better than one conceived using past observational data (i.e. climatology). At this point, the limit of predictability has been reached.

Observations of the current state of the atmosphere cannot accurately represent the current conditions at all points and on all scales. Thus, even if our NWP models were perfect, error in the ICs would render the forecasts useless after a short time. As an added complication, our NWP modeling systems are not perfect in that they cannot represent atmospheric phenomena on all spatial or temporal scales, forcing modelers to approximate many subgrid scale, unresolved processes. Thus even given perfect ICs, model deficiencies would again result in nonlinear error growth and limit predictability.

Forecasts of the future states must be looked at as uncertain events where there exists some chance of occurrence (Eckel 2008). The concept of ensemble forecasting (EF) was first introduced by Leith in 1974 (Leith 1974; Lewis 2005). Leith proposed using multiple perturbed NWP runs to produce a limited sample of possible future states. By using the mean value of forecasts from approximately 10 different NWP runs, Leith

was able to show improvements in forecasts with lead times out to 10 days (Sivillo et al. 1997). While requiring large computational resources, EF was seen as a viable method to estimate forecast uncertainty.

An EF is essentially a group of concurrent NWP forecasts, where each *member* of the ensemble is run using slightly different (perturbed) ICs and perturbations to the NWP model. The purpose of the ensemble forecast is to simulate the error growth associated with errors in the analysis of the current state and deficiencies in the NWP model, and to produce a sample of likely forecast states (Eckel 2008). Separating these two error sources in real-world ensemble prediction systems (EPS) may not be possible, as the first-guess used during the data assimilation (DA) process to produce the analysis for the next model run is typically a forecast state from the previous model run. This forecast state (*background*) is then updated using observations to nudge it closer to the current observed state of the atmosphere, ultimately providing an *analysis* of the current state that is more precise than either the observations or the background.

Anderson (1996), Eckel (2008), Toth and Kalnay (1993), and Traction and Kalnay (1993) describe the basic applications of EF data:

- EF mean accuracy is better on average than deterministic NWP;
- EF spread gives the confidence in a single deterministic NWP model run;
- Solution clusters can aid in narrowing the most likely evolution;
- Forecast probability of occurrence for some event can be calculated from the distribution of EF members.

Forecast probability is the ultimate product of EF data, since it provides the forecast user with objective uncertainty information regarding the event in question. The user can then complete a thorough risk analysis and optimize decision making.

There are many sources of error in NWP within the two general types, analysis error and model error. An ideal EPS will account for all sources of uncertainty associated with its modeling system. Any EPS deficiencies (errors not accounted for) result in errors in the ensemble forecast probability density function (PDF). If the ensemble forecast PDF is wrong, then measures of uncertainty in the forecast will be incorrect, including forecast probability. Sources of error in the ensemble forecast PDF include

limited sampling and poor simulation of IC errors and model errors. Model errors can be associated with the numerical techniques used in NWP as well as inaccuracies or uncertainty in subgrid scale, unresolved processes due to unrepresentative parameterizations or inadequate model resolution. In general, error can be separated into two categories, *systematic* and *stochastic*. Systematic error is bias or errors that consistently repeat. Stochastic error describes the variance of error about systematic error. Systematic and stochastic errors occur in all moments of the ensemble forecast PDF.

The following sections describe current state-of-the-art techniques used by operational forecast centers to generate IC and/or model perturbations in an EPS, while focusing on the limitations of the techniques and thus their contributions to stochastic error and ultimately ambiguity. Additionally, different sources of model error, horizontal resolution and the implications of limited sampling on ensemble forecasting are discussed.

## 1. Accounting for Analysis Error—IC Perturbations

Analysis error is any difference between the estimated and the true state of the system at initialization of the NWP model. Analysis error may result from errors in the observations due to instrument limitations or the inability to observe at all spatial and temporal scales. Additionally, analysis error may come about in data assimilation when an erred forecast state from a previous model run (i.e., the background) is combined with the observations. Also, when the background and observation information are combined, error may be introduced through interpolation or variable transformation. An analysis may be considered perfect (i.e., all grid point values accurately represent the average conditions within the grid box) and still be in error since it cannot represent sub-grid scale conditions or the numerical precision of actual atmospheric variable values.

The model analysis is a hyperdimensional vector containing the values of all state variables, where the values describe the instantaneous state of the system in phase space (i.e., a region where all state variables are represented by a unique dimension) (Eckel 2008). A perturbation or change to any state variable results in a change of location of

the state in phase space, which may be described as a change in the *direction* of the hyperdimensional vector pointing to the instantaneous state from a fixed origin. In ensemble forecasting, IC perturbations produce possible analysis states within the model's attractor (i.e., the collection of all naturally occurring states of the model in phase space) that are consistent with the analysis error covariance and structured to simulate the fastest error growth based on the analyzed state of the system, for example, perturbing the location of a baroclinic zone. Thus the goal is to produce perturbed ICs that are equally likely estimates of the true state that cover all scales of motion and lead to accurate simulation of error growth associated with the current flow (Eckel 2008).

Properly representing the sources of uncertainty relevant to the current flow is an important aspect of EF. Purely random ICs used with a finite member EPS will likely not adequately represent the analysis uncertainty, as error growth associated with many members will be too slow or even decrease early in the forecast (Magnusson et al. 2008). The generation of ICs for EF is intended to provide a range of analysis states that allow the EF solution to adequately disperse given the current uncertainty in the analyzed state. Given a perfect model, the $n$ members' states of the EF should encompass the true forecast state at some later lead time at a rate of $(n-1)/(n+1)$ (Eckel 2008). Several varying techniques are currently in use at the major operational forecast centers, but these techniques can be divided into two categories (Leutbecher and Palmer 2008—hereafter LP08).

The first category is described by LP08 as techniques designed to produce perturbed ICs using ensemble-based DA, such as the Ensemble Kalman Filter (EnKF) with perturbed observations. In this method, employed by the Canadian Meteorological Service, multiple DA cycles are performed using observations perturbed by random noise simulating observational error (LP08). EnKF produces an ensemble of analysis states that can be used as EF ICs, where the ensemble of perturbed analyses is created by optimally combining the perturbed observations with an ensemble of perturbed forecasts. Also, the mean of the EnKF member states may be used as the best-guess analysis from which to start a single NWP model run. EnKF is discussed in more detail in Chapter III.A.3. A limiting factor for the EnKF is estimation of background error covariance used

when updating the ensemble of perturbed forecasts. An EnKF ensemble that is too small may result in spurious, unrealistic correlations between locations in the model domain giving a noisy estimate of the background error covariance (Hamill et al. 2001; Lorenc 2003). Also, small ensemble sizes may result in background error covariance estimates that are too small leading to non-optimal estimates of the Kalman gain (Hamill et al. 2001; Lorenc 2003). Covariance estimate errors may also be introduced by errors in the NWP model used to integrate the EnKF members. The background error covariance problems described and/or a lack of quality, timely observations may lead the EnKF analyses to drift away from the true state of the system resulting in large analysis error.

The UK Meteorological Office (UKMO) uses a technique called the Ensemble Transform Kalman Filter (ETKF). The ETKF uses a transformation matrix to transform an ensemble of perturbed forecast states into an ensemble of perturbed analysis states (Wang and Bishop 2003). The transformation matrix rotates and scales the forecast perturbations based on observational information producing orthogonal analysis perturbations exhibiting variance that satisfies the Kalman filter error covariance update equation (Wang and Bishop 2003; Wei et al. 2006). The formulation of the ETKF does not allow it to be used to produce a best-guess analysis, so it must be used in conjunction with some DA technique (LP08; Wang and Bishop 2003). In this case, the background error covariance matrix used in DA will not strictly match the covariance matrix developed using the ensemble leading to errors in the estimate of the analysis error covariance since the ETKF assumes the matrices match (Wei et al. 2006). Like the EnKF, this perturbation generation method is sensitive to the ensemble size, thus covariance inflation is necessary to prevent underestimation of analysis error covariances for small ensembles (Wei et al. 2006). Underestimation of the analysis error covariance is also possible if model error is neglected. Importantly, the transformation matrix and the inflation factor as discussed by Wang and Bishop (2003) and Wei et al. (2006) are sensitive to the spatial and temporal variability of the observation network used. Routine changes in the observation density in an operational observation network can greatly affect the accuracy of the ETKF.

The second category of techniques includes those that attempt to select perturbations capturing the greatest error growth over some forecast period. According to LP08, the techniques "selectively sample initial uncertainty only in directions that are dynamically most important for determining ensemble dispersion." It is assumed that the growing modes found will continue to show the largest error growth beyond the forecast period used for selection. The bred vectors (BV) method, currently used in the National Center for Environmental Prediction's (NCEP) short-range EF (SREF), is in this category. In BV, a random perturbation is applied to an initial state, and both the perturbed and original states are evolved forward using the NWP model over some forecast period. At the end of the forecast period, the vector difference between the perturbed and original states is found. This difference vector is rescaled to match the typical analysis error magnitude and then used to perturb a new initial state. After several repeated cycles, the final perturbation direction is found. This process is repeated using several random perturbations to find multiple final perturbations that are used as the ensemble ICs (LP08). The BV method is limited by the fact that it attempts to find only the perturbations responsible for the greatest error growth, whereas other perturbation directions may also be important (Eckel 2008). In addition, the perturbation rescaling process can introduce errors, thus a regional or variable dependent rescaling may be necessary. Rescaling only certain variables that exceed a global analysis error value changes the direction of the hyperdimensional state vector describing the system and changes the direction of the perturbation found using BV (Eckel 2008).

The European Centre for Medium-Range Weather Forecasts (ECMWF) employs a technique that falls into the second category termed singular vectors (SV). The SV method finds the leading singular vectors or directions of maximum growth based on a linear version of the NWP model over some optimization period, typically taken as 48-hours for ECMWF ensemble forecasts (LP08). In other words, SV determines the directions of initial uncertainty that lead to the largest forecast uncertainty dynamically constrained by the NWP model (LP08). SV are sensitive to the choices made for the length of the optimization period and the norm used to evaluate the magnitude of the vector (e.g., Euclidean norm or total energy norm) (Kalnay 2003). Thus very different

SV may result from varying these two parameters. Another limitation of SV is the assumption of linear error growth over the optimization period (ECMWF 2009) requiring the use of a tangent-linear version and adjoint of the full, nonlinear NWP model. The tangent-linear model employed at ECMWF uses a simplified physics package without physical parameterizations (except for simple vertical mixing and friction), which may also result in suboptimal SV perturbations due to model deficiencies (LP08). Magnusson et al. (2008) found that SV are best for shorter time-scale forecasts, as their effectiveness degrades at longer time scales. Presumably, at longer lead times the SV associated with maximum error growth are different than those calculated over the optimization period.

Comparison studies performed to determine if one method or category of IC generation techniques is superior have had mixed results. Using current operationally produced data, it is difficult to separate the techniques from the numerical models they are applied to, which are of varying quality, thus no conclusive results have been found (LP08). In idealized studies, more interesting and informative comparisons between the categories have been achieved. Houtekamer and Derome (1995) found that the techniques in each of the categories produced equally skillful ensemble mean forecasts. Hamill et al. (2000) found the ensemble-based methods had superior statistical consistency (defined by Anderson 1996, 1997 and Talagrand et al. 1997), mainly early in the forecast period. In a more recent study, Descamps and Talagrand (2007) analyzed the skill and statistical consistency of ensemble forecasts made using a model of a low-order dynamical system as well as a quasi-geostrophic model in a perfect model context using EnKF, ETKF, BV and SV initial conditions. They found the skill of the ensemble mean was significantly higher for the EnKF and ETKF forecasts. Statistical consistency and other forecast skill and quality tests (i.e., Brier score and relative operating characteristic) also showed significant improvement when using the ensemble-based methods. These results confirmed tests by Bowler (2006) who found EnKF outperformed SV and BV in an EPS based on the same low-order model used by Descamps and Talgrand.

## 2. Accounting for Model Error—Model Perturbations

Model error is any difference between the model attractor and the true atmospheric attractor resulting from the design of the NWP model, including limits in model resolution, mathematical formulation, physics, and lateral and surface boundary conditions. For example, parameterizations are used within the NWP model to account for the effects of subgrid scale, unresolved processes on the forecast evolution. In some cases, the parameterized process may not be well understood or the availability of observational studies used to develop or train the parameterization may be limited. In these cases, forecast uncertainty may be high when the forecast trajectory is sensitive to the parameterization errors. The aim of perturbing the model is to introduce equally likely perturbations that represent probable model error covering all scales of motion (Eckel 2008), thus providing model diversity during the ensemble forecast that adequately represents the current flow's sensitivity to the model error.

Although a majority of the research into proper perturbations for an EPS has been focused on generation of ICs, the significance of model deficiencies to uncertainty in the ensemble forecast cannot be overlooked. Accounting for model error using one of the techniques described below can increase dispersion and improve overall skill particularly for surface, sensible weather phenomena of concern to users (Eckel 2003; Mylne et al. 2002). EPSs that do not account for model error are necessarily under-dispersive, as the model attractor does not mimic the true system attractor.

Descamps and Talagrand (2007) expanded their study of the quasi-geostrophic EPS to include model error. Once again, they found the ensemble-based IC perturbation techniques performed the best, but the skill and consistency of all methods was reduced by the introduction of model error. Their results also indicated that the gains made by using the ensemble-based IC generation methods did not last as long into the forecast period when model error was introduced. The authors explain this result as a consequence of rapidly growing transient instabilities (errors) in the flow generated early in the forecast. Thus at later lead times, deficiencies in the underlying forecast model

may rapidly dominate over the quality of the initial conditions in regards to forecast uncertainty. Therefore, uncertainty in the forecast due to model deficiencies must be accounted for.

### *a. Basic Techniques*

There are three basic techniques used to account for model error in an EPS. The first technique is called *stochastic-physics* and is used at ECMWF. Buizza et al. (1999) describe stochastic-physics as randomly perturbing the tendency of the state variables during integration with "some appropriate degree of spatio-temporal autocorrelation." The state variables are perturbed in an attempt to capture the influence of parameterization errors. Studies by Evans et al. (2000), Ziehmann (2000), and Richardson (2001) indicate this technique has limited effectiveness, likely due to each ensemble member using the same model attractor resulting in limited diversity. Random changes to the state variables move a member's trajectory off of the attractor, but it then converges back rapidly (Eckel 2003). Backscatter is another stochastic method used to account for unrepresented dynamical processes in the NWP model, where energy at subgrid scales is excited and transferred up-scale to resolved scales (Shutts 2005). Shutts provides support for the argument that energy dissipation in NWP models is excessive, thus arguing the need for local up-scale kinetic energy transfer. He found an improvement in ensemble spread, consistency and skill using an EPS with stochastic-physics and stochastic backscatter, while acknowledging the stochastic-physics contribution to increased spread was "small but consistently positive." Backscatter is limited by our ability to accurately estimate atmospheric energy dissipation and local up-scale energy transfer (Shutts 2005), which naturally leads to errors when exciting energy transfer in the NWP model.

The next model perturbation technique is termed *perturbed model*. In this method, a single NWP model is used, but parameterizations within the model are perturbed for each member. Understanding the uncertainty in the model associated with the parameterizations is a difficult question. Model parameterizations are perturbed within some estimate of the parameter uncertainty with the assumption that the correct

mean tendency can be achieved from one of the perturbations (LP08). In this way, it is assumed that the distribution of possible forecast states will encompass the true forecast state. However, each member shares many of the same model design features (i.e., the model core) and may not adequately reflect the uncertainty in the current flow (Eckel 2003). Using a stochastic (randomly perturbed) parameterization in a low-order model, Wilks (2005) found the stochastic parameterization outperformed a deterministic parameterization in representing the climatology of the true system, ensemble mean performance, and ensemble dispersion. However, the perturbed model EPS is limited by our understanding of the parameterized processes and thus our estimate of the associated uncertainty and the sensitivity of the forecast to errors in a given parameterization. Even if a parameterization is perturbed accurately, model error is inevitable since the single parameter value is used to represent a continuous spectrum of possible true values for a single model grid box.

The final approach used to account for model error is the *multi-model* technique, where each ensemble member is based on a different NWP model or model configuration. For example, two members of a multi-model EPS can be the NCEP and ECMWF control forecasts, or they may both be from the same model where a different convective parameterization is used in each. The different models will generally have different numerical schemes, physics schemes and parameterizations. In this way, each member model has a distinct attractor increasing ensemble dispersion. It has been shown that differences in skill among members for a given forecast is not a problem, as this also adds diversity to the distribution of forecast solutions. The primary assumption is that on any given day, any of the ensemble members has the potential to outperform the others. Thus a model that consistently exhibits low individual skill may not add skill to the ensemble forecast. Mylne et al. (2002) found that a multi-model EPS improves skill, while Ebert (2001) showed that multi-model ensembles were less likely to suffer from under-dispersion due to systematic errors. Multi-model ensembles are limited by the availability of different NWP models or by computational restrictions that do not allow all possible combinations of model configurations. When using a multi-model ensemble

where several members are designed around a single model core, similarities between the ensemble members will reduce model diversity.

Error in the NWP model can come from many different sources. The model perturbation techniques described approach the sources of model error from different perspectives in an attempt to simulate forecast uncertainty. Thus, an ideal EPS should use all of these methods in conjunction with one another to achieve the greatest model diversity and the most accurate estimate of forecast uncertainty. Additional sources of NWP model error that must be accounted for are described in the following sections.

### *b. Boundary Conditions*

Model boundary conditions are a significant source of model error that is normally accounted for separately (from the above basic techniques) in an ensemble. This source of error includes the handling of lateral boundary conditions (LBC) for a limited-area model (LAM) as well as the surface and upper boundaries of any NWP model. A LAM requires the use of LBC updates during model integration, generally supplied by a global NWP model, to transfer information from outside the LAM across the boundary. An EPS based on a LAM must perturb the LBCs to capture uncertainty flowing across the boundary into the LAM domain. Studies described in Nutter et al. (2004a and 2004b) indicated that LAMs showed greater sensitivity to changes in LBCs than in ICs, and that a SREF using perturbed LBCs produced forecasts with improved dispersion. LBCs may be taken from the members of a single- or multi-model, global EPS, where the differences between the global LBCs provide the perturbations. Perturbations to LBCs are limited by the coarse spatial and temporal availability of global information used to update the boundaries, thus missing mesoscale variability (Eckel 2008). Nutter et al. showed how the limitations may be mitigated using dynamically consistent, finescale random perturbations mimicking error growth at every time step between LBC updates, but there are no operational centers currently applying this technique.

The surface boundary is another aspect of NWP modeling that plays a significant role in producing model error. Surface boundary parameters, such as soil moisture, soil type, vegetation type and fraction, and snow cover for example, impact the evolution of the atmosphere. While these variables are continuous in nature, they are accounted for in the model using two-dimensional surface fields providing representative values on the NWP model grid (e.g., using seasonally-based land use tables). Thus the surface boundary parameters are a source of random error since they cannot accurately represent conditions at all scales and sensitivity to parameter errors are unknown for any given forecast. Another significant source of error at the surface boundary is the sea surface temperature (SST) field used in the NWP model. Most operational NWP modeling systems are not coupled to an ocean model and use only a static SST analysis throughout the forecast (Kalnay 2003). Perturbations to the surface boundary parameter fields within the estimated uncertainty may be used in an EPS to account for sensitivities associated with variations (Eckel and Mass 2005). The methods used to account for uncertainty in the formulation of surface boundary parameters are limited, as many of the surface processes taking place may not be well understood or well observed (spatially and temporally). Initializing and estimating the uncertainty associated with these processes in order to perturb them properly is difficult, and incomplete parameter field tables may potentially omit significant parameters.

Model error can also be produced by interactions at the model's upper boundary, where assumptions must be made regarding the evolution of conditions above the model's top during integration. In addition, the rigid lid or constant pressure surface employed by most NWP models results in gravity wave reflection, which can impact forecast conditions throughout the depth of the model. The gravity wave effects may be mitigated by absorption, damping or other techniques at the upper boundary. At this time, no operational EPSs consider the uncertainty associated with upper boundary conditions or interactions at the boundary during the forecast.

### *c. Horizontal Resolution*

Another critical source of model error is the effect of subgrid scale or unresolved *dynamical* processes. Grid point models can adequately resolve features on the scale of 7-8 grid points (Kalnay 2003), which leads to many dynamical processes taking place below the resolution of the model. These unresolved and therefore unaccounted for processes add uncertainty to the forecast, resulting in errors in the forecast PDF and under-dispersion in that the range of forecast states cannot reach the range of possible true states. Although the stochastic physics techniques described previously (Chapter II.A.2.a) attempt to account for these errors, they have been shown to produce marginal improvements and cannot fully simulate the subgrid scale uncertainty. Additionally, the model attractor will never exactly mimic the atmospheric attractor due to its limited dimensionality (i.e., resolution). Thus any given model state, where the value of the state variables at each grid point are the mean values for the grid box, may actually map to many different true atmospheric states creating model error.

Increasing horizontal resolution has been shown to improve the skill of the ensemble mean (Szunyogh and Toth 2002), thus providing a forecast PDF with reduced random error in location. An ensemble study conducted by Mullen and Buizza (2002) centered on the impacts of horizontal resolution and ensemble size on precipitation forecasts found a higher-resolution model performed better than a lower-resolution model for multiple consistency and skill measures (e.g. rank histograms, *BSS*, *ROC*). Their findings also indicated that using a lower resolution model while increasing ensemble size can outperform an EPS using higher resolution and fewer members, especially when forecasting rare events. However, given an equal number of members, the higher-resolution EPS will perform better.

Toth et al. (2002) assert the true value of increasing horizontal resolution is found when applied to ensemble forecasting. Their study showed improved ensemble mean and probabilistic forecasts for 500 hPa geopotential heights in the northern and southern hemisphere extratropics. Noise associated with small-scale features that have

lost predictability interacting with larger-scale, predictable features realistically represent natural processes and improves the ensemble's performance.

### 3. Limited Sampling

Computational constraints on an operational EPS force forecast centers to limit the number of ensemble members to ensure timely delivery of forecast products. An ensemble with few members cannot consistently reproduce the forecast PDF from which they are drawn (Figure 1). The mean and spread error for any one case due to limited sampling cannot be known *a priori*. Sampling distributions of error in the ensemble mean and error in the ensemble spread based on random draws from an $N(0,1)$ distribution for different ensemble sizes, show that error can vary greatly, especially for small ensembles (Figure 2). For both distributions in Figure 2, the potential error in the statistics decreases with increasing ensemble size, indicating that increased sampling provides a better estimate of the forecast PDF (Wilks 2006). Error in the estimated PDF due to limited sampling decreases exponentially with increasing ensemble size. The exponential decrease naturally leads to a leveling off of improvements to skill, where the added benefit may no longer justify the additional expense of adding more members. A similar effect is seen when comparing the skill of ensemble probability forecasts.

While the techniques described for perturbing the initial conditions and the NWP model in an EPS are sophisticated compared to purely random perturbations, they are still limited in their ability to fully cover the spectrum of possible error sources associated with an NWP modeling system. Even if an EPS were perturbed perfectly, limited sampling would generate random error in the forecast PDF. The inescapable existence of random or seemingly random error in the ensemble forecast means that ambiguity is inevitable in predictions of forecast uncertainty.

## B. AMBIGUITY

In general terms, ambiguity, or second-order uncertainty, can be described as the uncertainty associated with estimates of uncertainty (NRC 2006). Camerer and Weber (1992) defined ambiguity as "uncertainty about probability, created by missing

information that is relevant and could be known." Ensemble-based forecast probability provides uncertainty information regarding the future state of a system, specific to an event criterion. Ambiguity is therefore the uncertainty surrounding the forecast probability (NRC 2006; Eckel and Allen 2009), which can be described by a distribution of forecast probability values, referred to here as an *ambiguity distribution*.

Ambiguity may be found in ensemble forecasts that have limited sampling and insufficient simulation of sources of forecast uncertainty, i.e., the relevant, missing information. Ultimately, this leads to an inability of the ensemble to consistently reproduce or represent the true forecast PDF. The true forecast PDF is defined as the aggregate of all possible atmospheric states given a particular analysis using a specific NWP modeling system (Eckel and Allen 2009). Therefore, the true forecast PDF is specific to the EPS's underlying modeling system. Assume we have an infinite record of model analyses along with the resulting forecasts and atmospheric observations, where neither the model nor the atmospheric system has changed. To determine the true forecast PDF for a specific forecast lead time, we first search the record for all previous model forecasts matching the current model forecast. The analyses used to initialize each match will be the same, but their associated true states will be unique due to analysis error, thus the true state at each forecast lead time for each matching forecast will be unique. The true forecast PDF at the desired lead is then the combination of all verifying observations for the matching forecasts.

Limited sampling plays a large role in creating ambiguity. From Figure 1, an EPS with finite members cannot consistently represent the distribution from which the members are drawn, even if the EPS is otherwise ideal. The ensemble PDF may be calibrated to provide a reliable forecast on average, but the error for any specific case cannot be known before hand, which results in random error in the forecast PDF's moments (Figure 2). This random error exists even for large ensembles, thus sampling is a persistent source of random error and ambiguity in the ensemble forecast.

A non-ideal EPS misses simulation of some aspects of IC and model uncertainty, resulting in an ensemble forecast PDF with a variable and unknown ability to represent the true forecast PDF thus yielding ambiguity. Imagine an ideal EPS except that it is

designed with a single convective parameterization used for all members (i.e., uncertainty in modeling of convection is not represented). The ensemble PDF will be a close approximation to the true PDF if the error in the parameter value is low, or the sensitivity of the forecast to the parameter error is low, i.e., when convection is not present (Eckel and Allen 2009). The ensemble PDF may be a poor approximation when either parameter error or sensitivity is high. The variable representativeness of the forecast PDF for any one case cannot be predetermined, thus error in the forecast PDF appears random (Eckel and Allen 2009).

In this research, when considering ambiguity associated with ensemble forecast probabilities, systematic errors (bias) in the first two moments (mean and variance) of the PDF are removed through calibration leaving only the stochastic or random error. Although error may be present in higher moments, it is assumed that errors in the first two moments have the largest role in creating ambiguity. This may be explained by considering changes in probability density associated with changes to different moments of the forecast PDF. A change in the first moment (i.e., location) of the PDF results in a large shift of probability density and thus a relatively large change in forecast probability (depending on the placement of the event threshold within the PDF). While generally not as large as changes associated with the first moment, decreasing or increasing the variance (i.e., second moment) of the forecast PDF also has the potential to create a significant change in probability density, and therefore forecast probability. The ability to significantly adjust probability density relative to a given event threshold decreases with higher moments of the forecast PDF.

Camerer and Weber (1992) posit that uncertainty (first-order) and ambiguity (second-order uncertainty) are fundamentally different concepts. The magnitude of first-order uncertainty that can be measured as variance in the ensemble forecast PDF can be independent of the magnitude of ambiguity (i.e., misrepresentation of uncertainty in the ensemble forecast PDF due to random errors in the mean and/or variance). For instance, a well-designed EPS (i.e., well-sampled, well-perturbed) that is based on a poor NWP modeling system would produce forecasts with large uncertainty but very little ambiguity (Eckel and Allen 2009). Conversely, a poorly-designed EPS based on a highly skilled

NWP model would produce forecasts with low uncertainty and high ambiguity. However, Eckel and Allen (2009) assert that "larger uncertainty and/or more diversity in its sources may increase the opportunity for ensemble deficiencies, which can create ambiguity," thus correlating forecast uncertainty and ambiguity. Further evidence to support a relationship between ambiguity and the uncertainty in the current forecast is presented in the results section of this dissertation.

## C. FORECAST VALUE

The primary value of weather forecasts to users is the better consequences (economic or other benefits) realized from using the information in the decision making process (Zhu et al. 2002). Any new source of weather forecast information should add value to the decision maker. Value is added when the information allows the user to take actions that improve overall, long-term average consequences over many decision opportunities. In this research, we are concerned with the impact of introducing the ambiguity information into the user's decision making process. If users cannot effectively use the ambiguity information to add value, then the information holds merely entertainment value at best or confuses the user and detracts from optimal decision making at worst.

The analysis of value will be performed in the simple cost-loss ($C/L$) ratio scenario (Murphy 1985; Katz and Murphy 1997; Jolliffe and Stephenson 2003). In the basic $C/L$ scenario, the user will either decide to take protective action to mitigate the effects of some weather event or take no protective action based on the weather input. If the user decides to protect, he incurs a cost ($C$) for taking the protective action, regardless of whether or not the weather event occurs. If the user does not protect and the event does not occur, he incurs no expense. Otherwise, if the user does not protect and the event occurs, he will incur a loss ($L$). In this research, we assume the protective action is sufficient to guard against all loss. The results of the four possible preparation-outcome combinations over a forecast-observation dataset can be tallied in a $2 \times 2$ contingency table as shown in Table 1 (Jolliffe and Stephenson 2003). The expense ($E$) associated with each possible consequence is given in the Table 2.

For a deterministic forecast, the weather input to the decision making process is binary, whereas the stochastic forecast provides a probabilistic input. The stochastic input (i.e., the forecast probability, $p_e$) is converted to a binary input through application of a *decision threshold* or *decision rule* that expresses the amount of *risk* (i.e. the chance of getting an undesirable consequence) the user is willing to accept in the forecast of the weather event (Jolliffe and Stephenson 2003). In the *C/L* scenario, the goal is to use a decision rule that minimizes the total expense over many forecast cases, or the expected total expense.

The value score (*VS*), introduced by Richardson (2000), is a measure of the value of weather forecasts that can be explored through the *C/L* model. Using tallies (*a*, *b* and *c* defined in Table 1) accrued in the contingency table over *M* forecast-observation pairings, it is possible to calculate *VS* for any *C/L* ratio ($\alpha$):

$$VS = \frac{\frac{1}{M}\left(a\alpha + b\alpha + c\right) - \min\left(\alpha, \bar{o}\right)}{\bar{o}\alpha - \min(\alpha, \bar{o})} \tag{1}$$

where $\bar{o} = (a+c) / M$ is the sample's climatological rate of occurrence. In this form, the value of the forecast information is calculated assuming that in the absence of a forecast decisions will be made based on $\bar{o}$ (i.e., protecting when $\bar{o} \geq \alpha$). Additionally, decisions will be made using the ensemble forecast (i.e., protecting when $p_e \geq \alpha$). A perfect forecast has a *VS* = 1, while a $VS > 0$ indicates the forecast system adds value compared to following sample climatology. The forecast system has *VS* < 0 when it performs worse than climatology.

In the context of the *C/L* scenario where the goal is to minimize expected expense, optimal value is attained by a customer who chooses their decision rule or decision threshold to match their *C/L*. This fact can be demonstrated as follows. For many (*M*) instances in which the forecast probability ($p_e$) takes a specific value, a user

would either always protect or never protect based on their decision rule. For the two cases, the total expense (*E*) can be expressed respectively as:

$$E_{Protect} = M * C \tag{2}$$

$$E_{No\ Protect} = M * p_e * L \tag{3}$$

The user's decision should then be to protect when:

$$E_{Protect} \leq E_{No\ Protect}$$

$$M * C \leq M * p_e * L \tag{4}$$

$$p_e \geq \frac{C}{L}$$

Alternatively, the user's decision rule calls for taking no action when:

$$E_{No\ Protect} < E_{Protect}$$

$$p_e < \frac{C}{L} \tag{5}$$

As shown above, the user's optimal decision threshold is their *C/L*, prompting them to take protective action when $p_e$ is greater then *C/L* and to take no protective action when $p_e$ is less than *C/L* (Jolliffe and Stephenson 2003). Using this information, an analysis of the optimal *VS* obtainable by all customers can be determined. The curve in Figure 3 is the optimal *VS* created using data from the low-order model employed in this research (Chapter III.A). The *VS* for each *C/L* is calculated based on a unique contingency table (e.g., Table 2) for each user built using the *C/L* in question as the decision rule.

Ambiguity, or uncertainty in the forecast probability, adds another dimension to the decision making process, resulting in three possibilities given a user's *C/L*:

- The entire ambiguity distribution may be below the *C/L* (i.e., optimal decision threshold) so the user is convinced to take not protect.
- The entire ambiguity distribution is above the *C/L* so the user is convinced in their decision to protect.
- The ambiguity distribution *overlaps* the *C/L*. In this case, the appropriate decision is unclear to the user.

The term overlap is used here to refer to the total proportion of the ambiguity distribution that crosses the *C/L* in the direction opposing the decision based on the best-guess of the current risk, i.e., the chance of making the wrong decision. The ensemble forecast probability is taken as the *best-guess risk*, or alternately the *best-guess forecast probability*, since it represents the likelihood of the verifying observation crossing the event threshold resulting in a negative consequence. In Figure 4, the forecast probability indicates the user should protect. The ambiguity distribution overlap (hatched) in the figure describes the probability that the actual forecast probability is less than the *C/L* and the user should not protect.

In the *C/L* scenario, long-term expense may still be minimized by using the best estimate of risk even if ambiguity is present and ignored. The ensemble's estimate of risk (i.e., the forecast probability) is simply a random draw from a distribution of many possible forecast probability values (i.e., the ambiguity distribution). Given situations where overlap exists over many forecast cases, the best-guess risk will result in both positive and negative consequences, with the expectation that the forecast probability is truly a random draw from the ambiguity distribution and the selection process is not biased towards either of the consequence categories. Thus, the optimal user, when comparing the ensemble forecast probability alone as a measure of risk against the *C/L* (i.e., the optimal decision threshold), implicitly includes cases where overlap exists.

To this point, we have not addressed the question of value added via knowledge of ambiguity. This research introduces two approaches for attempting to add value to the

decision making process in situations where the decision input is unclear (i.e., overlap exists) using objective estimates of the ambiguity associated with the ensemble forecast.

### 1. Uncertainty-folding

The first approach to gain value from ambiguity information, called *uncertainty-folding*, combines the (first-order) uncertainty and ambiguity information to once again give the user a single probabilistic decision input based on the weather information. Given a sample of possible true forecast probability values ( $\hat{p}_T$ ) (i.e., ambiguity distribution or second-order uncertainty) estimated using some objective method, each $\hat{p}_T$ value is binned using a class interval of 1% over the range 0% to 100%. The relative frequency associated with each bin, $r(\delta)$, within the sample is determined. Note that $\delta = \{0.05, \ldots, 0.995\}$ (i.e., each bin's center value) is a possible true forecast probability value, and therefore represents a possible value of risk (i.e., first-order uncertainty). Each $\delta$ value is multiplied by its respective relative frequency then summed to produce a single estimation of the forecast probability ( $p_a$ ) that includes the ambiguity information.

$$p_a = \sum_{\delta} \delta r(\delta) \tag{6}$$

An example of this process is described in Figure 5.

As value studies in this research are focused on the *C/L* scenario, it is important to address whether or not the *C/L* is the optimal decision rule to minimize expense when using $p_a$. Samples from the ambiguity distribution (i.e., estimates of the true forecast probability, $\hat{p}_T$ ) are all equally plausible realizations of the forecast probability for an event given the EPS's sensitivity to the deficient simulation of uncertainty in the IC and model perturbations. As discussed, for any reliable ensemble forecast probability, the *C/L* is the optimal decision rule to minimize long-term expense, but while the forecast probability may be reliable on average, random error and ambiguity still exist for individual forecast cases. Thus, using the *C/L* with any random $\hat{p}_T$ value taken from the ambiguity distribution over many cases will minimize expected expense in much the

same way as using a single ensemble forecast. The $p_a$ value computed using uncertainty-folding is merely a combination of information from all of the $\hat{p}_T$ values and is therefore simply another plausible realization of the true forecast probability. In practice, this theory depends on obtaining an accurate objective estimate of the ambiguity distribution, which is discussed in the results section.

The control ensemble's calibrated forecast probability ( $p_e^*$ ) is a random sample taken from the ambiguity distribution. On the other hand, uncertainty-folding will produce a $p_a$ value close to the expected value of the ambiguity distribution. Thus the difference between $p_e^*$ and $p_a$ may be large enough to result in different decision inputs, i.e., they fall on opposite sides of the *C/L*. Over the long-term, $p_a$ should provide the best risk estimate and minimize expense by minimizing the error between the estimated risk used to make the decision and the true risk.

### 2. Secondary Decision Criteria

Dealing only with the economic value of information (i.e., the *C/L* scenario) neglects factors hard to quantify in terms of dollars that can also bring important consequences (e.g., loss of life, customer confidence, morale, mission effectiveness). Wallsten (1990) stated that ambiguity information was especially suited to decisions with multiple criteria. Thus, if the weather input to the decision is ambiguous, the user may be justified to take other factors into account to make the decision. This idea is used for the second approach to determine the value of ambiguity information. The simple *C/L* model is still applied, but when the ambiguity distribution overlaps the decision threshold (decision is unclear), the user *may* consider other (non-monetary) decision criteria to reverse the decision that would be made based purely on the best-guess risk. The option to include these secondary criteria comes with several questions:

- How much overlap of the ambiguity distribution across the optimal decision threshold (*C/L*) is necessary before the user should consider secondary decision criteria?
- How does the decision-maker decide whether or not to change their decision?
- How can we measure the improvement in secondary consequences?

Using this approach, the idea is not to increase the primary economic value (represented by the *VS*), but rather to add value to the user by improving consequences in terms of their secondary concerns. The goal is to add value to the secondary criteria without significantly decreasing the primary value achieved using the first-order criteria.

As an example of a secondary criteria, consider a user who cannot tolerate repeat false alarms (i.e., the event is forecast to occur but does not occur). An example may be a base commander, who previously evacuated aircraft and personnel when a typhoon was forecast to strike the base, but the typhoon track changed and it missed the base. The commander's decision, although justified by risk analysis, resulted in degradation of mission effectiveness and unnecessary expense. As the next typhoon approaches, the commander desperately wishes to avoid another unnecessary evacuation. If the commander is given a risk clearly exceeding his *C/L*, he should again evacuate. But, given ambiguity and overlap, he may choose to stay put.

Using an estimate of the ambiguity, it may be possible to reduce the likelihood of repeat false alarms by going against the decision based on the best-guess forecast probability, while not significantly changing the *VS* based on minimizing total expense. The idea is to reshuffle the outcomes to break up repeated false alarms, while keeping *VS* nearly constant. Changes to the decision based on including secondary decision criteria result in a different contingency table as compared to basing decisions only on the primary decision criteria (control $p_e$) (Table 3). In order to prevent changes in the primary value, the secondary criteria decision rule must produce changes that preserve the overall balance between positive and negative consequences, while not biasing towards one extreme. The user essentially trades the expense associated with a number of false alarms for the expense of a few extra misses as far as negative consequences are concerned. Value is then measured as a significant decrease in the number of repeat false alarms for a user who employs the ambiguity information compared to a user who bases

decisions solely on the best-guess forecast probability. Value is not gained if reducing the number of repeat false alarms results in a significant decrease in the primary value (i.e., increase in expenses).

It is important to stress that this scenario is just one example of using secondary criteria to add value when the decision input is unclear. There are many possible criteria that can be explored, where the criteria are user or context dependent.

There has been a great deal of effort put into designing EPSs to efficiently sample the uncertainty associated with an NWP modeling system, but the EPSs still have limitations that result in random error in the uncertainty estimates (i.e., ambiguity). The purpose of this research was not to explore EPS design, but rather to investigate methods for objectively estimating the ambiguity associated with an EPS and to understand how EPS deficiencies influence the magnitude of ambiguity. Additionally, while most research has focused on the decision maker's attitude towards ambiguity in the decision, we apply objective ambiguity estimates in the decision making process in an effort to add value compared to a user who simply uses the ensemble's uncertainty estimate. Thus, this research will attempt to show: (1) it is possible to produce reasonably accurate objective estimates of the ambiguity associated with an EPS and (2) the ambiguity information can add value to the decision making process.

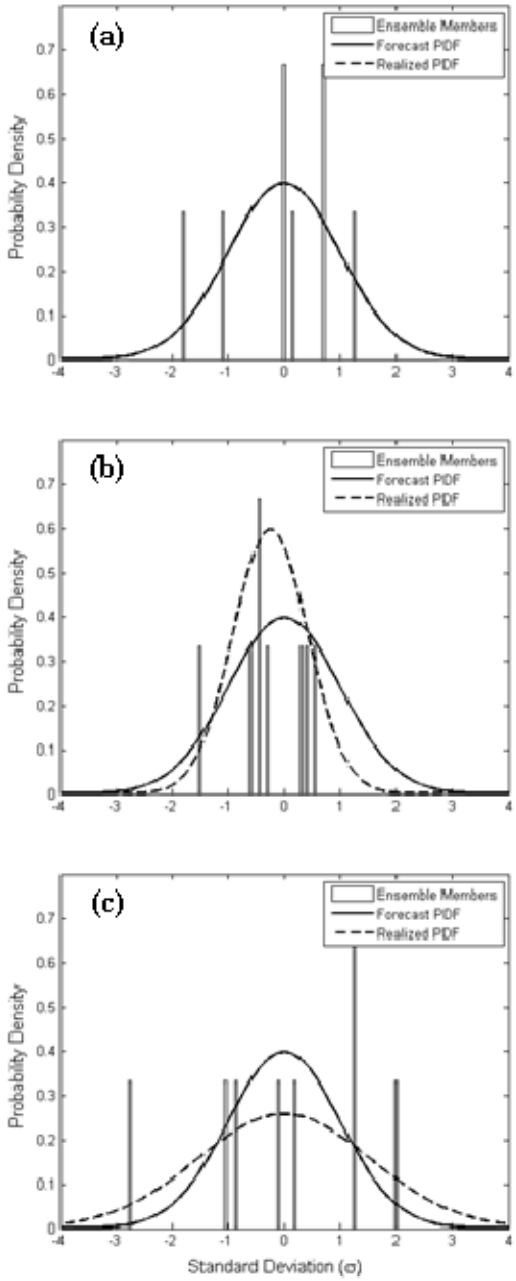



Figure 1. Three simulated attempts to represent the forecast PDF using an eight member "perfect model" ensemble. The forecast PDF (solid) being sampled is $N\left(0,1\right)$, while the realized ensemble PDF (dashed) is normal with parameters values calculated based on random ensemble members (a) mean and variance close to true values. (b) negatively biased mean and variance too small. (c) mean close to true and variance too large. Vertical lines represent the location of ensemble members.

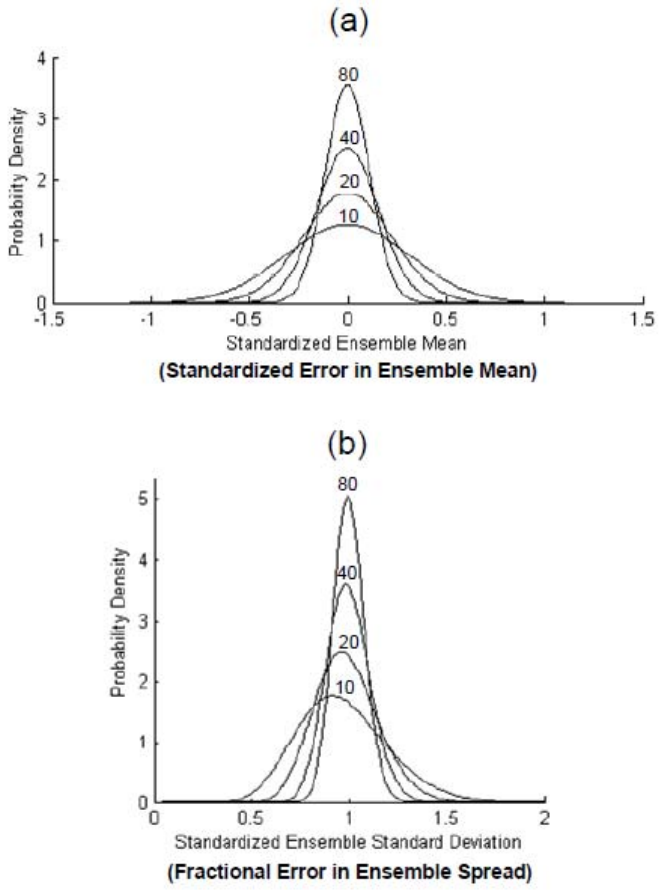



Figure 2. Sampling distributions of the (a) standardized error in ensemble mean and (b) fractional error in ensemble spread, dependent on the number of ensemble members. Results are shown for ensemble sizes of 10, 20, 40 and 80 members (labeled) [From Eckel and Allen 2009].

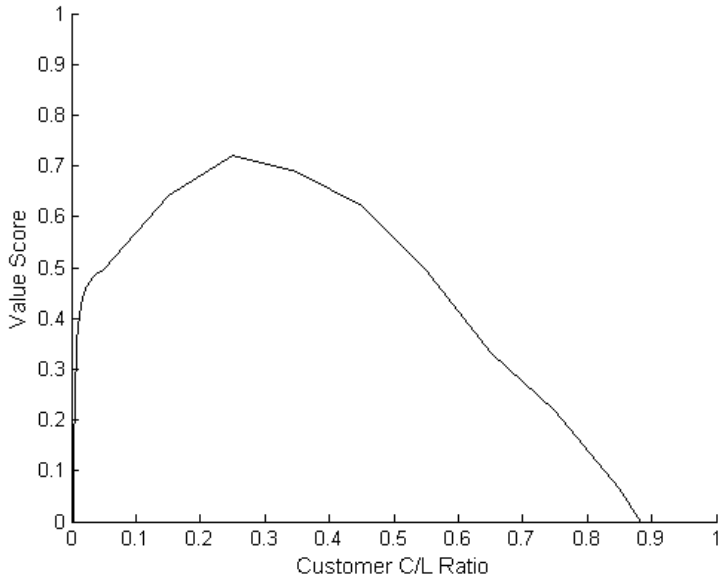



Figure 3. Optimal value score across the range of *C/L* values. The value score for each *C/L* is calculated using the *C/L* as the decision threshold. The climatological rate of occurrence ( $\bar{o}$ ) is 29.5%.

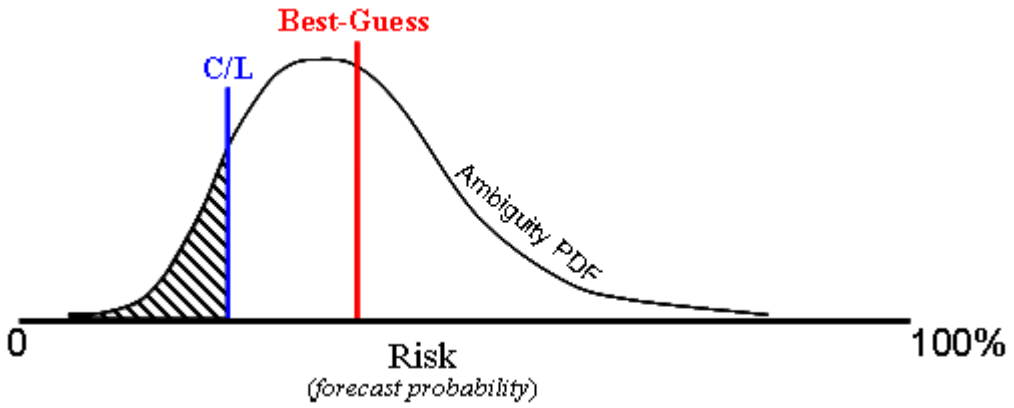



Figure 4. Ambiguity distribution overlap in the *C/L* scenario. The hatched area represents the overlap of the ambiguity distribution beyond the *C/L* (blue line), which would result in a different decision than that found using the best-guess or control forecast probability (red line).

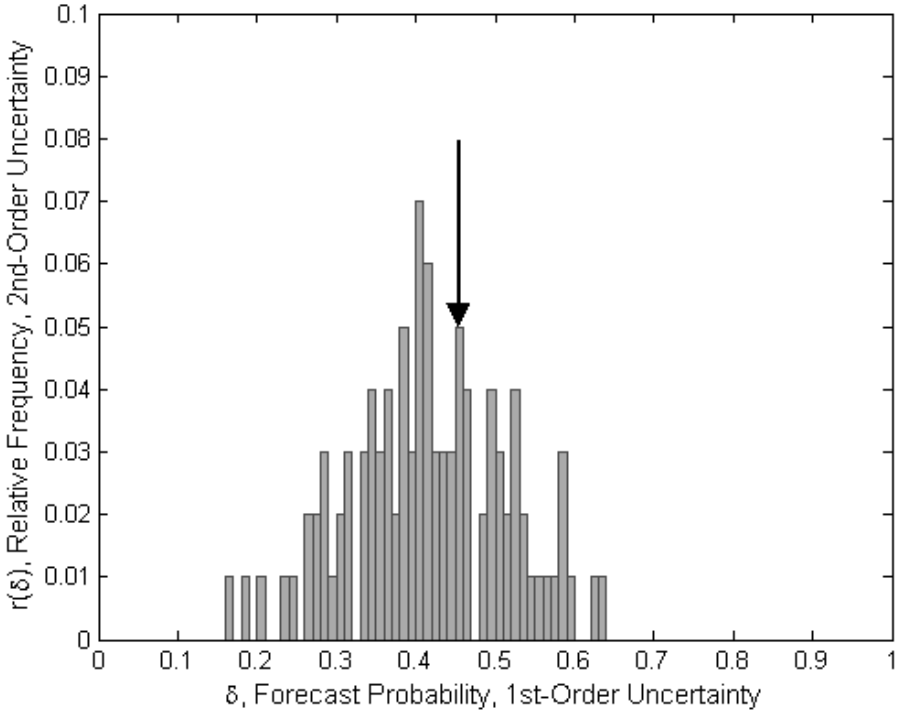



Figure 5. Histogram of possible first- and second-order uncertainty associated with some event used for calculating the uncertainty-folding forecast probability estimate ( $p_a$ ). As an example, the bin of forecast probability values $44\% < p_e \leq 45\%$ (arrow) has a relative frequency of 5%, thus contributing $44.5\% \times 5\% = 2.23\%$ to the summation in Equation (6).

Table 1. Contingency table used to tally the number of consequences associated with a forecast-observation dataset. A hit (*a*) is tallied when the weather event is forecasted to occur and the event does occur. When the event is forecasted to occur and is not observed, the resulting consequence is a false alarm (*b*). Alternately, when a weather event is not forecasted to occur is observed, the consequence is a miss (*c*). Lastly, a correct rejection (*d*) is counted when the weather event is not forecast to occur and the event is not observed.

| | | **Weather Event Observed** | |
|---|---|---|---|
| | | Yes | No |
| **Event Forecast and/or Decide to Prepare** | Yes | *a* (# of hits) | *b* (# of false alarms) |
| | No | *c* (# of misses) | *d* (# of correct rejections) |

Table 2. Contingency table of consequences measured as the expense (*E*) associated with each forecast-observation pair within the *C/L* framework. *C* is the cost of taking protective action to mitigate the loss (*L*) if the event occurs.

| | | **Weather Event Observed** | |
|---|---|---|---|
| | | Yes | No |
| **Event Forecast and/or Decide to Prepare** | Yes | $E = C$ | $E = C$ |
| | No | $E = L$ | $E = 0$ |

Table 3. Contingency table of possible changes in the repeat false alarm secondary decision criteria scenario. The change shown by the solid circle results in a positive consequence (correct rejection), while the change shown by the dotted circle results in a negative consequence (miss).

| | | **Weather Event Observed** | |
|---|---|---|---|
| | | Yes | No |
| **Event Forecast and/or Decide to Prepare** | Yes | *a* | *b* |
| | No | *c* | *d* |

# III. METHODOLOGY

This chapter describes the methods used during this research to accomplish the stated research goals. Specifically, Section A gives a detailed look at the design of the EPS and the low-order model it was based on. Section B provides an overview of data postprocessing. Section C provides a description of the ambiguity estimation techniques, while Section D covers the validation of the techniques. The final section discusses the processes and scenarios used for determining the value of the ambiguity information. The primary programming platform used during this research was Matlab version 7.0 or later.

## A. L96 ENSEMBLE PREDICTION SYSTEM

### 1. L96 Model Design

In order to fully study the ambiguity associated with EF, it is necessary to have access to an EPS, a large forecast dataset, and suitable observation information. As a portion of this research will involve running multiple parallel EPS forecasts, using an EPS of an atmospheric model is impractical due to computational and storage limitations. Therefore, we use an EPS of a more simple dynamical system model to mimic an operational EPS. For this research, we chose the low-order, chaotic model first introduced by Lorenz (1996) as a suitable proxy for atmospheric NWP models.

The model, hereafter L96, includes a set of symmetric, coupled equations describing the evolution of variables on two distinct time scales (Lorenz 1996; Wilks 2005).

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b} \sum_{j=J(k-1)+1}^{kJ} Y_j \; ; \qquad k = 1 \ldots K \tag{7}$$

$$\frac{dY_j}{dt} = -cbY_{j+1}(Y_{j+2} - Y_{j-1}) - cY_j + \frac{hc}{b} X_{\mathrm{int}|(j-1)/J|+1} \; ; \qquad j = 1 \ldots JK \tag{8}$$

The model emulates atmospheric processes in that the linear and forcing (*F*) terms provide internal dissipation and external forcing, and the quadratic terms simulate advection (Lorenz 1996). Results garnered from experiments using the L96 model can therefore reasonably be assumed to apply to atmospheric modeling systems. To further ensure the validity of this research, we designed the L96 EPS to operate using state-of-the-art methods for data assimilation, ensemble perturbations, and numerical techniques.

The $X_k$ variables in the L96 model can be thought of as describing large-scale, slow moving processes, and the $Y_j$ variables thought of as small-scale, fast moving processes, where energy is transferred between the two scales of motion (Lorenz 1996; Wilks 2005). A possible physical explanation of the modeled process would be to consider the $Y_j$ variables as representing convective-scale values while the $X_k$ variables represent large-scale static instability (Lorenz 1996). Described another way, the $X_k$ variables are resolved on the model grid (latitude circle), while the $Y_j$ variables are unresolved or subgrid scale variables (Wilks 2005).

The basic setup of the L96 model for this research follows Wilks (2005), with some modifications. After Wilks, $K = 8$ and $J = 32$, which corresponds to eight resolved variables and 256 unresolved variables (Figure 6). Scaling constants *h*, *c*, and *b* are taken as 1, 10, and 10, respectively, which ensures both scales are chaotic (Lorenz 1996; Wilks 2005). To mimic operational atmospheric models, the unresolved variables are not modeled explicitly and must be parameterized in some fashion since they influence the evolution of the large-scale, resolved variables. We assume the physical laws governing the resolved variables are known completely, but the effects due to the unresolved variables are not precisely known and must be parameterized (Wilks 2005). In this configuration, the last term on the right side of Equation (7) is replaced by a parameterization term ( $g_U$ , described below) (Wilks 2005; Orrell 2003),

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - g_U(X_k); \qquad k = 1 \ldots K \tag{9}$$

This experiment design gives us a model with random error (i.e., uncertainty) as well as the ability to be omniscient and know the true evolution of the system beginning from a known initial condition. The full, coupled L96 equations provide the "ground truth" or "true" state of the system. To provide the true trajectory, Equations (7) and (8), henceforth termed the L96 *System* (L96S), are integrated forward using the fourth-order Runge-Kutta (RK4) (Weisstein 2009) numerical scheme at a time step of 0.001 (non-dimensional). In comparison, the parameterized L96 equations, Equation (9), henceforth the L96 *Model* (L96M), are integrated forward using the second-order Runge-Kutta (RK2) numerical scheme at a time step of 0.005. Forecasts derived using L96M exhibit model errors as a result of using an inferior numerical integration scheme (RK2 vs. RK4) and from parameterization of the unresolved scales.

The parameterization scheme in L96M is stochastic and based on the unresolved tendencies found between integrations of the L96S at time steps equivalent to the L96M (Wilks 2005). Developing the parameterization involves first integrating the L96S forward over some long trajectory with time step of 0.001, while storing all data. Then, for each time step over this long trajectory, the current resolved variable's value ( $X_k(t)$ ) and the value at a time equivalent to one L96M time step ( $X_k(t+\Delta t)$, $\Delta t = 0.005$ ) are found in the L96S data. The unresolved tendencies $U(t)$ are then calculated as:

$$U(t) = \underbrace{\left[ -X_{k-1}(t)\left\{ X_{k-2}(t) - X_{k+1}(t) \right\} - X_k(t) + F \right]}_{\mathbf{A}} - \underbrace{\left[ \frac{X_k(t+\Delta t) - X_k(t)}{\Delta t} \right]}_{\mathbf{B}} \qquad (10)$$

Term A represents the model tendency of $X_k$ over $\Delta t$, while term B gives the true tendency of $X_k$ over $\Delta t$.

The range of subsequent values associated with each possible $X_k$ value represents the unresolved tendency in $X_k$, or values that could be missed without explicitly

modeling the unresolved $Y_j$ variables. The symmetry of the governing equations produces very similar *U* for each resolved variable, so we can combine all results. The unresolved tendencies in $X_k$ are shown in Figure 7 for all eight resolved variables. The data are fit with a fourth-order polynomial regression (solid line in Figure 7). The unresolved tendency depends "strongly and nonlinearly on the value of the resolved variable" (Wilks 2005). Thus, for each value that the resolved variable can take on, there is a distribution of unresolved tendency values, centered on the regression curve. The parameterization function $\left( g_U \right)$ is thus given both a deterministic and a stochastic component:

$$g_U(X_k) = b_0 + b_1 X_k + b_2 X_k^2 + b_3 X_k^3 + b_4 X_k^4 + q_k \tag{11}$$

where $b_0 = 0.293$, $b_1 = 1.55$, $b_2 = -0.0201$, $b_3 = -0.0106$, and $b_4 = 0.000565$ are the regression coefficients. The deterministic component (the first five terms on the right hand side of Equation (11) is the regression equation. The $q_k$ term on the right side represents the stochastic component that allows for parameter values off of the regression curve.

The simple stochastic term is white noise produced by a normal distribution with zero mean and standard deviation of 2.32, which is equal to the average standard deviation of unresolved tendencies across all possible $X_k$ values. We rescale the stochastic component following Hansen and Penland (2006) who found that combining stochastic components with deterministic differential equations requires scaling by the square root of the time step $\left( 1 / \sqrt{0.005} = 14.14 \right)$. Initial tests of the ensemble resulted in ensemble forecasts that were nearly perfectly dispersive. Since the EPS needed to mimic current operational ensembles that are typically under-dispersive (Wilks 2005; Buizza 1997; Toth and Kalnay 1993; Hamill and Colucci 1997; Eckel 2008), we reduced the standard deviation of the white noise distribution to 1.2 to reach a suitable, albeit subjective, level of under-dispersion (Chapter III.A.4). By decreasing the range of white

noise in the L96M, additional model error was created, which resulted in under-dispersion since it was not accounted for in the EPS.

## 2. L96 Climatology

This section provides an understanding of the climatology of the L96S and L96M and notes any differences. For comparison and to show the advantages of using the stochastic parameterization, we introduce a simple deterministic parameterization where the stochastic component ($q_k$) in Equation (11) is set to zero. Also discussed are the correlations between the resolved ($X_k$) variables to motivate the decision to consider each as independent when determining the ensemble's error characteristics and for use in the validation of the ambiguity estimation techniques.

To determine the climatological statistics, we ran the L96S and L96M using both the deterministic and simple stochastic parameterization schemes over a period encompassing 5000 time units beginning from a random, transient-free initial state. For each of the model climatologies (deterministic and stochastic), all eight of the $X_k$ variables were stored. For the L96S climatology, both the $X_k$ and $Y_j$ variables were maintained to understand the climatology of unresolved variables as well. A summary dataset was created for each configuration by assuming independence of the eight resolved variables (discussed below) and combining them. Climatological statistics for each dataset are displayed in Table 4. Comparing the range of values the resolved variable trajectory visited, it is clear the stochastic parameterization provided a climatology closer to L96S. The probability density of possible $X_k$ values is shown in Figure 8. Both model configurations do a reasonable job of representing the distribution of resolved variables, but the range and shape of the stochastic distribution is superior. The mean and standard deviation of the resolved variables are significantly closer to the "true" system values for the stochastic parameterization as well. These results further strengthened the case for implementing the L96M as described.

We may consider the location of the resolved variables in the L96M as grid points on a single latitude circle (Lorenz 1996; Wilks 2005), where the forecast values at a

specific lead time are essentially values at *K* adjacent grid points. Thus the resolved variables are analogous to variables in an operational NWP model (e.g., 2m temperature or 10m wind speed). Verification of operational deterministic and ensemble forecasts for a variable such as 2m temperature over a certain domain can be accomplished by using a high quality model analysis. An "observation" is available at each model grid point, so verification takes place at each grid point. These individual point statistics are then combined for an overall assessment of the modeling or ensemble system. In order to combine data from the individual grid points, the data should be independent and uncorrelated. In many cases, these assumptions do not strictly hold, but the correlations may be weak. In practice, the grid point data is typically combined under the assumption of independence to increase the size of the forecast-observation dataset used for verification.

Evaluating the error characteristics and skill associated with an EPS requires an extensive set of forecasts and observations. Following Descamps and Talagrand (2007), who found that "cross correlation between the $X_k s$ is negligible" in the L96 system, we chose to assume independence between the $X_k$ variables. Although our testing indicated a pattern of moderate correlation between the variables, we proceeded under the assumption of independence to increase the size of the verification and climatological datasets. By making this choice, we may have underestimated the uncertainty associated with our results due to the increase in the size of the datasets.

## 3. L96M EPS Design

This section describes how L96M was incorporated into an EPS for this research. As described earlier, the goal of an EPS is to effectively account for all sources of uncertainty in the modeling system (Chapter II.A). Thus, state-of-the-art techniques were used to account for analysis errors and model deficiencies.

To generate a control analysis and a suite of ensemble ICs, the process begins with a uniform random draw for each of the eight $X_k$ variables and each of the 256 $Y_j$ variables from their respective climatological ranges. Using the L96S, the random state

is integrated forward to converge upon an arbitrary state on the true system attractor. This state is taken as the current true state of the system. The L96S is then integrated forward from this state over the data assimilation period and the entire forecast period to provide a true trajectory for the system. The ICs for the ensemble members are found using an Ensemble Kalman Filter (EnKF) data assimilation scheme.

Data assimilation is the process by which imperfect information (i.e., observations and model background) about the current state of a system are combined optimally to produce an analysis of the current state that is more precise than the original information (Kalnay 2003; Reichle et al. 2002). The Kalman Filter (KF) (Kalman 1960; Cohn 1997) is "an approximation of Bayesian state estimation which assumes linearity of error growth and normality of error distributions" (Hamill 2006). The KF process is generally divided into two steps, an update step and a forecast step.

The update step involves adjusting an estimated state (e.g., background) and associated error statistics to new observations to form a new analysis state and uncertainty estimate. In the forecast step, the new analysis and uncertainty estimate are propagated forward to the next observation time using the full nonlinear dynamical model and tangent linear model and adjoint, respectively. Ultimately, the traditional KF is computationally too expensive for practical use in atmospheric data assimilation due to the high dimensionality of atmospheric modeling systems.

The EnKF process is a sequential data assimilation technique that uses an ensemble of perturbed forecasts to provide the statistical information needed to produce the new analysis (Evensen 1994, 1997; Burgers 1998). The process is an approximation of the traditional KF or extended KF where the background-error covariance is not explicitly propagated forward in time but is estimated using the variance of the ensemble of background states (Evensen 1997; Reichle et al. 2002). In addition to not needing the tangent linear model and adjoint for explicit prognosis of the forecast-error covariance, the EnKF does not require the assumptions of linear error growth and normality of error distributions (Hamill 2006; Kalnay 2003; Tippett et al. 2003; Reichle et al. 2002). Determination of the background-error covariance using the ensemble provides a flow-

dependent estimate of the background error allowing the EnKF to more optimally update the background to new observations (Whitaker & Hamill 2002; Hamill 2006).

Ensemble-based data assimilation can be put into two categories, deterministic and stochastic. The basic difference depends on "whether or not random noise is applied during the update step to simulate observation uncertainty" (Hamill 2006). The EnKF used for this research is a stochastic data assimilation technique, in that it involves an ensemble of parallel data assimilation cycles where each member of the ensemble is updated using an observation set perturbed by white noise while still being a plausible realization of the observed state of the system (Hamill 2006). The process used to generate the set of perturbed observations is described below.

Burgers et al. (1998) showed that for EnKF analysis to work properly, the observations must be considered random variables. Otherwise, the ensemble error covariances (background and analysis) will be underestimated since using the same observation to update each member results in spurious correlations. Underestimation of the error covariances may lead to filter divergence (i.e., the analysis drifts away from truth) as observations are underweighted in the update step (Burgers et al. 1998; Whitaker & Hamill 2002). Using optimal DA, the analysis should typically be more precise than the information used to create it. Several methods have been developed to account for error covariance underestimation, such as covariance inflation and localization (e.g., Anderson and Anderson 1999; Anderson 2003; Houtekamer and Mitchell 1998).

Importantly, the underestimation problem is a function of the ensemble size used during EnKF (Hamill 2006; Whitaker and Hamill 2002; Reichle et al. 2002). Burgers et al. (1998) noted that using too small an ensemble resulted in large analysis errors, and more benefit could be gained by using an optimal interpolation data assimilation scheme. According to Kalnay (2003), research using a quasi-geostrophic model found that 25-50 members were enough to benefit from using EnKF, but Houtekamer and Mitchell (1998) found ensembles on the order of 100 members were necessary. Due to the inexpensive computational cost of implementing the EnKF with the L96M, an ensemble size of 500 members was used for this research. We tested the EnKF over many scenarios starting from different locations in the model attractor to ensure filter divergence was not

occurring. The Euclidean difference between the best-guess analysis and observation vectors and the true state vector averaged 0.3 and 1.05, respectively. Thus covariance underestimation was not a problem, likely due to the large ensemble size chosen.

The EnKF data assimilation process is presented here following notation used by Hamill (2006) and is shown in Equation (12) (a)-(e). Let $\mathrm{X}^{\mathrm{b}} = \left( \mathrm{x}_1^{\mathrm{b}}, \ldots, \mathrm{x}_{\mathrm{m}}^{\mathrm{b}} \right)$ describe an ensemble of background state vectors $\left( \mathrm{x}_{\mathrm{i}}^{\mathrm{b}} \right)$ with *m* members in which each member's data is a column vector covering all state variable values. Ensemble perturbations $\left( \mathrm{x}_{\mathrm{i}}^{\prime \mathrm{b}} \right)$ from the mean $\left( \frac{1}{n} \sum_i \mathrm{x}_{\mathrm{i}}^{\mathrm{b}} \right)$ are found in the matrix given by Equation (12) (a).

$$\mathrm{X}'^{b} = \left( \mathrm{x}_1'^{\mathrm{b}}, \ldots, \mathrm{x}_{\mathrm{m}}'^{\mathrm{b}} \right), \qquad \mathrm{x}_{\mathrm{i}}'^{\mathrm{b}} = \mathrm{x}_{\mathrm{i}}^{\mathrm{b}} - \overline{\mathrm{x}}^{\mathrm{b}}, \qquad i = 1, \ldots, m \qquad (a)$$

$$\hat{\mathrm{K}} = \hat{\mathrm{P}}^{\mathrm{b}} \mathbf{H}^{\mathrm{T}} \left( \mathbf{H} \hat{\mathrm{P}}^{\mathrm{b}} \mathbf{H}^{\mathrm{T}} + \mathbf{R} \right)^{-1} \qquad (b)$$

$$\hat{\mathrm{P}}^{\mathrm{b}} = \frac{1}{m-1} \mathrm{X}'^{\mathrm{b}} \mathrm{X}'^{\mathrm{b}^{\mathrm{T}}} \qquad (c) \qquad (12)$$

$$\mathrm{x}_{\mathrm{i}}^{\mathrm{a}} = \mathrm{x}_{\mathrm{i}}^{\mathrm{b}} + \hat{\mathrm{K}} \left( \mathrm{y}_{\mathrm{i}} - \boldsymbol{H}(\mathrm{x}_{\mathrm{i}}^{\mathrm{b}}) \right) \qquad (d)$$

$$\mathrm{x}_{\mathrm{i}}^{\mathrm{b}} = \boldsymbol{M} \left( \mathrm{x}_{\mathrm{i}}^{\mathrm{a}} \right) \qquad (e)$$

The update process begins by calculating an estimated Kalman gain [ $\hat{\mathrm{K}}$, Equation (12) (b)], which gives the optimal weights for the update based on the observation- and background-error covariances (Reichle et al. 2002). To calculate the Kalman gain, the background-error covariance ($\hat{\mathrm{P}}^{\mathrm{b}}$) from Equation (12)(c) must be estimated diagnostically from the ensemble of background states using Equation (12)(a). The over-hat (^) is used to denote that the covariance found is an estimate of the true error covariance since the ensemble size is finite. The **H**-term is the linear transformation

matrix used to interpolate model data to the observation locations and transform the model state variables to match the observed variable. In this research, the observation and model data are the same quantity and are collocated, thus **H** is the identity matrix. **R** is the observation-error covariance matrix describing the typical observation error. The superscript $^{T}$ denotes the transpose of a vector or matrix.

With the estimated Kalman gain, Equation (12) (d) is used to update each member of the ensemble of background state vectors $\left( x_i^b \right)$ individually using random (stochastic) realizations of the observation information $\left( y_i \right)$ to find the ensemble of analysis state vectors $\left( x_i^a \right)$. The nonlinear transformation matrix, ***H***, performs the same function as **H**, and was again simply the identity matrix. Following the update step, each member of the ensemble of analysis states is integrated forward in time using the full nonlinear model (***M***) to the time of the next observation using Equation (12) (e). The process is repeated when new observation data is available.

For this research, the initial EnKF state estimate (eight $X_k$ variable values) used to initialize the spin-up cycle (described below) was taken as an observation of the current true state of the system. We created an observation by adding a random draw from an $N\left(0, \sqrt{\mathbf{R}}\right)$ distribution to the current true state taken from the L96S trajectory. The standard deviation of the observation error was taken as 5% of the climatological standard deviation in $X_k$, thus $\sqrt{\mathbf{R}}$ was a diagonal matrix with 0.2 at all locations on the diagonal. We generated 500 additional perturbed observations by adding random draws from $N\left(0, \sqrt{\mathbf{R}}\right)$ to the original observation. These 500 perturbed observations were used as the initial EnKF members. The same process was used to produce perturbed observations of the current true state for all subsequent filter updates.

We ran the EnKF through a one-time spin-up cycle consisting of 1000 model time steps (0.005 time units each) where perturbed observations were available to update the filter every 10 steps, or one data assimilation cycle. Each data assimilation cycle is approximately equal to receiving observations every six hours, according to Lorenz

(1996) who found that one time unit was approximately equal to five days. The spin-up cycle is necessary to allow the EnKF to achieve dynamical stability, ensuring its mean is close to the true state and its perturbations have evolved to more accurately estimate the background error covariance. The final forecast states from each of the 500 EnKF members following the spin-up cycle were updated to produce the EnKF analyses used for the first ensemble forecast. Additionally, these 500 EnKF analyses were used as the starting point for the next data assimilation cycle.

We chose to separate the ensemble forecast runs by 20 data assimilation cycles (i.e., 200 model time steps). The length of this separation period was chosen empirically to allow sufficient time for the trajectory to reach a different region of the model attractor, thus reducing correlation between ensemble forecasts and producing forecasts that span as much of the attractor as possible. We monitored the total vector difference between the starting mean analysis and the mean analysis found following each data assimilation cycle over a number of cycles for many different starting conditions. The vector difference generally increased over a period of 15-25 data assimilation cycles before starting to decrease, which led to our choosing 20 cycles as the separation period.

Following each data assimilation cycle, we took the mean of the EnKF members as the best-guess analysis of the current state of the system. Burgers et al. (1998) explained that the EnKF mean is a state estimate minimizing the root mean square error of the forecast. The best-guess analysis provided the initial condition for the deterministic forecast. The 21 ensemble forecast members' initial conditions were taken as uniform random draws (without repeats) from the 500 EnKF members, all of which are equally plausible (Hansen 2009). We chose a 21-member ensemble to coincide with NCEP's Global Ensemble Forecast System (GEFS), which we used during our value studies (Chapter III.F).

Model deficiencies are simulated in the L96M EPS using the perturbed parameter approach, which is applied through the simple stochastic parameterization. As described previously, a perturbed parameter EPS uses a single NWP model (i.e., the L96M) where parameter values within the model are perturbed for each ensemble member (Chapter II.A.2.a). The stochastic parameterization randomly varies the parameter value for each

member at every time step while maintaining the deterministic component of the parameterization as the average parameter value.

We also tested a multi-model configuration of the L96M EPS. The parameter coefficients were designed to be static for each ensemble member, analogous to the deterministic portion of the stochastic parameterization (red line in Figure 7). First, we binned the unresolved tendency values (blue dots in Figure 7) across all $X_k$ values using a class interval of $0.5$. To determine a single member's coefficients, a uniform random draw for each bin was taken from a range equaling four times the standard deviation of unresolved tendency values within the bin centered on the bin's average unresolved tendency value. We found each ensemble members parameter coefficients using a fourth-order polynomial fit to the values drawn for each bin. The result was *n* static deterministic parameterizations that are similar in nature but perturbed within the known uncertainty of the parameter (Figure 9). Testing of the multi-model EPS consistency and skill (not shown) showed mostly negligible differences between the perturbed parameter EPS and the multi-model EPS configurations. One large difference was found when comparing each EPS's limit of predictability. The perturbed parameter EPS showed skill through ten time units, while the multi-model EPS maintained skill through only eight time units. We chose the perturbed parameter approach for this research since it was previously proven to work well in the context of the L96 system (Wilks 2005).

## 4. L96M EPS Performance

Uncalibrated and calibrated forecast data is used to evaluate the consistency and skill of the L96M EPS to ensure it behaves similarly to a real-world EPS. Consistency (i.e., statistical consistency) is a measure of how well on average the ensemble forecast PDF matches the true forecast PDF (Anderson 1996, 1997; Talagrand et al. 1997). We evaluate consistency using the error variance diagram, dispersion diagram, verification rank histograms (VRH), and the verification outlier percentage (*VOP*). The error variance diagram is used to understand the predictability and benefit of using the ensemble forecast by displaying the average error growth and comparing the limits of predictability of the deterministic and ensemble forecasts (Eckel 2008). The dispersion

diagram directly compares the mean square error of the ensemble with the average ensemble variance at each forecast lead time, where the ratio of these two values should equal one for statistical consistency (Eckel 2008). This diagram is also useful in diagnosing ensemble dispersion (i.e. rate of change in ensemble spread with increasing time) and ensemble spread problems (under or over) (Eckel 2008). The VRH aides in visualizing dispersion and consistency characteristics by tracking the location of the verifying observation amongst the ranked ensemble members over many trials. Ideally, the frequency of occurrence in each rank is equal. Hamill (2001) described interpretation of various VRH shapes, but also demonstrated how EPS problems may be masked in the VRH by interactions with other issues. We employ *VOP* as a measure of the ensemble's ability to portray truth by finding the percentage of verifications that fall outside three standard deviations from the ensemble mean (Eckel 2008). *VOP* is calculated as:

$$VOP = \frac{1}{M}\sum_{m=1}^{M}\begin{Bmatrix} 0:\ 3\left(\sigma_e\right)_m \geq \left|V_m - \overline{e}_m\right| \\ 1:\ 3\left(\sigma_e\right)_m < \left|V_m - \overline{e}_m\right| \end{Bmatrix} \tag{13}$$

$M$ is the total number of verifications, $\overline{e}_m$ and $\left(\sigma_e\right)_m$ are the mean and standard deviation of the ensemble members for a single verification, respectively, and $V_m$ is a single verification value. Lower *VOP* values indicate an ensemble PDF that more consistently portrays the true state of the system.

The error variance diagram created from L96M forecast data (Figure 10) shows that L96M accurately models the L96S climatology. Over a long forecast trajectory, the deterministic forecast's error variance should asymptote to twice the climatological variance (Eckel 2008), which is seen in the figure. The deterministic limit of predictability due to error growth is found at $\tau \approx 3.8$ ($\tau$ equals forecast time), where the deterministic error variance increases above the climatological variance ($\sigma_C^2$). Once the ensemble mean error variance reaches $\sigma_C^2$ ($\tau \approx 10.2$), the ensemble forecast has lost predictability, and a forecast based on climatology is in order. The extension of the

ensemble mean error variance above $\sigma_C^2$ was not expected, but it is consistent with results seen in Tribbia and Baumhefner (2004) using a real-world EPS to forecast 500-mb height. Maximum ensemble dispersion is indicated between $\tau \approx 0.6$ and $\tau \approx 2.0$ by the average variance between ensemble members, which is a measure of how the ensemble members diverge with respect to one another. In a consistent EPS, this measure should match the forecast error growth (i.e. rate of increase of deterministic forecast error variance) (Eckel 2008). Thus, the L96M EPS is under-dispersive, but it was designed this way on purpose to imitate the performance of a real-world EPS.

Dispersion diagrams are provided for both the uncalibrated (Figure 11) and calibrated (Figure 12) ensemble forecast data (see Chapter III.B.1 for specifics on the calibration technique). As stated, the dispersion diagram gives a direct look at the consistency of the EPS by comparing of mean square error of the ensemble mean and the average ensemble variance (Chapter III.B). As expected, the dispersion diagram for the uncalibrated data indicates under-dispersion of the ensemble forecast on average. The bulk calibration is able to correct for the dispersion deficiencies and give near-perfect dispersion at all forecast lead times. VRH for various forecast lead times are provided for the uncalibrated (Figure 13) and calibrated data (Figure 14) as well. In Figure 13, the L96M EPS displays the characteristic U-shaped VRH of being under-spread, where more verifications fall into the outer ranks than expected. The indication of a slight positive bias (i.e., more verifications in the left-hand ranks) is also present. This positive bias is seen in the L96M error statistics (Chapter III.B.3). Calibration is able to flatten out the VRH (i.e. make the ranks more uniform) throughout the forecast period (Figure 14). The remaining lack of uniformity seen in the calibrated VRH may be explained by the lack of calibration on higher moments of the ensemble PDF.

The *VOP* values (Figure 15) indicate the calibrated ensemble forecast PDF does a better job representing the true forecast PDF compared to the uncalibrated data. Both datasets show low *VOP* values early in the forecast period, which rapidly increase as error growth increases. Since the calibrated data has a better handle on the dispersion on average, its *VOP* value does not grow to the extent of the uncalibrated data. Although the

calibrated data shows near perfect dispersion on average, the *VOP* does not reach the perfect line in Figure 15 since dispersion is not perfect for all individual cases.

We now evaluate forecast skill using the entire forecast dataset to examine the performance of forecast probabilities (see Chapter III.B.2 for specifics on calculating forecast probability). In this research, two representative event thresholds were chosen to be verified, one to represent a fairly common event and the other a rare event, based on the climatology of the L96S (Figure 8(a), page 84). The threshold for the common event was taken as $X = 6.31$, which is exceeded 30% of the time. The rare event threshold was $X = 9.98$, which is exceeded only 10% of the time.

We verified probability forecasts using the Brier Skill Score (*BSS*) using sample climatology as the reference forecast (Jolliffe and Stephenson 2003). *BSS* decomposition provides a measure of the *reliability* and *resolution* of the ensemble forecasts for a given event threshold. Taken over many verifications, reliability is a measure of how well forecast probabilities match observed relative frequencies for the event in question (Wilks 2006). For example, over many cases where the probability of occurrence is 20%, we expect to observe (verify) that event 20% of the time. The resolution of the ensemble forecasts provides a measure of how well the forecasts distinguish between events and non-events (i.e. the sharpness of the forecast PDF) (Wilks 2006).

The *BSS* we employed is the decomposed form, which uses discrete, contiguous bins of forecast-observation data pairs allowing calculation of the component reliability and resolution values (Eckel 2008). To calculate the *BSS*, we must first define the Brier Score (*BS*) (Wilks 2006):

$$BS = \frac{1}{M}\sum_{i=1}^{I} N_i \left( \left( p'_e \right)_i - \overline{o}_i \right)^2 - \frac{1}{M}\sum_{i=1}^{I} N_i \left( \overline{o}_i - \overline{o} \right)^2 + \overline{o}\left(1 - \overline{o}\right) \qquad (14)$$

$M$ is the total number of forecast-observation pairs, $I$ is the total number of bins, and $N_i$ is the number of forecast-observations pairs in the $i^{\text{th}}$ bin. Also, $\left( p'_e \right)_i$ is the

representative forecast probability value for the $i^{th}$ bin (i.e., bin's average $p_e$ value), $\overline{o}_i$ is the observed relative frequency of the $i^{th}$ bin, and $\overline{o}$ is the sample climatology. The first term on the right hand side of the equation is the reliability (*rel*) component of the *BS*, while the second term is the resolution (*res*) component. The final term is a measure of the *uncertainty* (*unc*) in the forecast of the event in question and is solely dependent on the event climatology. *BSS* may then be computed by (Wilks 2006):

$$BSS = \frac{res - rel}{unc} \tag{15}$$

For the common event, the *BSS* indicates forecast skill through $\tau = 9.6$ for the uncalibrated data (Figure 16) and $\tau = 10.2$ for the calibrated data (Figure 17). Calibration appears to have significantly improved the reliability of the ensemble forecasts for this event throughout the forecast [Figure 18(a)], while a small and likely insignificant improvement in resolution was also seen [not apparent in Figure 18(b)]. The combination of improvements provided a small gain in *BSS* scores throughout the forecast, thereby extending the period over which the L96M EPS showed skill. For the rare event, the *BSS* indicates forecast skill through $\tau = 7$ for both the uncalibrated (Figure 19) and calibrated (Figure 20) data. Calibration resulted in an improvement in reliability through $\tau = 2.6$, but degraded reliability after that time [Figure 21(a)]. It should be noted that scaling of the figure makes the decrease in reliability appear large, but changes are in the thousandths decimal place. Although hard to discern in Figure 21(b), resolution is actually improved by the calibration, which likely offset the decrease in skill due to worse reliability.

Based on this analysis of the EPS performance, we have further confirmed the ability of the L96M to simulate the L96S climatology and demonstrated the effectiveness of the calibration technique used during this research. More importantly, we have shown that the L96M EPS appears to behave like a real-world EPS. Additionally, we have seen that the uncalibrated and calibrated forecasts for both the rare and common events have skill out to approximately seven and ten time units, respectively. This feature plays a

crucial role when we consider the value of the ensemble forecasts and the ambiguity information, as it would not make sense to assess the value of a modeling system compared to climatology once the modeling system has lost skill compared to climatology. Taking this analysis in conjunction with data processing constraints, we only consider forecast lead times less than five time units for exploring the research objectives.

## B. POSTPROCESSING OF ENSEMBLE FORECAST DATA

This section describes the postprocessing of the ensemble forecast data. We used the L96M EPS to generate 3,000 ensemble forecasts each consisting of 8 resolved variables, giving a total of 24,000 verifications available at each forecast lead time. The forecasts were run out to five time units (non-dimensional), and postprocessing was performed at a time increment of 0.2 time units, which totaled 51 lead times including the analysis. For the purpose of determining the L96M EPS error characteristics and calibration coefficients, the postprocessing described in the following subsections was performed using all forecast data at each forecast lead time. Verifying observations were taken directly from the L96S trajectory without adding error (based on the typical observation error) even though erred observations are a source of random error. By $\tau = 0.2$, the typical observation error was only a small fraction of the total error, thus the observation error was inconsequential at later lead times.

### 1. Calibration

We calibrated the L96M ensemble forecast data to correct for systematic errors. Once the systematic error is removed, the remaining error is the random error associated with the forecast, which is the primary cause of ambiguity. In this research, a simple bulk calibration was performed to correct the average errors associated with the first and second moments of the forecast PDF. We chose to use a bulk calibration technique versus a more sophisticated technique to allow for a fair comparison during the estimate validation. A calibration technique that introduces additional information (e.g., downscaling) may reduce ambiguity, thus applying a more sophisticated calibration to

one of the practical estimation techniques (discussed in Chapter III.C) without performing the same calibration on the theoretical estimation method used as the validation standard would bias the results. Calibration was performed at each forecast lead time.

We used a simple *shift-and-stretch* calibration technique described by Eckel (2008). The shift adjusts the first moment of the ensemble forecast PDF by correcting each ensemble member individually by the negative of the mean error in the ensemble mean defined as:

$$ME_{\bar{e}} = \frac{1}{M}\sum_{m=1}^{M}\left(\bar{e}_m - y_m\right) \tag{16}$$

$M$ is the total number of verification points, $\bar{e}_m$ is the mean of a single ensemble forecast, and $y_m$ is the observation. Using $ME_{\bar{e}}$, the shift calibration is performed as follows:

$$\tilde{e}_i = e_i - ME_{\bar{e}}\ , \qquad i = 1 \ldots n \tag{17}$$

$\tilde{e}_i$ is a single shifted ensemble member, $e_i$ is a single uncorrected member, and $n$ is the number of ensemble members. This approach assumes the bias is the same for each ensemble member, making it unacceptable for use with a multi-model EPS.

The second moment calibration or stretch is performed to increase (or decrease) the spread (defined here as standard deviation) of the ensemble forecast PDF in accordance with the fractional error in ensemble spread:

$$\sigma' = \sqrt{\frac{\overline{\sigma_{\tilde{e}}^2}}{MSE_{\bar{\tilde{e}}}}} \tag{18}$$

The numerator is the average ensemble variance, and the denominator is the mean square error of the bias-corrected (shifted) ensemble mean, each respectively defined as:

$$\overline{\sigma_{\tilde{e}}^2} = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{n-1}\sum_{j=1}^{n}(\tilde{e}_{i,j} - \overline{\tilde{e}}_i)^2\right] \tag{19}$$

$$MSE_{\overline{\tilde{e}}} = \left(\frac{n}{n+1}\right)\frac{1}{M}\sum_{i=1}^{M}(\overline{\tilde{e}}_i - y_i)^2 \tag{20}$$

$n$ is the number of ensemble members, $M$ is the total number of individual verifications, $\overline{\tilde{e}}_i$ is the bias-corrected ensemble mean, $y_i$ is the observation, and $\tilde{e}_{i,j}$ is the $j^{th}$ bias-corrected ensemble member (Eckel and Mass 2005; Eckel 2008). Eckel and Mass showed the $n/(n+1)$ factor in Equation (20) is required for small ensemble sizes. The stretch calibration is performed using the previously shifted ensemble members:

$$e_i^* = \overline{\tilde{e}} + \left(\tilde{e}_i - \overline{\tilde{e}}\right)\frac{1}{\sigma'}, \qquad i = 1 \ldots n \tag{21}$$

$e_i^*$ is the $i^{th}$ fully calibrated ensemble member.

## 2. Calculating Forecast Probability

Unless otherwise noted, we based all forecast probability calculations during this research on probability of exceedance of the event threshold. The results presented would not change if the probability of precedence were used.

We calculated forecast probability values using the uniform ranks method (Hamill and Colucci 1997). Uniform ranks assumes the output from each of the *n* ensemble members for a variable at one grid point is equally likely, or that there is a uniform

probability distribution of the rank-ordered values. The total probability is then divided into $n+1$ bins, each with equal probability of containing the verifying observation.

The forecast probability is calculated as the sum of the rank-probability bins greater than the event threshold, plus the partial probability of the bin containing the event threshold. For an event threshold in bins 2 through $n-1$, the forecast probability ( $p_e$ ) is calculated as:

$$p_e = \Pr\left(\Theta < V < e_i\right) \; + \; \Pr\left(V \geq e_i\right)$$
$$p_e = \left(\frac{e_i - \Theta}{e_i - e_{i-1}}\right)\frac{1}{n+1} \; + \; \frac{n-i+1}{n+1} \tag{22}$$

$\Theta$ is the event threshold value, $V$ is the observation value, $e_i$ is the value of the ensemble member with rank $i$, and $n$ is the number of ensemble members (Eckel 2008). This process is depicted in Figure 22 .

If the event threshold falls in either rank 1 or rank $n+1$, it is not possible to use Equation (22) since no ensemble value is available to calculate the partial probability. For example, if $\Theta$ lies in rank $n+1$, then $e_{i-1} = e_n$ is the largest ensemble value and no $e_i = e_{n+1}$ ensemble member is available. In this case, the forecast probability is calculated as:

$$p_e = \; \Pr\left(V \geq \Theta\right) \; = \; \left(\frac{1-G(\Theta)}{1-G(x_n)}\right)\frac{1}{n+1} \tag{23}$$

$G\left(\;\right)$ represents the Gumbel cumulative density function (CDF) (Wilks 2006) given by equations:

$$F(x) = \exp\left\{ - \exp\left[ \frac{(x-\xi)}{\beta} \right] \right\},$$

$$\beta = \frac{s\sqrt{6}}{\pi}, \tag{24}$$

$$\xi = \overline{x} - \gamma\beta$$

The Gumbel CDF parameters are estimated using the sample mean ($\overline{x}$) and standard deviation ($s$) of the ensemble members and $\gamma = 0.57721$ (Euler's Constant). If $\Theta$ falls in rank 1, the reverse Gumbel, $G'(\ )$, is used (Eckel 2008).

$$\begin{aligned} p_e &= \Pr\left(\Theta < V < e_1\right) \ + \ \Pr\left(V \geq e_1\right) \\ p_e &= \left( \frac{G'(e_1) - G'(\Theta)}{G'(e_1)} \right) \frac{1}{n+1} \ + \ \frac{n}{n+1} \end{aligned} \tag{25}$$

The Gumbel distribution was chosen to represent the tails of the ensemble distribution because of its ability to capture extreme events.

### 3. L96M EPS Error Characteristics

Estimating ambiguity in the ensemble forecast requires knowledge of the error characteristics of the EPS. For the 3,000 forecast cycles, each EPS forecast run consisted of 21 members describing plausible realizations of the 8 resolved variables at each time step. From previous discussion, the eight resolved variable forecasts are considered independent forecasts and are combined to create a total dataset of 24,000 ($3{,}000 \times 8$) ensemble forecasts. The postprocessing procedure used for the ensemble forecast data is depicted in Figure 23.

Using the large calibrated dataset, Equations (16), (18), (19), and (20) are applied to diagnose the overall or bulk EPS error characteristics, giving the mean error of the ensemble mean, fractional error in ensemble spread, and average ensemble variance at each time step. As stated, uncertainty in the forecast probability is actually a function of the error variance and not the bulk error. We therefore remove the bulk (i.e. systematic) error using the shift-and-stretch calibration method to reveal the remaining random error which contributes to ambiguity. The practical ambiguity estimation techniques (described in Chapter III.C) use the error variance statistics (i.e., the random error) to produce their ambiguity distributions.

To determine the variance associated with the relevant error statistics, it is necessary to subset the large ensemble forecast dataset (Eckel and Allen 2009). For this research, we chose to subset the large dataset of 24,000 verifications (per time step) into 3,000 sets of 8 forecasts, where each set is an individual EPS run. Each ensemble forecast, consisting of 21 possible values of the eight variables, describes the uncertainty about a unique trajectory within the model attractor. Errors associated with each ensemble forecast PDF are sensitive to the flow or current location in the attractor. Thus sub-setting based on complete EPS runs ensures flow-dependent error characteristics from different locations around the model attractor are used to find the error variance statistics. This sub-setting strategy also follows the analogy of relating the L96M EPS to an operational EPS running once per day. If we assume each L96M EPS run is the same as an operational EPS run then we are essentially looking at a 3,000 individual (one per day) ensemble forecasts. The subset error statistics are thus equivalent to determining the error on a daily basis, which is the same as that chosen by Eckel and Allen (2009). Once the error statistics ($ME_{\bar{e}}$, $\sigma'$ and $\overline{\sigma_e^2}$) are calculated for each of the 3,000 subsets using the same equations as above, the variance of the subset values provides the variance of the error distributions about the bulk values. The L96M EPS error statistics (bulk and variance) for each time step are provided in Table 5. The error statistics indicate that the L96M EPS had a small positive bias that was consistent throughout the forecast. The fractional error was less than one throughout the forecast, but this was expected since the EPS was designed to be under-dispersive to mimic an operational EPS.

## C. ESTIMATING AMBIGUITY

This section provides a description of three ambiguity estimation techniques used during this research. The first estimation method is a fundamental approach that would produce the true distribution of forecast probabilities given unlimited sampling. The remaining two techniques estimate the ambiguity based on the error characteristics of the EPS and are a practical approach to estimating ambiguity. The explanation of the practical methods follows that given in Eckel and Allen (2009) for ease of writing, where real-world data from the Japanese Meteorological Agency EPS (hereafter JM) was used over the same domain and forecast period described in Chapter III.F.

### 1. Ensemble-of-Ensemble

The ensemble-of-ensemble (EoE) method is the theoretical and impractical approach to estimating the ambiguity associated with an operational EPS. The calibrated forecast probability ( $p_e^*$ ) from a non-ideal EPS can be considered a single sample from a distribution of true forecast probabilities ( $p_T$ ), since the ensemble forecast PDF is a single realization of many plausible forecast PDFs, given the limited sampling and unaccounted for uncertainty in the EPS. The EoE approach builds an estimate of the $p_T$ PDF by running *N* parallel EPSs (termed *constituents*) each with unique ICs and each with unique model perturbations, all of which are similar in nature to the original, control EPS. The result is *N* equally probable forecast PDF realizations for any single forecast timeframe, giving *N* unique $\hat{p}_T$ samples (i.e., estimates of the true forecast probability). The distribution of $\hat{p}_T$ reflects the uncertainty in the forecast probability (i.e. the ambiguity) in the forecast. This approach is unrealistic and absurd for operational use given the large computational expense of running multiple, parallel ensemble forecasts, and if the computational resources were available, they would be better served improving the EPS through additional members and/or higher resolution.

To produce the EoE ambiguity distribution, we ran multiple parallel ensemble forecasts (i.e., different versions of the L96M EPS) from the same control analysis state

while allowing initial condition perturbations and model parameterization values to vary. Each constituent gave a different yet equally plausible set of *n* members as well as a different forecast probability of occurrence of some event threshold. Taken all together, the constituents provided a distribution of forecast probability values (i.e., an estimation of the $p_T$ PDF) for a given event that was flow dependent or sensitive to the uncertainty in the EPS perturbations.

An important consideration for the EoE was the number of constituents required to provide a thorough statistical sampling of the ambiguity distribution. This question is analogous to the problem of determining an appropriate number of members to use for an ensemble forecast. Too few constituents may lead to misrepresentation of the desired ambiguity distribution even when the distribution the constituents are sampling is perfect (Figure 1). For ensemble forecasts, error in forecast probability decays exponentially with increasing ensemble size with the most dramatic decrease for sizes ranging from 2-20 members, whereas the decrease in error for sizes $>20$ members dropped off significantly. This suggests that an ensemble size $\geq 20$ is needed to reasonably represent the underlying true forecast PDF. Using this reasoning, it was assumed that an EoE with $\geq 20$ constituents would adequately represent the ambiguity distribution. As computational costs were not a significant limiting factor during the EPS runs, the EoE was configured to produce $N=100$ constituent ensemble forecasts in order to minimize sampling error.

The setup of the L96M EPS used for the EoE forecast runs is shown in Figure 24. In contrast to the setup used for determining the EPS error characteristics, the initial state fed into the data assimilation for each of the 100 constituent runs is identical (outside the dashed blue box in Figure 24). In this way, the DA process creates a unique set of perturbed initial conditions for each constituent based on the same forecast situation. The differences in the perturbed states are due to random processes within the DA process varying the outcome within the realm of possible analysis error. Additionally, the model parameterization values vary randomly (by design in the L96M) throughout the

constituent runs. The combination of varying initial condition and model parameter perturbations results in varied but equally likely realizations of the uncertainty in the future state of the system.

Creation of an EoE forecast case dataset begins similarly to the runs used for finding the EPS error characteristics, described in Chapter III.A.3. From an initial set of 500 perturbed observations, we completed a spin-up cycle of 1000 model time steps updating the filter with new perturbed observations every 10 steps. Following the spin-up cycle, the process then runs an additional 20 data assimilation cycles (10 model steps each) from the final spin-up cycle analyses. The first EoE constituent forecast is run using the analyses found at the completion of the last data assimilation cycle, where *n* ensemble members are selected as before. For the next EoE constituent, another 20 data assimilation cycles are run once again starting from the final set of spin-up cycle analyses. Thus the next constituent forecast is run over the same forecast period as the previous constituent. This process is repeated to produce a single dataset of *N* EoE constituents for a specific forecast scenario. Subsequent EoE forecast case datasets are separated from initial state of the previous forecast case by the standard separation period (i.e., 200 model steps) to find cases from different regions in the L96M attractor.

### 2. Calibrated Error Sampling

The calibrated error sampling (CES) method uses information on past performance of the ensemble to estimate ambiguity. Errors in ensemble forecast probability ( $p_e$ ) may be the result of errors in any moment of the ensemble forecast PDF. For CES, we focus on errors in the first two moments, as they are believed to be the largest contributors. Possible errors in the ensemble PDF may be described by error distributions for the ensemble mean and spread based on long-term verification. Such error distributions reflect error due to finite sampling as well as error due to unaccounted for uncertainty in the EPS (Eckel and Allen 2009). For error in the first two moments, we find the error distributions for the mean error of the ensemble mean [ $ME_{\bar{e}}$ , equation (16)] and fractional error in ensemble spread [ $\sigma'$ , equation (18)]. The mean values for

the $ME_{\bar{e}}$ and $\sigma'$ error distributions are computed over a full verification dataset, while the spread of the distributions are calculated using subsets of the full dataset, as described in Chapter III.B.3. Note that the following explanation of CES follows that given in Eckel and Allen (2009) using real-world JM 2-m temperature 5-day forecast data.

We may estimate the distribution of possible $p_e$ errors by converting potential PDF errors into $p_e$ errors. Consider the following example of translating from a PDF error to a $p_e$ error using an arbitrary ensemble 2-m temperature forecast, defined as a Gaussian with a mean of 2.8°C and a spread of 1.8°C, or $N\left(2.8^{\circ}\mathrm{C}, 1.8^{\circ}\mathrm{C}\right)$ (Figure 25). For this example, we assume that the true forecast PDF is known (which is generally untrue), and it is $N\left(2.2^{\circ}\mathrm{C}, 2.6^{\circ}\mathrm{C}\right)$. The errors in the ensemble PDF mean and spread due to finite sampling and/or ensemble deficiencies are 0.6°C and –0.8°C, respectively. The fractional error in spread is thus 1.8 / 2.6 = 0.69. The $p_e$ error can be calculated for any chosen event threshold by comparing the ensemble forecast probability and the true forecast probability ( $p_T$ ). For the event of temperature ≤ 0°C, the error is –13.9% since $p_e$ = 6.0% and $p_T$ = 19.9% (Figure 25).

Performing the same type of calculation over many different event thresholds (i.e., different values of 2-m temperature) for the same ensemble and true distributions, produces different $p_e$ error values for each threshold chosen [Figure 26 (a)]. Similarly, we may employ a single event threshold while allowing the location of the ensemble PDF to vary (i.e., an ensemble PDF with the same mean and spread errors placed in different locations with respect to the event threshold). In our example, the positive bias of the ensemble PDF results in primarily negative $p_e$ error values (since we are considering probability of preceding the event threshold). Positive $p_e$ errors are present for high event thresholds once the true PDF's density becomes larger to the right due to the under-spread ensemble PDF. When the event threshold moves deeper into the PDF tails on either side, $p_e$ error asymptotes to zero as the outcome of the event for both the ensemble and true forecast PDFs become more certain (i.e., $p_e$ closer to 0% or 100%). Our goal is

to provide an estimate of ambiguity as a function of ensemble forecast probability, so we replot the results in Figure 26(b) as true probability versus ensemble forecast probability.

The ensemble PDF in Figure 25 exhibits merely one of many possible errors in ensemble mean and spread. Different PDF error values can produce different $p_e$ errors. For a given EPS, the spectrum of ensemble PDF errors is described by error distributions (one each for ensemble mean and spread) created by evaluating the EPS's long-term error characteristics as described above. To remove systematic error leaving only the random error component, we bulk calibrate the forecasts as described in Chapter III.B.1. Example error distributions for $ME_{\bar{e}}$ and $\sigma'$ are shown in Figure 27 (a) and (b), where the $ME_{\bar{e}}$ distribution is fit using a normal distribution, and the $\sigma'$ distribution is fit to a gamma PDF.

CES also requires a distribution for the average ensemble spread, shown in the example Figure 27(c). Error in the ensemble mean affects $p_e$ error, but the actual value of the ensemble mean does not impact $p_e$ error. On the other hand, $p_e$ error is affected by both error in the ensemble spread and the magnitude of the ensemble spread itself. Wider ensemble PDFs produce smaller values of $p_e$ error since differences in the ensemble and true probability densities become smaller. The distribution of average ensemble spread is computed following the same methodology used to find the error distributions. We then fit the ensemble spread distribution with a gamma distribution.

Scatter plots between these various parameters in Figure 28 show no strong correlations between the three variables (average ensemble spread, error in ensemble mean, and error in ensemble spread). The spread-skill relationship suggests that the variability of errors in the forecast PDF increase with increasing ensemble spread, which would result in larger ambiguity. Thus we must determine if a significant correlation exists between ensemble spread and the variability (i.e., standard deviation) of errors in the ensemble mean and spread. Figure 29 shows a plot of mean error and spread and indicates the variability of both errors remains fairly constant regardless of ensemble spread (thus independent). Additionally, the standard deviations of both errors generally

match the standard deviation of the error distributions in Figure 27 (a) and (b). The spread-skill relationship is likely not seen due to computing the domain averaged errors for each forecast case. Since the variables are independent, we can sample randomly from each variables' distribution to give a set of possible values, which may then be used in CES.

To summarize the CES method, for a given set of random samples from distributions as in Figure 27, the $p_e$ error value for a specific calibrated forecast probability value is found as follows. The randomly drawn ensemble mean error and ensemble spread values are assumed to describe a Gaussian distribution, where the mean error value is an error in location away from zero. Using this distribution, the event threshold value giving the forecast probability in question is located. This event threshold is then used to find the true forecast probability, where the true PDF is a Gaussian distribution with zero location error and spread equal to the ensemble spread divided by the randomly drawn fractional error. In this way, the spread of the true forecast PDF will be greater than the ensemble PDF if the fractional error is less than one. The $p_e$ error is then the calibrated forecast probability in question minus the true forecast probability. The true forecast probability is actually a single estimate from the distribution of estimated true probabilities ( $\hat{p}_T$ ), since it was found using a single set of error distribution and spread values representing plausible variations in the ensemble forecast PDF based on past performance. It is important to note that the CES ambiguity estimate is not based on knowing the true forecast PDF, but rather on knowing possible values of its mean and spread relative to the ensemble PDF described by the EPS error characteristics, as well as possible values of ensemble spread.

In Figure 30, we see that possible $\hat{p}_T$ values vary dramatically for five sets of random draws from the Figure 27 PDFs. For $p_e$ = 55%, the $\hat{p}_T$ values range from 47% to 80%, yielding a rough estimate of ambiguity (i.e., the actual $p_T$ may randomly occur within that range). Looking at the same $p_e$ value (55%), a robust ambiguity estimate can be created by repeating the CES process using 50,000 random samples from the error and

ensemble spread distributions, thus producing a specific ambiguity distribution [Figure 31(a)]. CES produces an ambiguity distribution for all values of calibrated forecast probability. We define *total ambiguity* as the 90% CI (i.e., the maximum likely value minus the minimum likely value) of the distribution of $\hat{p}_T$ values for a specific calibrated forecast probability value,

$$total\ ambiguity = p_{95} - p_5 \tag{26}$$

where $p_5$ and $p_{95}$ represent the 5th and 95th percentile of the rank-ordered $\hat{p}_T$ values, respectively.

Computing the total ambiguity for each calibrated forecast probability value yields the results in Figure 32, conveying the general, overall ambiguity of our example EPS temperature forecasts. In general, these results do not make sense when considering an ensemble forecast at a specific point, as the results were produced for all possible values of ensemble spread. A specific forecast has a specific ensemble spread. Figure 33 shows that CES runs for fixed values of ensemble spread, but the same variability of mean and spread error described by the distributions in Figure 27 (a) and (b), have very different amounts of ambiguity. Thus, in real-world applications, specific CES ambiguity distributions must be generated for the full range of observed ensemble spread values.

Therefore, CES takes two forms in this research, thus making a distinction between developing the CES ambiguity distributions using randomly varying ensemble spread values or using specific ensemble spread values. The former, termed CES Global ($CES_G$), produces a bulk estimate of the ambiguity distributions for any calibrated forecast probability, independent of ensemble spread. The later method, termed CES Local ($CES_L$), provides a flow-dependent estimate of the ambiguity distributions specific to ensemble spread values.

CES requires a significant up-front computational expense to produce the ambiguity lookup tables for each calibrated forecast probability value. Real-time application involves simply calculating the ensemble spread then accessing the lookup

table to get the ambiguity data. The crux lies in developing error distributions for the ensemble mean and spread, as these distributions are likely sensitive to changes in forecast lead time, season, location, weather patterns, etc. Subsetting of the verification datasets must be accomplished in such as way as to avoid combining dissimilar signals while maintaining a large enough sample size to get accurate results.

### 3. Randomly Calibrated Resampling

The second practical method for estimating ambiguity, randomly calibrated resampling (RCR), employs bootstrap resampling, which is designed to estimate the uncertainty in sample statistics (Wilks 2006). In application here, the sample dataset is the set of ensemble members and the sample statistic is the forecast probability for a given event. A single resampling of the *n*-member ensemble values consists of making *n* random draws with replacement resulting in a new version of the dataset and a different $\hat{p}_T$ for the event. Repeating this process 10,000 times gives a distribution of $\hat{p}_T$ values. It is important to note that the original $p_e$ from the control ensemble forecast will be near the mean of the resampled $\hat{p}_T$ distribution since averaging the alternative datasets reproduces the original (Eckel and Allen 2009). Note that the following explanation RCR follows that given in Eckel and Allen (2009) using real-world JM 2-m temperature 5-day forecast data.

Resampling alone will not provide an accurate estimate of the ambiguity associated with a given ensemble forecast, since the resampling process accounts for only one source of ambiguity, finite sampling. The resampled ambiguity distribution is dependent on the size of the ensemble used to represent the true forecast PDF. Resampled ensemble forecasts from a small ensemble are likely to produce very different PDFs and subsequently very different $\hat{p}_T$ values [Figure 34(a)], resulting in a wider ambiguity distribution. The resampled datasets from a well-sampled, large ensemble are more likely to give similar PDFs, reducing the range of $\hat{p}_T$ values [Figure 34(b)]. Resampling does not account for random error due to deficient simulation of sources of

uncertainty. Possible forecast solutions missing among the members due to deficiencies of an imperfect ensemble would never show up in any resampling, thus the $\hat{p}_T$ PDF will be too narrow (Eckel and Allen 2009).

Since forecast probability values are confined between 0% and 100%, systematic bias can also affect the width of the $\hat{p}_T$ PDF. For example, an ensemble with a negative bias will shift the $\hat{p}_T$ PDF towards 0% erroneously, causing a decrease in variance as $\hat{p}_T$ values are unable to cross the lower bound. To provide an accurate estimate of ambiguity, the effects of random and systematic error must be included.

Each resampled dataset can be calibrated using information from the EPS's error characteristics by applying the 'shift-and-stretch' technique described previously (Chapter III.B.1.). As before, the bulk mean error in the ensemble mean [$ME_{\bar{e}}$, Equation(16)] is used to correct the first moment of the ensemble PDF. Corrections to ensemble spread are made using the average fractional error in ensemble spread [$\sigma'$, Equation(18)] to stretch (or compress) the bias-corrected members about their mean. Each resample dataset is calibrated individually, giving calibrated forecast probability values for each resample.

Bootstrap and calibration account for systematic errors and random errors due to finite sampling, but not random errors due to unrepresented sources of uncertainty. To include these effects, the calibration applied to each resample dataset is varied by using random draws from the EPS's $ME_{\bar{e}}$ and $\sigma'$ error distributions. Random calibration takes into account the variation in the ensemble forecast error statistics, which result from the EPS's inability to simulate all of the uncertainty associated with the forecasts. Thus calibrating based on the random errors brings in possible forecast solutions that would otherwise be absent in the resampled datasets. As the distributions from which the random deviations are drawn are centered at the average $ME_{\bar{e}}$ and $\sigma'$, the original calibrated forecast probability value is maintained as the central value of the $\hat{p}_T$ PDF (Eckel and Allen 2009).

Before the random deviations are drawn from the distributions of $ME_{\bar{e}}$ and $\sigma'$, we must remove the error variance due to finite sampling (Eckel and Allen 2009). Otherwise, finite sampling will be accounted for twice (i.e., once by bootstrap resampling and once by random calibration) leading to overestimation of ambiguity. The sampling distributions in Figure 2 reflect the contributions purely from finite sampling to error in the ensemble mean and spread associated with calibrated PDFs for various ensemble sizes. Since we are concerned with adjusting the raw error distributions, the spread of the distributions in Figure 2 must be de-standardized.

When calibrating ensemble spread towards an average fractional error of one, the spread of the fractional error distribution is adjusted by nearly the same proportion as the average fractional error. Thus to de-standardize the spread of the fractional error distribution for an *n*-member ensemble from Figure 2, we reduce the spread value by a factor equal to the raw average fractional error (from EPS forecast verification) divided by the average fractional error for an *n*-member ensemble (from Figure 2). The spread of the EPS's fractional error distribution is then reduced by the de-standardized spread due to finite sampling to give the reduced error distribution for random calibration.

For the contribution of finite sampling for an *n*-member ensemble to variance in $ME_{\bar{e}}$, Figure 2(a) gives a standardized (i.e., calibrated) value based on $1/\sqrt{n}$, which must be inflated by $RMSE_{\bar{\varepsilon}}$ to de-standardize to the $ME_{\bar{e}}$ distribution. The $RMSE_{\bar{\varepsilon}}$ represents the best estimate of the standard deviation of the true forecast PDF ($\sigma_T$), since both $\sigma_T$ and $RMSE_{\bar{\varepsilon}}$ represent the average error in observations away from the true mean ($\mu_T$) or the bias-corrected ensemble mean (Eckel and Allen 2009). Therefore, $RMSE_{\bar{\varepsilon}}/\sqrt{n}$ is subtracted from the standard deviation of the EPS's $ME_{\bar{e}}$ error distribution to arrive at the PDF for random calibration. Examples of the reduced error distributions of $ME_{\bar{e}}$ and $\sigma'$ are shown in Figure 35.

Each resample is thus randomly calibrated using information on the long-term variability of the ensemble's error, which generates a wider $\hat{p}_T$ PDF (Figure 36). The width of the RCR ambiguity distribution is strongly dependent upon the spread of the

original ensemble forecast, giving the RCR estimate flow-dependent characteristics. An ensemble forecast with less uncertainty (low spread) will typically have a wider $\hat{p}_T$ PDF compared to a forecast with greater uncertainty. For a low spread forecast, the adjustment in location for each resample forecast PDF due to the random calibration results in a larger range of $\hat{p}_T$ values for a given event threshold (discussed in detail in the Results chapter).

Although RCR appealingly produces a more flow-dependent ambiguity estimate, it comes with a significant, real-time computational cost. Generating and analyzing the resampled datasets (10,000 at each grid point for every variable of interest) may be too computationally demanding for operational application.

## D. VALIDATION

There is a fundamental difference between the EoE approach and the other two ambiguity estimation techniques. In EoE, the original, calibrated, control forecast probability value ( $p_e^*$ ) represents a single, random draw from the theoretical $p_T$ PDF, which is estimated by the EoE $\hat{p}_T$ PDF. Thus, the original $p_e^*$ can fall anywhere within the $p_T$ distribution. The other two estimation techniques use information on past ensemble performance to provide a $\hat{p}_T$ distribution (i.e., $p_T$ PDF estimate) that is centered on the original $p_e^*$. Because of this difference, our validation efforts were confined to determining how well the practical estimation techniques captured the variance of the EoE ambiguity distribution.

The EoE produces a spectrum of possible forecast PDFs and a $\hat{p}_T$ PDF for any particular event at some particular lead time, and it dynamically captures the EPS limitations (i.e., limited sampling and inadequate simulation of uncertainty). EoE reflects the flow-dependent deficiencies in the perturbations associated with the different regions in the attractor. $\mathrm{CES}_G$, on the other hand, produces a $\hat{p}_T$ PDF for any particular $p_e^*$, which could come from any event. The $\hat{p}_T$ distribution is a generic ambiguity distribution based solely on the EPS's average error characteristics, which are taken as

the same over the entire attractor. $CES_L$ produces a somewhat flow-dependent ambiguity distribution based on the same general error distributions but dependent on specific ensemble spread. RCR again uses the same general error distributions, and its ambiguity estimate is somewhat flow-dependent since the estimate is sensitive to the distribution of members in the ensemble PDF.

The validation strategy considers aggregated ambiguity distributions built over many locations on the L96M attractor in order to determine the overall effectiveness of the ambiguity estimation methods. This strategy was necessary because the sample PDFs used to find the $\hat{p}_T$ values for the $CES_G$ ambiguity distribution were created using the long-term, average error distributions. Thus the $CES_G$ ambiguity distribution reflected the forecast uncertainty from a combination of many possible events or locations on the L96M attractor. We created the aggregates by combining data from all of the EoE forecast cases used for validation into a single dataset. Accordingly, the same aggregation had to be done for the $CES_G$ and RCR datasets. $CES_L$ was developed based on the evolution and validation results in this research (Chapter IV), which unfortunately resulted in its omission from the validation study due to time and processing constraints.

For this validation strategy, we must consider what each ambiguity estimation method regards as the expected value of its ambiguity distribution, $E(\hat{p}_T)$. The $E(\hat{p}_T)$ for $CES_G$ and RCR is the calibrated forecast probability value ($p_e^*$), which is the best-guess forecast probability value from the control ensemble forecast. The EoE $E(\hat{p}_T)$ is the expected value of its $\hat{p}_T$ PDF, which may be very different from $p_e^*$. Thus, to validate $CES_G$, a certain $p_e^*$ is chosen and then many cases where EoE $E(\hat{p}_T)$ matches $p_e^*$ are found. The aggregate of EoE $\hat{p}_T$ distributions from the many cases should then match the generic $CES_G$ $\hat{p}_T$ PDF.

The validation approach is similar for RCR, with one notable difference. In the case of $CES_G$ where the ambiguity distributions are static, the EoE data does not have to coincide with any particular forecast scenario. RCR on the other hand requires the EoE

results specifically coincide with that of the resampled ensemble due to the flow-dependent aspect. From an EoE forecast case of *N* constituents, a single constituent is drawn to act as the control ensemble forecast. The control is then used to create the RCR ambiguity estimate, which is centered on the calibrated control forecast probability ( $p_e^*$ ). The complete set of *N* constituents is then used to create the EoE ambiguity distribution. Validation is performed where the $E\left(\hat{p}_T\right)$ for both the EoE and RCR ambiguity distributions are equal. As RCR uses random deviates of the long-term average error of the EPS, its $\hat{p}_T$ PDF may be over- or under-spread compared to EoE for any one case. Therefore, it is again necessary to aggregate many forecast scenarios from across the attractor. Thus for validation, the RCR $\hat{p}_T$ distributions and the associated EoE $\hat{p}_T$ distributions are aggregated separately for comparison.

These comparisons show how well the estimated $\mathrm{CES_G}$ and RCR ambiguity distributions capture the variance of the EoE ambiguity distribution when the EoE $E\left(\hat{p}_T\right)$ equals $p_e^*$. However, we cannot validate the estimation methods' ability to consistently capture the location of the EoE ambiguity distribution. Both the CES and RCR ambiguity distributions will be centered on the calibrated forecast probability from the control ensemble forecast, which is a random sample from the EoE ambiguity distribution, thus a random error in location exists.

### 1. Processing of Ambiguity Data

We used the EoE to create 100 sets of 100 constituent ensemble forecasts to be used during validation and value testing. All of the sets were used during the evaluation of value, but computational constraints associated with RCR allowed only 20 of these sets to be used during the validation of the estimation techniques. For each of the 20 EoE forecast cases used for validation, the data were processed to ensure comparisons were performed using ambiguity distributions with equal expected values. The overall processing scheme for the ambiguity data used for validation is shown in Figure 37.

To find the EoE $\hat{p}_T$ distributions for a single set of 100 constituent forecasts, we tested each of the eight $X_k$ variables sequentially across the range of forecast probability values shown in Figure 37. The forecast probability values tested represent possible $p_e^*$ values found using the control ensemble forecast. We will describe the postprocessing procedure for $p_e^* = 50\%$ and $X_k = X_1$ for a single EoE forecast case of 100 constituents here as an example of how all $p_e^*$ values and $X_k$ variables are processed at each lead time.

We must determine the event threshold *X*-value that yields $E\left(\hat{p}_T\right)$ across the 100 constituents equal to the forecast probability value being tested. Thus for variable $X_1$ within the set of EoE constituents, there exists an *X*-value such that the distribution of $\hat{p}_T$ values calculated using that *X*-value as the event threshold with each constituent forecast creates an $E\left(\hat{p}_T\right)$ equal to 50% within 0.01%. Once this *X*-value is located, we know the distribution of constituent $\hat{p}_T$ values for $X_1$ and $p_e^* = 50\%$ in our EoE forecast case. The *X*-value and $\hat{p}_T$ distribution will be different for different EoE forecast cases or even different $X_k$ variables within a single dataset.

We employed an iterative-bisection method for determining the *X*-value (Figure 38). Here, the control EF, taken as the first constituent of the EoE forecast case, was used to find the range of $X_1$ values based on the ensemble members. This range was then expanded on either side by an arbitrary amount, Figure 38(a). We expanded the range after initial tests had difficulties converging on an *X*-value for extreme forecast probability values. The average of the largest and smallest $X_1$ values was then taken as the first test value used to calculate $\hat{p}_T$ across the constituents, Figure 38(b). We then tested the $E\left(\hat{p}_T\right)$ from the 100 constituents against the desired $p_e^* = 50\%$, which resulted in some error value. When the magnitude of the error was too large compared to the tolerance (set to 0.01%), we used the signed value of the error to determine which direction to move when adjusting the range of *X*-values used for determining the next test

*X*-value, Figure 38(c). The process repeats until the algorithm converges. The final output available for use during validation is a distribution of 100 possible $\hat{p}_T$ values for $p_e^* = 50\%$ and $X_k = X_1$ for this EoE forecast case. Again, this process was repeated for all $p_e^*$ values listed in Figure 37 for each $X_k$ at all required lead times for every EoE forecast case.

Application of $\text{CES}_\text{G}$ during the validation process required a control forecast probability value about which the appropriate preprocessed $\text{CES}_\text{G}$ ambiguity distribution was placed. We took the control ensemble forecast for $\text{CES}_\text{G}$ as the first constituent from a set of EoE forecasts. The $\text{CES}_\text{G}$ static ambiguity distribution associated with each possible $p_e^*$ value was found using the process described in Chapter III.C.2.

Similarly to $\text{CES}_\text{G}$, the RCR estimation required defining a control ensemble forecast, which again we took as the first constituent of each EoE set. The resampling process was then performed using the *n* ensemble members. For RCR, we used the uncalibrated control ensemble forecast data, since each resample must be calibrated using randomly drawn calibration coefficients. Once each of the 10,000 re-sampled ensemble datasets had been randomly calibrated, we employed the iterative-bisection method again to find the *X*-value giving $E\left(\hat{p}_T\right)$ equal to some desired $p_e^*$ within 0.01% error. After converging, the desired RCR ambiguity distribution consisting of 10,000 $\hat{p}_T$ values was known. Since this process had to be completed for each variable within each EoE set at every lead time for every desired $p_e^*$ value, the processing time was extraordinary. Therefore for this research, we confined the RCR calculations and thus validation to forecast lead times out to 5 time units at a time increment of 0.2 time units, and processing was only accomplished for a limited number of $p_e^*$ values (see Figure 37). Additionally, we did not perform validation of the later forecast lead times as changes to the ambiguity distributions from all three estimation techniques were insignificant beyond 5 time units.

### 2. Comparing Ambiguity Estimates

The forecast times and probability values available from the computationally expensive RCR data constrain the comparison of the ambiguity estimates from the three methods. Thus, we made comparisons only through forecast lead times of 5 time units at a time increment of 0.2 and for $p_e^*$ values listed in Figure 37. Comparisons were made using total ambiguity [Equation (26), page 61]. So, for each $p_e^*$ value at each time, we found the upper and lower bounds of the 90% CI for each estimate type. As described in the previous section, the expected values of the estimated ambiguity distributions match by design, so comparing the 90% CI ranges provides a measure of the similarity in the variance of the ambiguity distributions. Even if the total ambiguity is equal, we cannot conclude that the ambiguity distributions are the same, since one of the distributions may exhibit differences in higher moments. Thus we are limited to validating only the variance of the ambiguity distributions.

In accordance with the validation theory, we validated specific $p_e^*$ values using aggregates of the EoE and RCR ambiguity distributions. Since 20 EoE sets are used, this resulted in an EoE distribution of 16,000 $\hat{p}_T$ values and an RCR distribution of 1,600,000 $\hat{p}_T$ values. The CES $\hat{p}_T$ distribution contained 50,000 values. The lower and upper CI bounds for a certain distribution were found by sorting the $\hat{p}_T$ values into ascending order and taking the $5^{th}$- and $95^{th}$-percentile based on the size of the dataset, respectively. We then computed the total ambiguity for the distributions as the upper bound minus the lower bound. We compared the $CES_G$ and RCR ambiguity estimates to the EoE "standard" by subtracting the EoE estimate from the $CES_G$ or RCR estimates.

## E. VALUE USING UNCERTAINTY-FOLDING

We applied the uncertainty-folding approach to ambiguity distributions developed using the EoE, $CES_G$ and RCR estimation techniques. Additional testing was performed using what is termed the *grand ensemble*, which consisted of combining the ensemble members from the 100 constituents for a given EoE forecast case into a large 2100-

member ensemble. The grand ensemble may provide evidence that EPS designers would be better off allocating resources towards improving the EPS (i.e., running more members) versus devoting resources to implementing the impractical EoE technique to estimate ambiguity. We ran tests using all of the 100 EoE constituent forecast cases with the two event thresholds previously described (Chapter III.A.4). Thus for each event threshold, we had 800 control ( $p_e^*$ ) and grand ( $p_g$ ) ensemble forecast probabilities and 800 uncertainty-folding forecast probabilities (i.e., one for each of the eight variables in each of the 100 constituents) for each ambiguity estimation method available at each forecast lead time. As before, the control ensemble forecast was always taken as the first constituent of each EoE forecast case. We confined the analysis of value to lead times up to 5 time units because changes in the ambiguity distributions were insignificant beyond this time and because the L96M EPS was shown to lose skill shortly after this time.

For a specific forecast case, we developed the EoE ambiguity distributions by determining the forecast probability for the two selected thresholds for each of the 100 EoE constituents. The $\mathrm{CES_G}$ and RCR ambiguity distributions were found using on the control ensemble forecast (members from constituent #1), using the same procedure described in Chapter III.C. For the grand ensemble, we found a single $p_g$ value using uniform ranks with the 2100-member ensemble for each of the event thresholds. To find the grand ensemble's $p_g$, each of the constituent forecasts was calibrated separately using the average error characteristics for the 21-member L96M EPS. It may have been more appropriate to combine the constituent members and then calibrate using calibration coefficients for a 2100-member L96M EPS, but the computational expense of finding the error characteristics prevented this approach. We then applied uncertainty-folding (Chapter II.C.1) with the EoE, $\mathrm{CES_G}$ and RCR ambiguity distributions to find the $p_a$ value associated with each estimate.

We analyzed value using an extension of the optimal *VS*, called the *integrated optimal VS* (*IOVS*):

$$IOVS = \sum_i \Delta x \begin{cases} 0: \ VS_i \le 0 \\ VS_i: \ VS_i > 0 \end{cases}, \ i = 0.005 \ldots 0.995 \qquad (27)$$

where $\Delta x = 0.01$ and $VS_i$ is the value score attained using *C/L* = *i* as the decision threshold. The summation computes the positive area under the *VS* curve for a given forecast technique at a specific lead time (Figure 39) by breaking the area into sections of width $\Delta x$ and length $VS_i$. In this approach, the optimal *VS* found using $p_e^*$ from the control ensemble, $p_g$ from the grand ensemble or $p_a$ from EoE, $CES_G$ or RCR at a specific lead time is integrated across all *C/L* giving a single *IOVS* value for each source at each forecast lead time.

Using *IOVS* allowed uncertainty-folding from EoE and RCR and the forecast probability from the grand ensemble to be easily compared for all lead times. For comparison, we standardized the *IOVS* values associated with the ambiguity estimation techniques and the grand ensemble with respect to the *IOVS* based on the control forecast probability from the first constituent in each EoE forecast case. Thus, scores greater than one indicate improved value over the control ensemble forecast, while scores below one indicate a reduction in value.

## F. VALUE USING SECONDARY DECISION CRITERIA

We undertook the study of value associated with application of secondary decision criteria using a real-world operational EPS. In the study, we developed a process for applying ambiguity information towards improving the secondary criteria of minimizing repeat false alarms at all locations (i.e., grid points). We used the $CES_L$ ambiguity distributions for this portion of the value study because it is the most practical approach to use operationally over a large domain. Thus we attempt to use ambiguity and decision thresholds, both based on past performance, to add value in a real-world decision context.

## 1. Description of Real-world EPS and Ground Truth Data

We obtained historical ensemble forecast data from the THORPEX Interactive Grand Global Ensemble (TIGGE) database. TIGGE is a collaborative project where ensemble data is made available for scholarly research in support of the THORPEX goals of improving accuracy of 1-day to 2-week forecasts (WMO 2009). The database holds surface and upper level variable data from ten operational centers from around the globe dating as far back as 2001. We retrieved the ensemble forecast and ground truth data through the ECMWF TIGGE internet portal (TIGGE 2009). The following EPS and model descriptions were also obtained from the TIGGE portal.

For the secondary criteria value studies, we chose the Global Ensemble Forecast System (GEFS) provided by NCEP. GEFS is a 21-member, single-model ensemble based on the NCEP Global Forecast System (GFS). The model horizontal resolution provided is T126 (or ~ 110 km) with 28 vertical levels. GEFS forecast data is provided on a $1^o \times 1^o$ grid initialized daily at 00Z, 06Z, 12Z and 18Z over the forecast period T+0 to T+384 hours at a six-hour increment. Initial condition perturbations are produced using an ensemble transform method that incorporates regional rescaling of perturbations with an optimization period of 48 hours. At the time of the study, GEFS contained no model perturbations or surface boundary conditions perturbations, thus ignoring a significant source of uncertainty. Due to the limited number of members and the lack of model perturbations, we expected the ambiguity associated with GEFS to be high, which is why it was chosen for this portion of the research.

We focused the secondary criteria value studies on GEFS 120-hour (T+120) forecasts of 2-m temperature over a CONUS domain with corner points 50N, 125E and 24N, 66E. Based on the $1^o \times 1^o$ grid, this gave us 1,620 forecast-verification points per forecast case, where independence among data points was assumed. Two separate verification periods were used during this study. The first was a training period where we verified the GEFS forecasts to determine the calibration coefficients as well as the forecast error characteristics used for estimating ambiguity. The training period ran from 15 Dec 2007 to 15 Feb 2008 using only 12Z forecasts for a total of 63 forecast days

during the winter season. The period was chosen to avoid seasonal transitions where error characteristics may change dramatically over short periods of time, while providing a large enough dataset for robust estimates of the error characteristics. The second period was an independent application dataset where we performed the value studies employing the error characteristics obtained during the training period. We therefore had to assume stationarity and seasonal dependence of the systematic and random error in order to apply the error statistics to the application period. The application period covered 1 Jan 2009 through 31 Jan 2009 using only 12Z forecasts for a total of 31 forecast days.

To determine the error characteristics associated with the GEFS EPS, we chose the ECMWF global model analysis (T+0 hours) as the ground truth to use for verification. The ECMWF analysis, originally run at horizontal resolution T799 (or ~ 25 km) with 91 vertical levels, is archived on the TIGGE portal using an N200 reduced Gaussian grid. We requested the data on a $1^o \times 1^o$ grid through the TIGGE portal, where the portal software automatically interpolates the data to the user's requested format using a bilinear interpolation (Fuentes 2008). We retrieved 12Z 2-m temperature analyses for 20 Dec 2007 through 20 Feb 2008 for a total of 63 days over the training period. Analyses for the application period consisted of 31 days from 6 Jan 2009 through 5 Feb 2009 for 12Z.

## 2. Metrics used in Secondary Criteria Value Study

From previous discussion, we want to add value to the secondary criteria while leaving the primary value significantly unchanged. The primary value of the forecasts was established using the optimal *VS*. The main secondary value metric was simply the number of repeat false alarms. Other metrics (described below) were used to ensure the primary value of the forecasts was not significantly degraded. We compute all metrics to diagnose any change in primary and secondary value for decisions based on the control ensemble forecast alone versus consideration of the ambiguity information.

In order to ensure the primary value was not significantly reduced, we used the following additional metrics. *Probability of detection* (*POD*) is defined as the proportion

of correctly forecast occurrences (Jolliffe and Stephenson 2003). Based on the nomenclature introduced in the contingency table (Table 2, page 31),

$$POD = \frac{a}{a+c} \tag{28}$$

A metric not normally found in verification readings was defined for this research, the *probability of missed detection* (*POMD*). This metric is defined as the proportion of incorrectly forecast occurrences.

$$POMD = \frac{c}{a+c} \tag{29}$$

The significance of changes to the metrics was evaluated using the 95% CI about the scores. For determining optimization, we considered a change insignificant if the expected value of metric from the alternate behavior fell within the upper and lower bounds of the control's 95% CI.

### 3. Secondary Criteria Value Study Scenario

As described above, we used the secondary criteria of reducing the number of repeat false alarms for this research, where a repeat false alarm is defined as two unnecessary protections in a row at a specific forecast location. Repeat false alarms were chosen because of the tangible and intangible effects they may produce, such as loss of customer confidence in the forecast and degraded mission effectiveness, among others. While a user may employ the ambiguity information to help prevent repeated misses, we considered only repeat false alarms as a secondary criterium. As this was a preliminary study into assessing the value associated with secondary criteria, we chose to focus on a single criterium to show the potential benefits of using ambiguity information in the decision making process.

Here, the goal was to minimize the total number of repeat false alarms over the entire domain by allowing the user to incorporate the ambiguity information to alter the decision to protect at any grid point where the previous consequence was a false alarm (i.e., an unnecessary protection). In other words, at a specific grid point in the domain, if the previous forecast at that location resulted in a false alarm, the user may change the current "protect" decision to "do not protect" if the ambiguity distribution indicates that the decision input is unclear (i.e., overlap exists). The occurrence of any other consequence following the first false alarm breaks the sequence preventing the user from reversing decisions. An important aspect of the study is that we must maintain the primary value associated with using the best-guess forecast probability to minimize expected expense while simultaneously gaining value based on the secondary criteria. The metrics used to monitor the primary value were discussed in Chapter III.F.2. Using the 2-m temperature data, we focused the study on the event threshold of temperature $\leq 0^{o}C$, which is critical to a variety of users in the real world. It is easy to imagine the issue of repeat false alarms extending to any event of interest.

In this scenario, we explored several possible decision rules employed by the forecast user when the chance of having a repeat false alarm is possible (Table 6). Decisions based on users who follow these decision rules are evaluated in relation to a normative user who consistently makes decisions based on the best-guess ensemble forecast (the 'Control' user in the table). The basic decision flow for using the ambiguity distribution overlap to reduce repeat false alarms is shown in Figure 40. It's important to emphasize that changes to the decision can only happen following a previous false alarm at the same grid point. When an opportunity to reverse the decision arises and the decision is unclear, the overlap is compared to an overlap threshold value to determine if the user's action will be changed. If the current overlap is greater than the overlap threshold value, the user will reverse the decision (i.e., choose not to protect).

The conceptual model, depicted in Figure 41, was our first guess at how the overlap threshold value should vary according to *C/L*. We assumed that high *C/L* users needed to reverse the decision less often (i.e., higher overlap threshold) since they are generally not as concerned about false alarms, while users with low *C/L* may be anxious

to reverse the decision since false alarms may occur frequently. For a low *C/L* user, the forecast probability is more apt to indicate that protective action is required, which is likely to result in more negative consequences (i.e., false alarms and repeat false alarms). Thus we assumed the low *C/L* user will want more leeway (i.e., smaller overlap threshold) when the option to reverse the decision is available.

We found the empirical *optimal* overlap threshold by building contingency tables for overlap thresholds varying from 50% to 0.5% at an increment of -0.5% for each *C/L* while also measuring secondary criteria value. When using one of the practical estimation methods (e.g., $CES_L$) to create the ambiguity distribution, the best-guess forecast probability value is generally located at the center of the distribution, thus the maximum overlap value is taken as 50%. An overlap greater than 50% would necessarily result in a different initial decision, and the user would not have the option to change (i.e., the current decision would be to not protect).

For each *C/L*, primary value metrics (*VS*, *POD*, and *POMD*) based on the control user were compared with the metrics computed using the overlap thresholds with the ambiguity information to reverse appropriate decisions. For example, at *C/L* 1%, the control user's metrics were first compared to metrics found using an overlap threshold of 50%. If no significant difference (defined in Chapter III.F.2) was found, 50% was stored as the optimal overlap threshold. Primary value metrics based on subsequent overlap thresholds (49.5% to 0.5% at a -0.5% increment) were also compared to the control user until a significant difference was found, at which point the optimal overlap threshold was taken as the previously stored value. The comparison process (Figure 42) repeated at each *C/L* resulted in an empirically derived optimal overlap threshold value for each *C/L* representing the lowest overlap threshold value that did not significantly degrade the primary value. This optimal overlap threshold has the potential to deliver significant changes to and add value for our secondary criteria.

### 4. Processing of Real-World EPS Data

Using the training dataset of 63 days of 2-m temperature forecasts and observations, we processed the data using the same procedures described previously for

determining the error characteristics of the L96M forecast data (Chapter III.B.3). The GEFS EPS error characteristics (Table 7) were used to provide calibration coefficients for both the training dataset and the independent application dataset. The bulk-calibrated training dataset gave the random error distributions associated with the GEFS forecasts, which were used to generate the required $CES_L$ ambiguity distributions. The variance of the error distributions was determined using subsets based on forecast days to capture the likely flow-dependent sensitivities and the EPS's inability to adequately sample IC and model errors.

The calibration coefficients (shift = 0.0319°C, stretch = 1.64) derived from the training dataset indicated that the GEFS forecasts were on average negatively biased and under-spread. Calibration of the training dataset resulted in a $ME_{\bar{e}}$ of zero and a fractional error in ensemble spread of 0.976 (increased from 0.596), thus even with calibration, the ensemble forecasts were still slightly under-spread. Using the reliability diagrams for the raw and calibrated training dataset forecasts (Figure 43), we see that the calibration improved the reliability and the forecasts are now highly reliable. We used the reliability diagram to compute the reliability and resolution components of the *BSS*. The reliability (*rel*) was improved from $2.04\times10^{-3}$ for the raw data to $1.23\times10^{-4}$ after calibration. From the bin usage histograms, it appears calibration marginally decreased resolution (*res*) (i.e., more forecasts falling outside bins 1 and 11), but both the raw and calibrated data had *res* equal to 0.186. A decrease in resolution was expected since the spread of the forecasts was increased (i.e., made less sharp) during the calibration process. Overall, the calibrated forecasts of 2-m temperature at 120-hr were quite skillful with a *BSS* of 0.756 (increased from 0.752 for the raw data) when compared to the sample climatology.

We also calibrated the application dataset using the coefficients given above, which resulted in a $ME_{\bar{e}}$ of -0.124°C (shifted from -0.156°C) and a fractional error of 1.015 (increased from 0.620). Thus the independent forecasts were still negatively biased after calibration, but the ensemble spread was increased slightly too much. Looking at the reliability diagrams in Figure 44, we see that the calibration performed quite well. In

general, the assumption of stationary (systematic) error characteristics appeared to hold (i.e., difference in $ME_{\bar{e}}$ between the datasets is less than 0.1°C and the spread correction resulted in near perfect fractional error). Because of the stationarity in systematic error, we assume similar stationarity in the random error, indicating that ambiguity estimates computed using the training dataset should apply well to the application dataset. Component *rel* values for the raw and calibrated data calculated from the reliability diagrams were $3.72\times10^{-3}$ and $1.20\times10^{-3}$, respectively, showing the improvement in reliability. The *res* components for the raw and calibrated data were 0.169 and 0.167, respectively, reflecting a slight decrease in resolution. Thus the calibrated forecasts in the application dataset are highly reliable and display fairly high resolution compared to the maximum *res* of 0.245 possible based on the uncertainty (*unc*) [Equation (14), page 47] of the forecast associated with the sample climatology. The *BSS* for the calibrated dataset was 0.677 (increased from 0.673), which indicates quite skillful forecasts that should provide value.

From the description of the NCEP EPS, we expect that ambiguity may be high due to the limited number of ensemble members and its complete lack of model perturbations, but we have seen that ambiguity also varies by forecast error growth and the ensemble spread. We determined the stage of forecast error growth for the 5-day forecasts by comparing the *MSE* of the control forecast (i.e., the first ensemble member) to the climatological variance $\left(\sigma_C^2\right)$ of 2-m temperature taken over the sample dataset. Our comparison of $\sigma_C^2$ (154.44) and $MSE_{\text{det}}$ (20.00) resulted in a value of 12.94%, indicating that the GEFS forecasts were on average still in the early stages of error growth (i.e., since $MSE_{\text{det}}$ may grow to as large as twice $\sigma_C^2$). This result provided more evidence to support our assumption of high ambiguity in the ensemble forecasts of 2-m temperature at the lead time chosen, as the average ensemble variance at this time would still on average be relatively small (discussed in detail in the Results chapter).

For this value study, we employed ambiguity distributions created using $\text{CES}_{\text{L}}$ (Chapter III.C.2), which was configured to produce 50,000 $\hat{p}_T$ values for all forecast

probability values from 0.5% to 99.5% incremented by 0.5%. Ensemble spread values were binned using a class interval of 0.1°C over the range of values from 0-11°C to bin forecasts exhibiting similar uncertainty. In application, any ensemble spread value greater than 11°C used the 11°C bin. The resulting $CES_L$ ambiguity distribution tables were provided at a 1% interval from 1%-99%, with a specific table for each combination of forecast probability and ensemble spread. All together, there are 2,231,100 elements in the tables, which is much too large to show here, but Figure 45 and Table 8 display some sample data. In Table 8, which shows a sample of three ambiguity distributions for $p_e^* = 15\%$, we see the distributions narrow for forecasts with larger ensemble spread. For example, the ambiguity distribution for $\sigma_e = 2\ ^\circ C$ ranges from 0% to above 55%, while the distribution for $\sigma_e = 8\ ^\circ C$ ranges from 0% to 40%, as seen in Figure 45.

We used the training dataset to determine the empirical overlap threshold value for each *C/L* (at an increment of 0.01) using the method described above (Figure 46). We computed the first-order and secondary criteria value metrics based on the "control" user as well as for each of the possible overlap threshold values (50%-0.5% at -0.5% increments). For each *C/L*, we then compared the metrics for each overlap threshold against the control's scores to find the optimal overlap threshold (Figure 42). An example of the comparisons made for *C/L* 0.01 is shown in Figure 47. Looking at the *VS* alone in this example [panel (a)], we would conclude that the optimal overlap threshold was 21.5%, but this threshold still shows a significant difference for the *POD* and *POMD* metrics [panels (b) and (c)]. The lowest overlap threshold value where no significant difference exists for all three metrics is 31.5%, which was taken as the optimal overlap threshold value for *C/L* 0.01 (Figure 46).

The resampling process was not entirely straightforward for this study. For metrics such as *VS*, *POD* and *POMD*, we were able to use the standard approach to resampling, where resample draws may be taken from any grid point on any day. For example, using the training dataset (63 days × 1620 grid points = 102,060 forecasts) resampling may be accomplished by placing the forecasts sequentially in a single column vector and performing 102,060 resamples with replacement for each resampled dataset.

For the secondary criteria value metric of repeat false alarms, the sequential resampling method proved to be inadequate, where results showed that the score for the original dataset was an extreme outlier from the resampled datasets. For repeat false alarms, we found it was necessary to maintain the time-series of forecast-observation pairs at each grid point when performing the resampling. In other words, resampling was performed using only the 1620 grid points (i.e., only 1620 draws with replacement performed), but when a location was drawn, its entire time-series of forecasts and observations over all of the forecast days was taken. This process maintained the consistency of repeat false alarms at a single grid point while allowing the total number of repeat false alarms to vary over the domain for each resampled dataset. The time-series resampling alleviated the problem of the control being an outlier for the secondary criteria value metric, but we found that it underestimated the first-order value metrics. This is most likely because time-series resampling effectively reduces the variance of the results since as it pulls large chunks of data with each draw, and the associated uncertainty within each chunk is not sampled.

The optimal overlap thresholds shown in Figure 46 are close to the reverse of our original conceptual model, indicating that low *C/L* users require a higher overlap threshold than mid to high *C/L* users. As the *C/L* increases into and beyond the mid-range values, the certainty of the forecast required to take protective action increases, which decreases the likelihood of false alarms and repeat false alarms. For these users, the size of the overlap was less important, as any overlap threshold used resulted in minimal and insignificant changes to the primary value because the difference in expense between a false alarm and a miss is small ($C \approx L$). Thus our algorithm resulted in smaller values for the optimal threshold. The low *C/L* users required a larger overlap threshold as a consequence of the large number of opportunities to change, since changing too often with a small overlap threshold likely resulted in an increase in misses, which significantly degraded the scores based on the primary value metrics.

Once the optimal overlap was determined, we then used the application dataset to compute the primary and secondary criteria value metrics using all of the decision rule types described in Table 6. The optimal and conceptual model overlap thresholds were

applied to the forecasts using ambiguity distributions based on the $CES_L$ technique as before. Results from each of these decision rules was then compared to the control user to find any improvement in the secondary criteria while not significantly altering the primary value, and these comparisons are reported in the Results chapter.

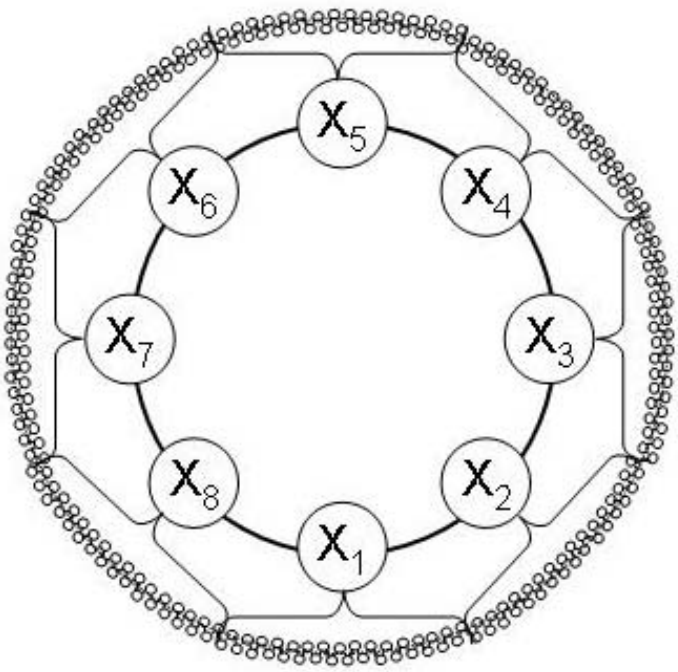



Figure 6. Lorenz 96 System schematic with 8 resolved variables (large circles) and 256 unresolved variables (small circles). The unresolved variables are grouped with the resolved variable to which they belong in sets of 32 [From Wilks 2005].

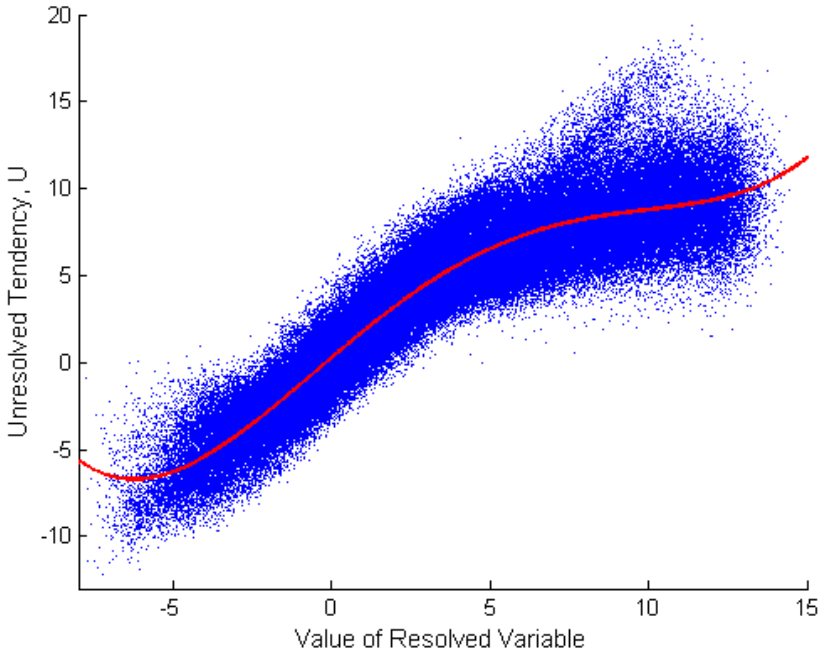



Figure 7. Scatterplot of the unresolved tendency $U$ from all resolved variables as a function of the resolved variable. The fourth-order polynomial regression best-fit (solid line) is the deterministic portion of the parameterization. The average variance of $U$ across all $X$ values about the best-fit line is used for the stochastic portion of the parameterization.

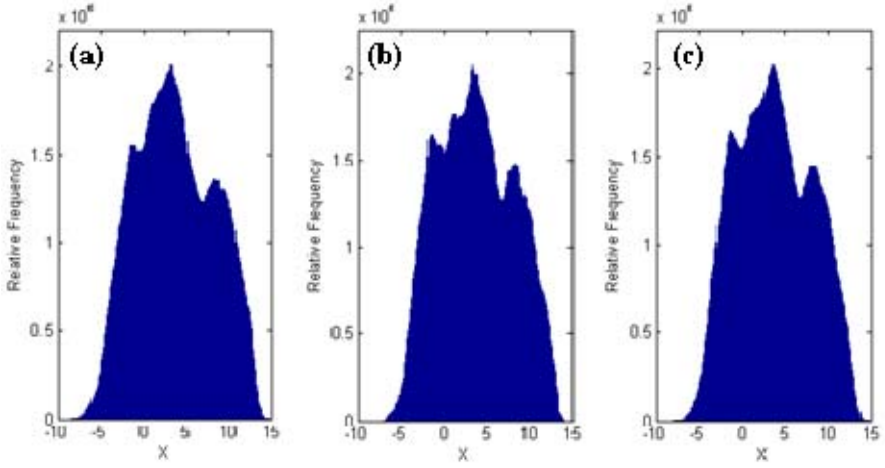



Figure 8. Probability density of resolved ( $X_k$ ) variable using (a) L96 System, (b) L96 Model with deterministic parameterization, and (c) L96 Model with stochastic parameterization.

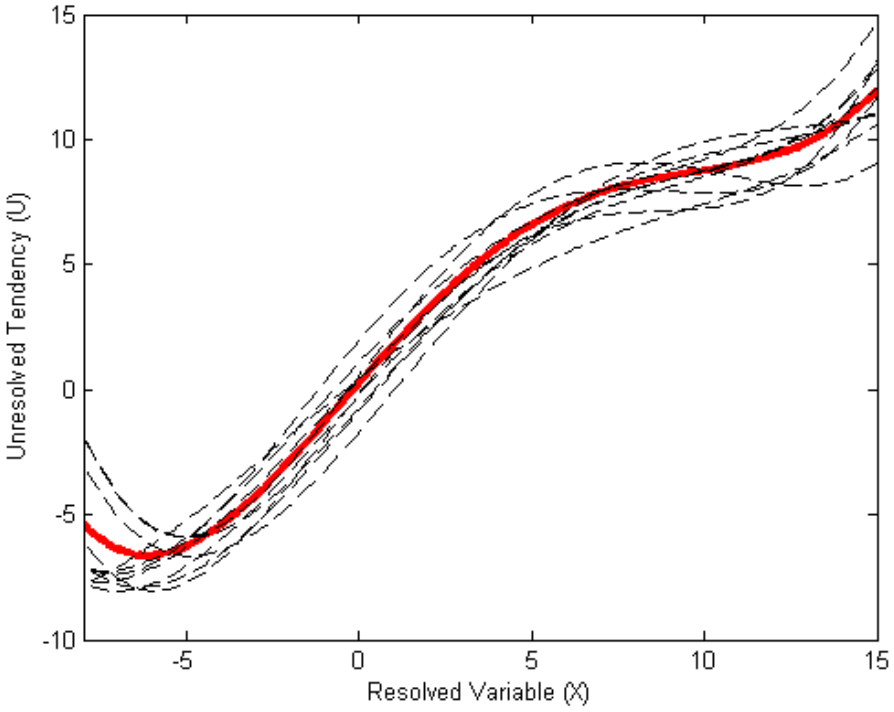



Figure 9. Multi-model EPS deterministic parameterizations. The solid line is the deterministic portion of the stochastic parameterization shown in Figure 7. Dashed lines are static deterministic parameterizations, where each is associated with a specific ensemble member. Only ten members are shown for clarity.

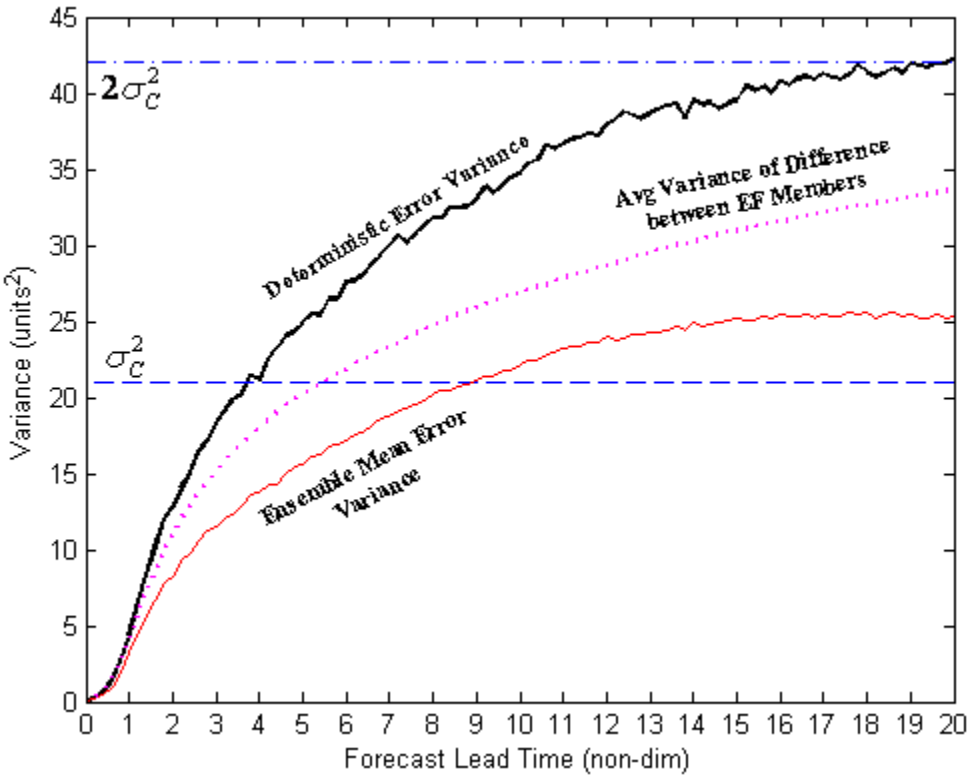



Figure 10. Error variance diagram using L96M deterministic and ensemble forecast data from 24,000 forecast-observation pairs.

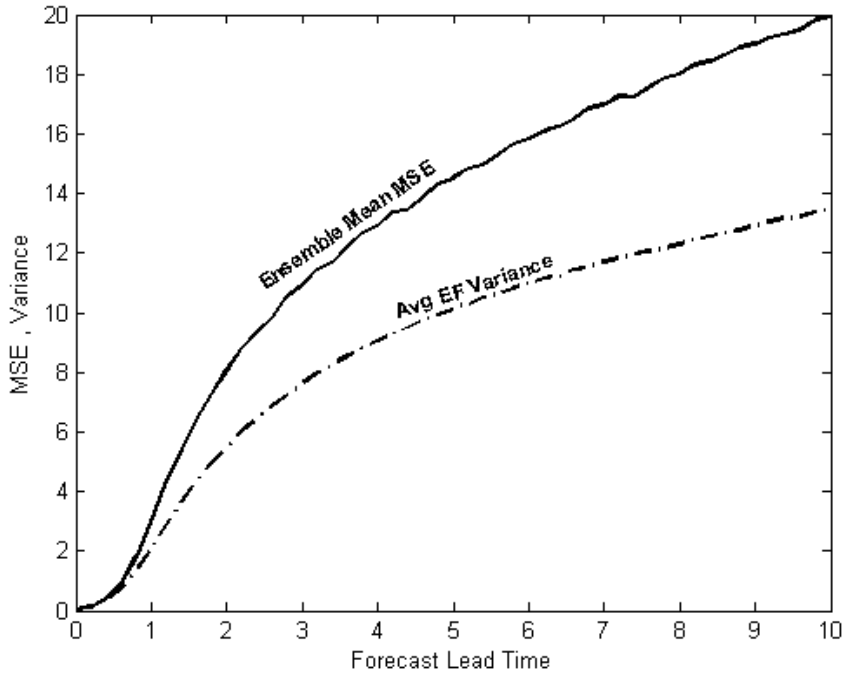



Figure 11. Dispersion Diagram using *uncalibrated* L96M EPS forecast data from 24,000 forecast-observation pairs.

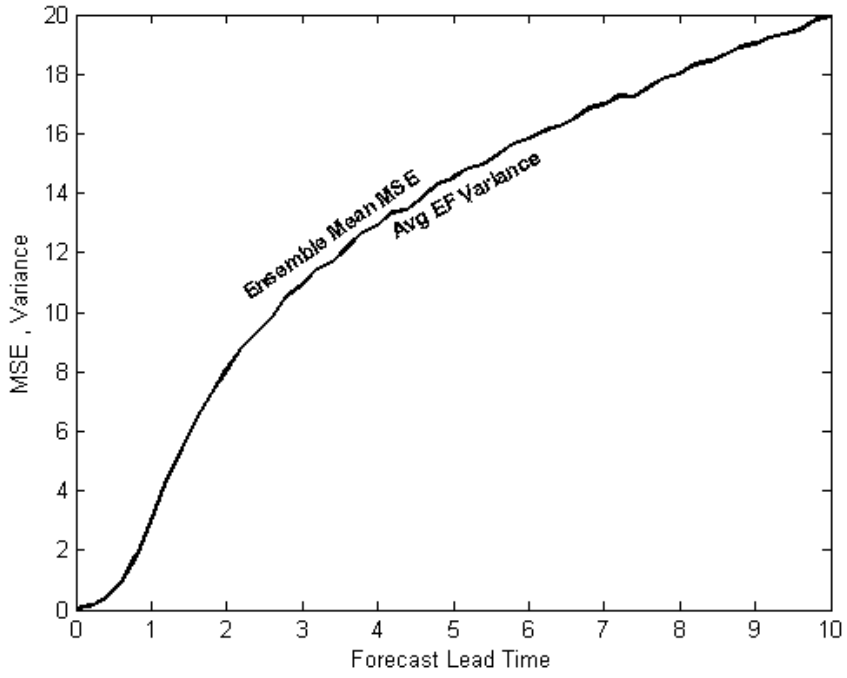



Figure 12. Dispersion diagram using *calibrated* L96M EPS forecast data from 24,000 forecast-observation pairs.

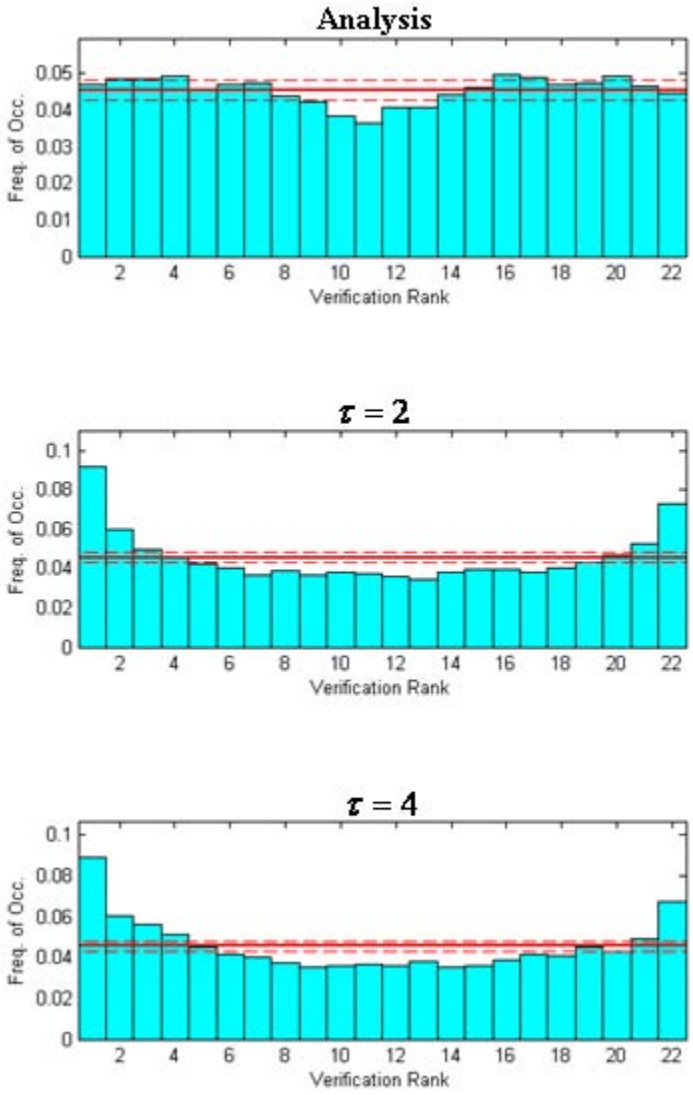
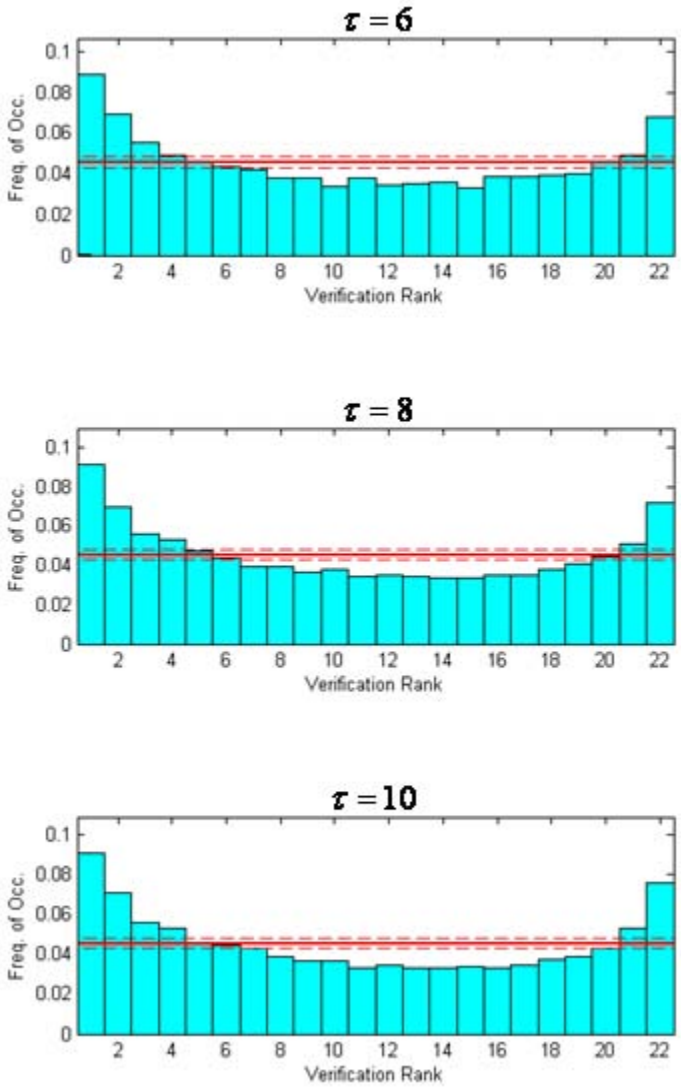



Figure 13. Verification rank histograms using *uncalibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs for various forecast lead times. The solid red line indicates the uniform probability of any rank given a 21-member ensemble. The dashed red lines are the bounds of the 95% CI about the uniform probability given the number of ensemble forecasts (*M*). (Continued, next page.)




(Figure 13, continued.)

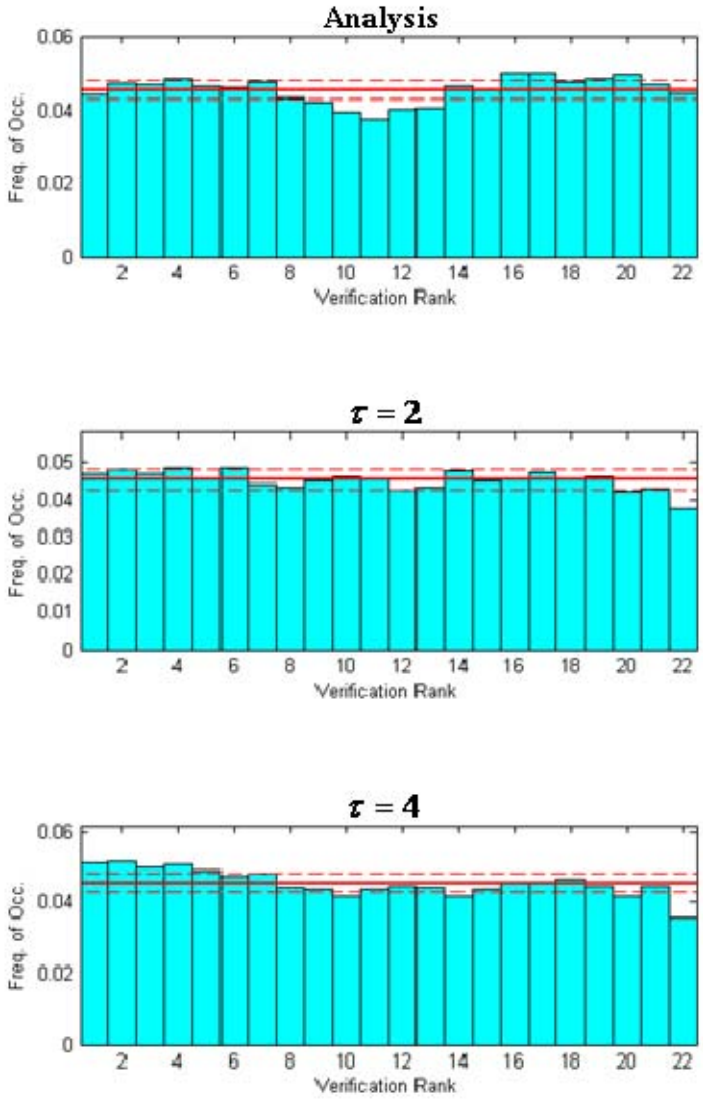
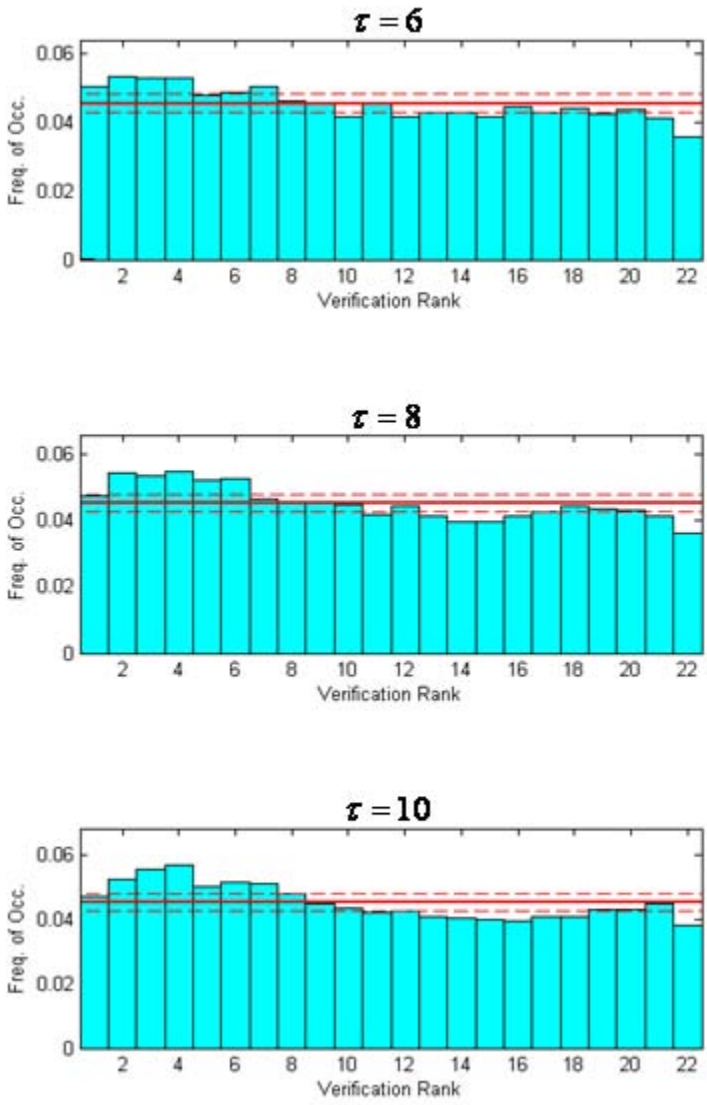



Figure 14. Verification rank histograms using *calibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs for various forecast lead times. Same as Figure 13.




(Figure 14 continued.)

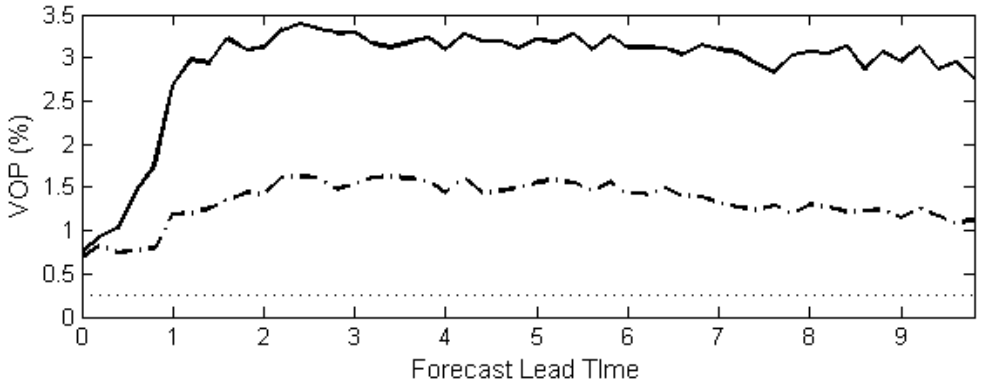



Figure 15. Comparison of Verification Outlier Percentage (*VOP*) values based on the *uncalibrated* (solid) and *calibrated* (dot-dash) L96 EPS ensemble forecast data from 24,000 forecast-observation pairs. The perfect VOP-line of 0.26% is shown by the dotted line.

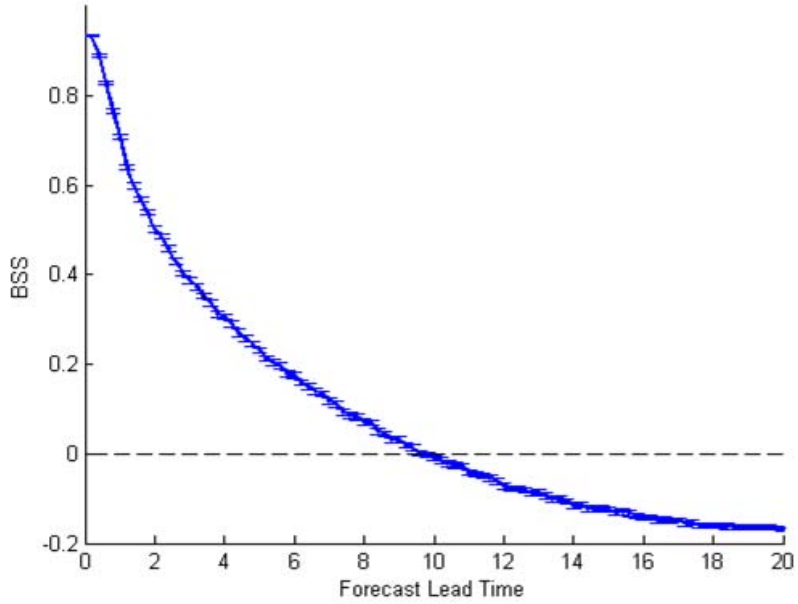



Figure 16. Brier skill score (*BSS*) for the common event using *uncalibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs. Error bars created using bootstrap resampling represent the 95% CI about the BSS value at each forecast lead time. The dashed line is the zero-skill line.

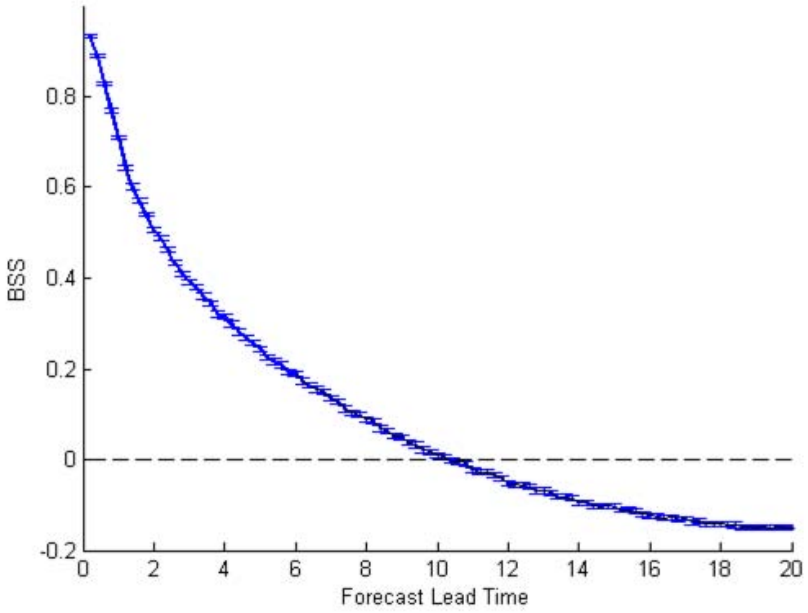



Figure 17. *BSS* for the common event using *calibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs. Same as Figure 16.

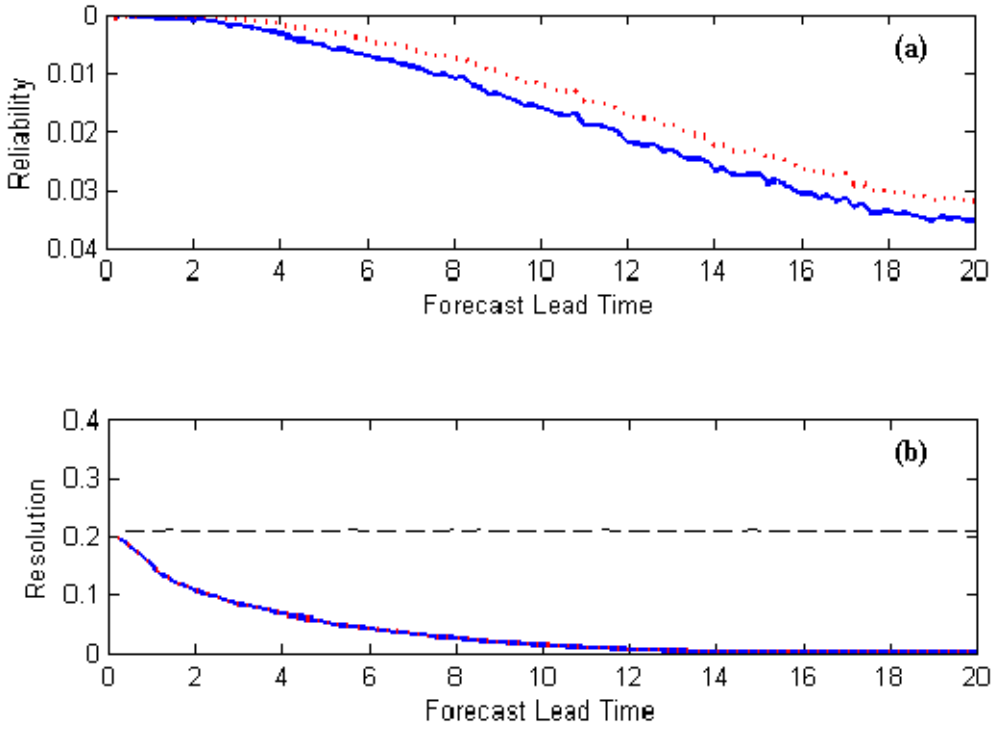



Figure 18. Comparison of (a) reliability and (b) resolution components of *BSS* for both *uncalibrated* (blue solid line) and *calibrated* (red dashed line) for the common event.

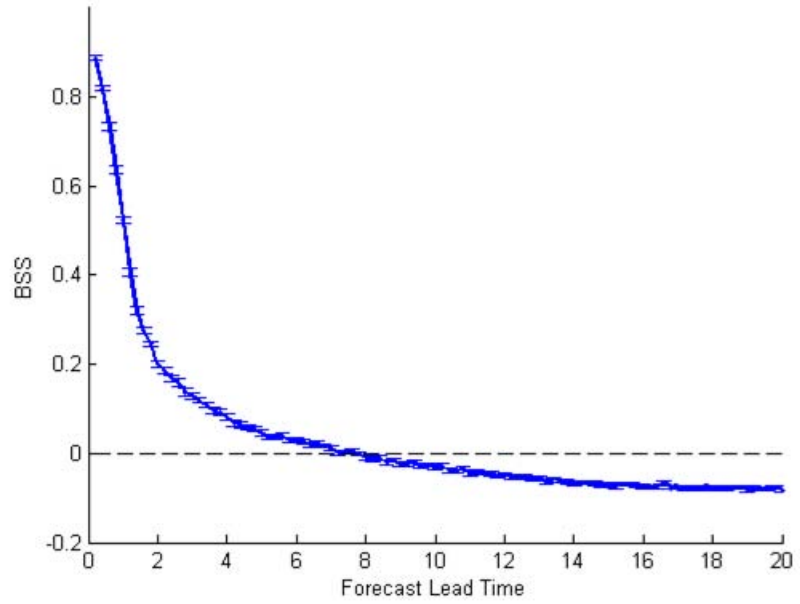



Figure 19. *BSS* for the rare event using *uncalibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs. Same as Figure 16.

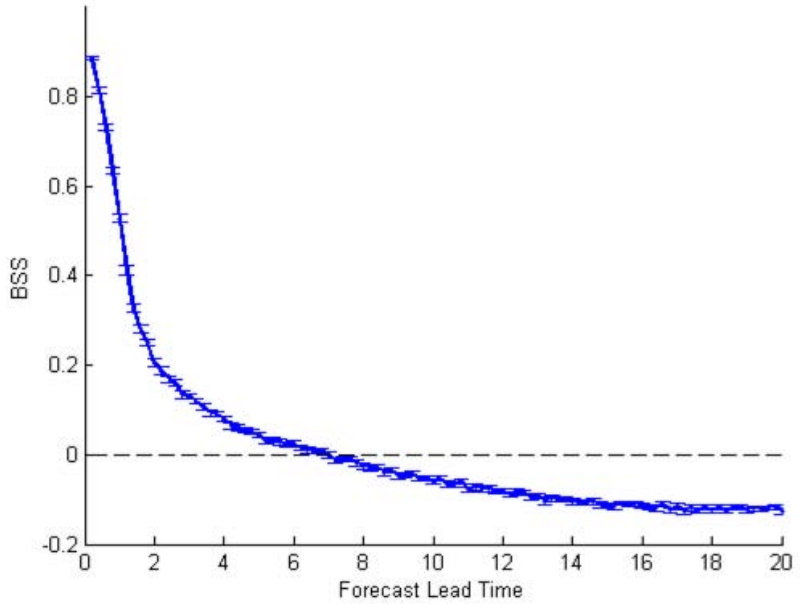



Figure 20. *BSS* for the rare event using *calibrated* L96 EPS ensemble forecast data from 24,000 forecast-observation pairs. Same as Figure 16.

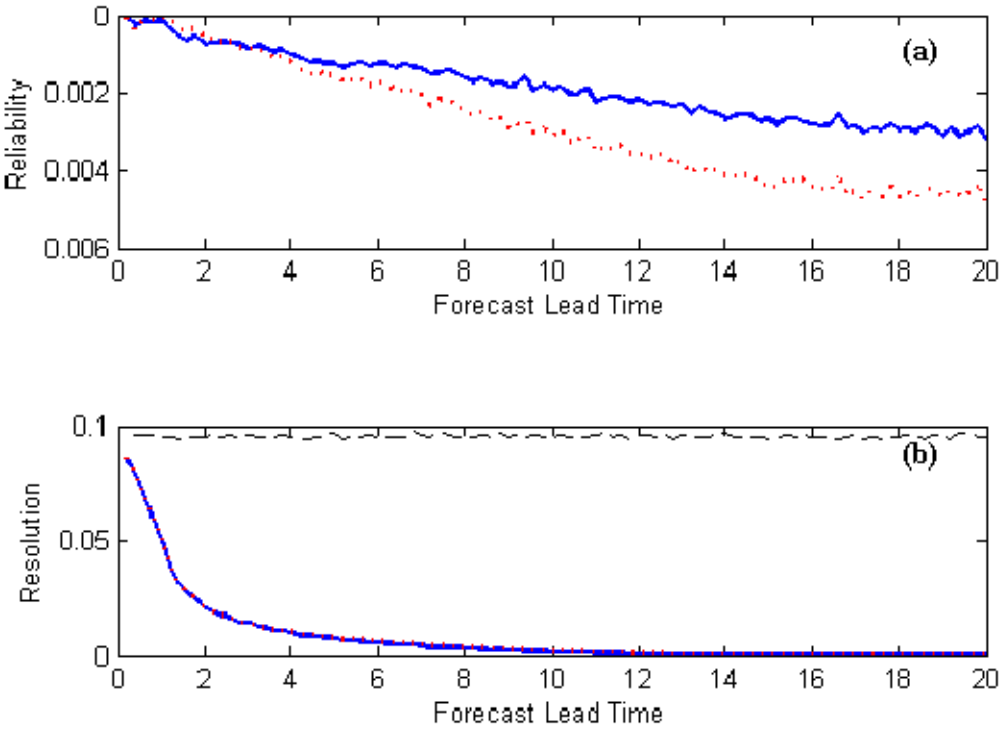



Figure 21. Comparison of (a) reliability and (b) resolution components of *BSS* for both *uncalibrated* (blue solid line) and *calibrated* (red dashed line) for the rare event.


Figure 22. Uniform Ranks method. Calculating forecast probability for $X \geq 5.0$ using a 10-member ensemble. The probability value of 77% is represented by the hatched area [After Szczes 2008].

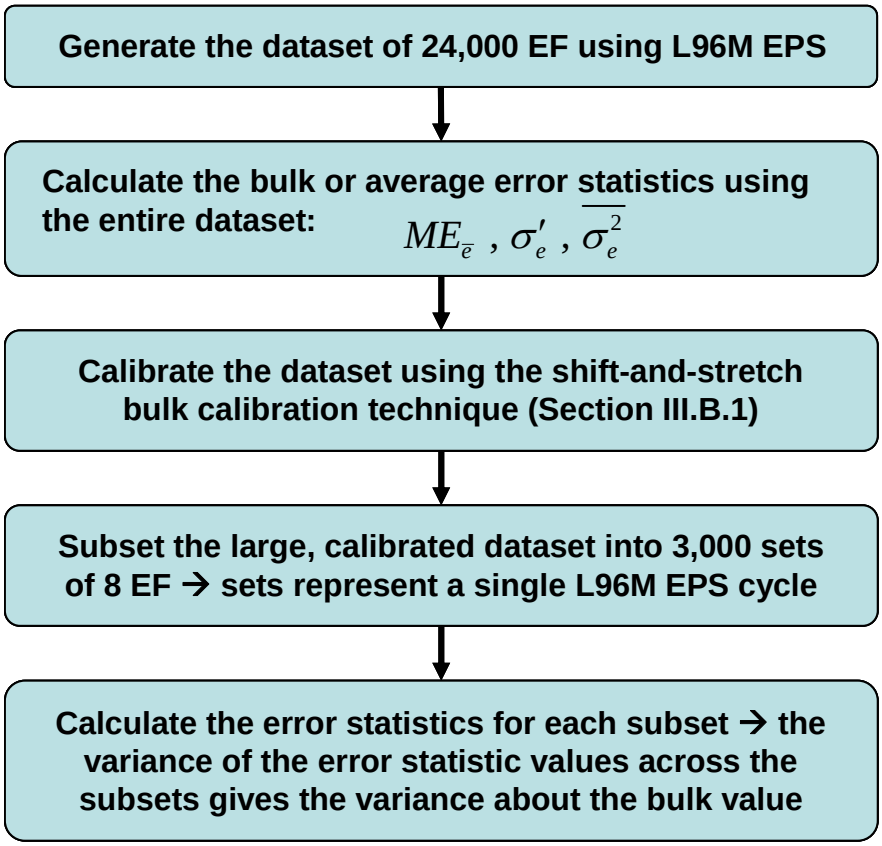



Figure 23. Postprocessing steps for L96 EPS Data.

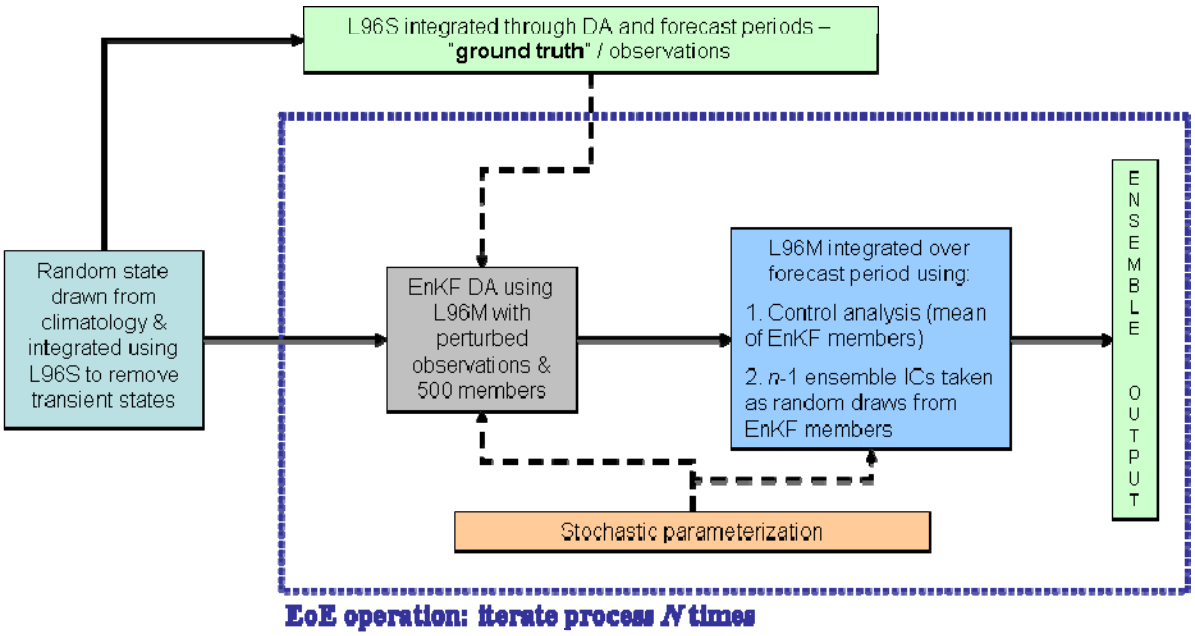



Figure 24. L96M EPS EoE Schematic. After the random starting state is determined, this state is integrated forward through the data assimilation and forecast periods using the L96S. The process inside the dashed box is repeated *N* times using the L96M with the same random initial state to generate the EoE constituents.

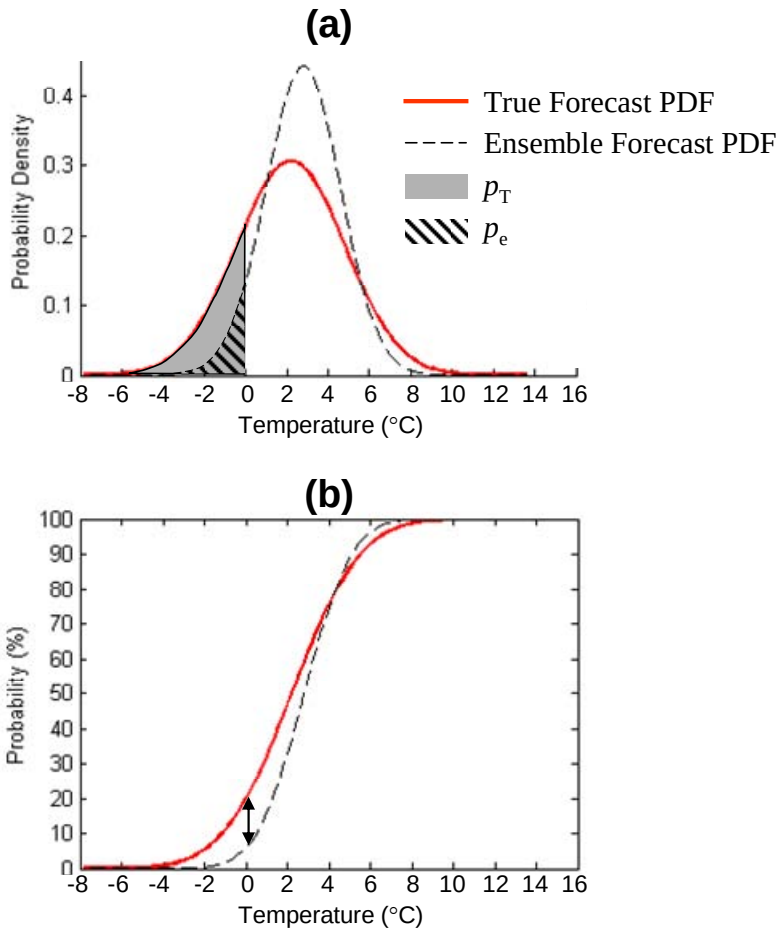



Figure 25. Example comparison of a true and an ensemble forecast PDF (a) and CDF (b) defined as $N$(2.2°C, 2.6°C) and $N$(2.8°C, 1.8°C) respectively. An error of –13.9% in $p_e$ for the chance of temperature ≤ 0°C is the difference in the PDFs' shaded areas, or the difference in the two CDFs (double arrow) [From Eckel and Allen 2009].

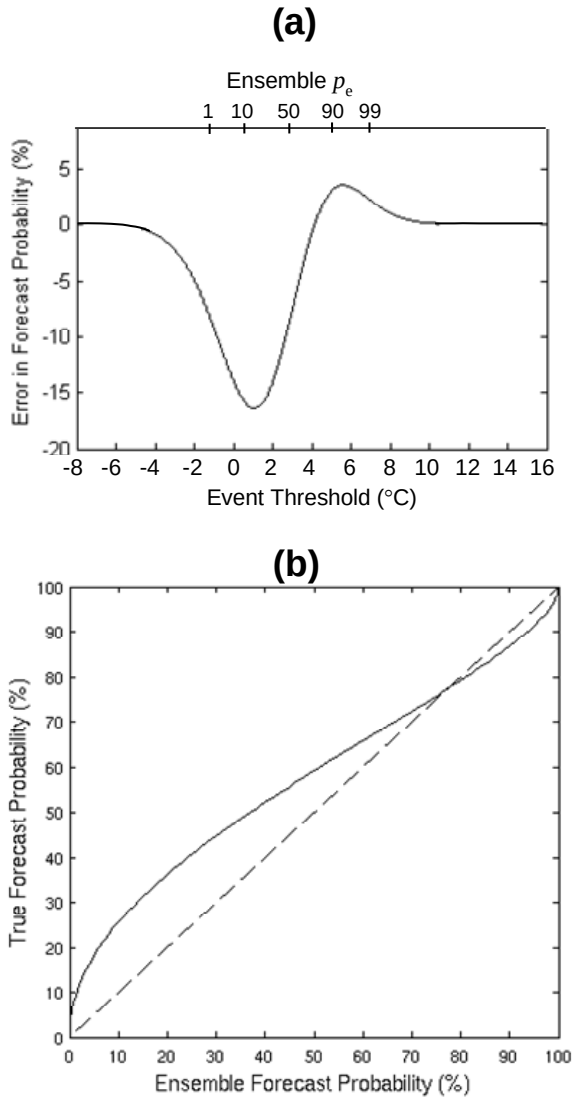



Figure 26. (a) Error in $p_e$ for a range of temperature values for the event threshold, calculated as the difference in the two CDFs of Figure 25. The top axis is the nonlinear $p_e$ scale. (b) Plot of $p_e$ vs. true forecast probability (solid), where the dashed line indicates perfect correlation [From Eckel and Allen 2009].

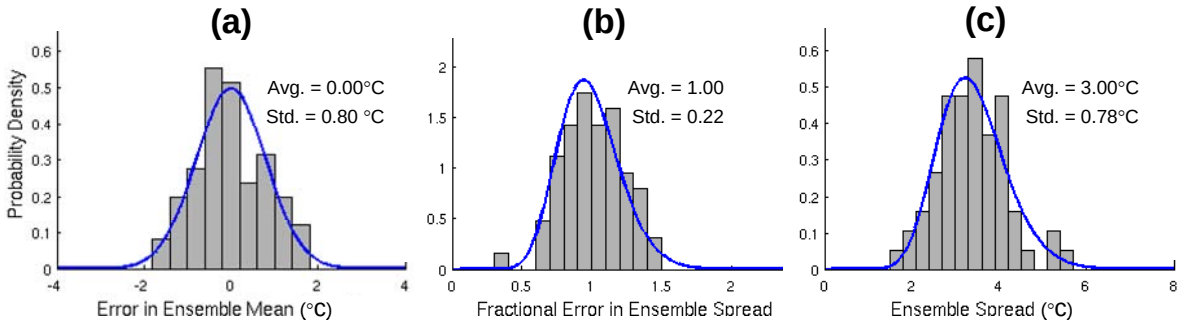



Figure 27. Histogram and fitted PDFs of results from an example bulk-calibrated ensemble forecast dataset for (a) error in ensemble mean, (b) fractional error in ensemble spread, and (c) ensemble spread. The data are based on statistics from the JM 51-member EPS. The domain and forecast period are the same as described in Chapter III.F. [From Eckel and Allen 2009].

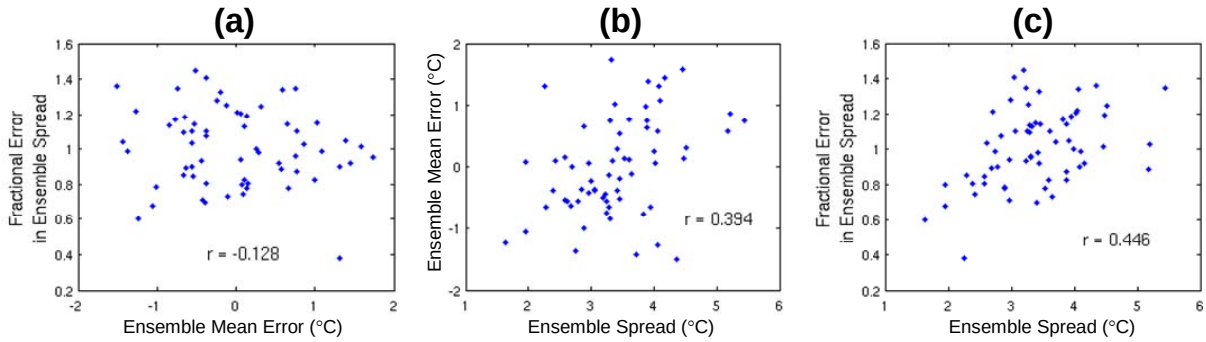



Figure 28. Scatter plots showing relationships between the variables in Figure 27. Correlation coefficient (r) is inset in each plot [From Eckel and Allen 2009].

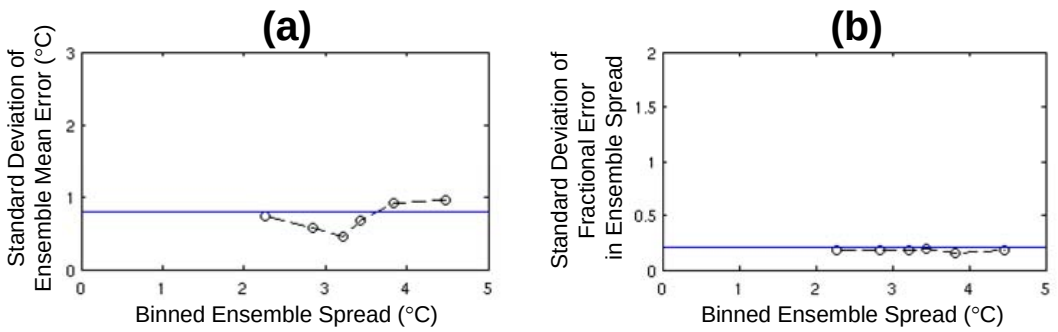



Figure 29. Relationship of ensemble spread with variability (standard deviation) of (a) ensemble mean error and (b) fractional error in ensemble spread. Solid line in each plot indicates the standard deviation of the error distributions in Figure 27 (a) and (b). [After Eckel and Allen 2009].

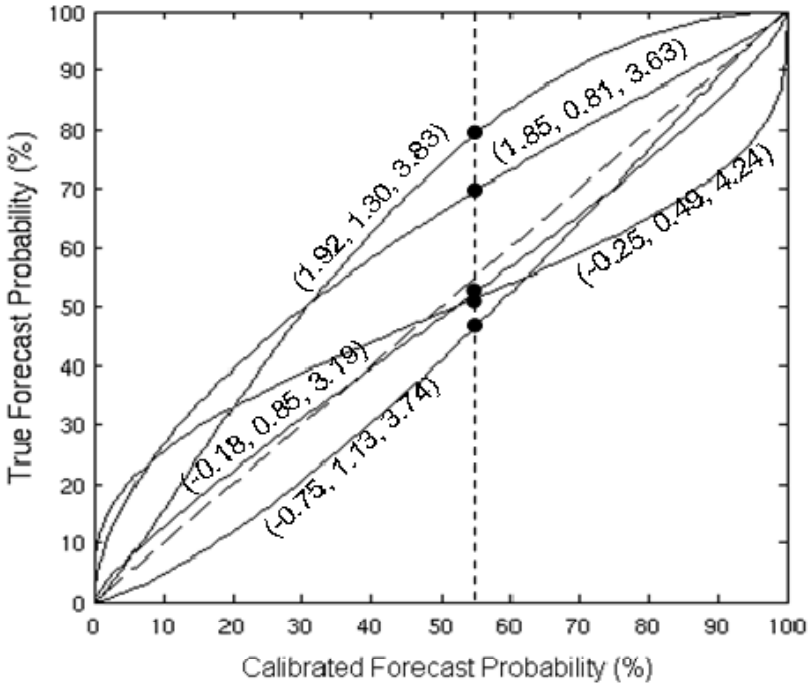



Figure 30. True forecast probability for five sets of random draws from the PDFs in Figure 27 where each curve is labeled with its associated ensemble mean error, ensemble spread error and ensemble spread. The five possible values of true forecast probability (marked by dots) for a $p_e$ of 55% are 79.1, 69.6, 52.4, 51.3, and 46.7% [After Eckel and Allen 2009].

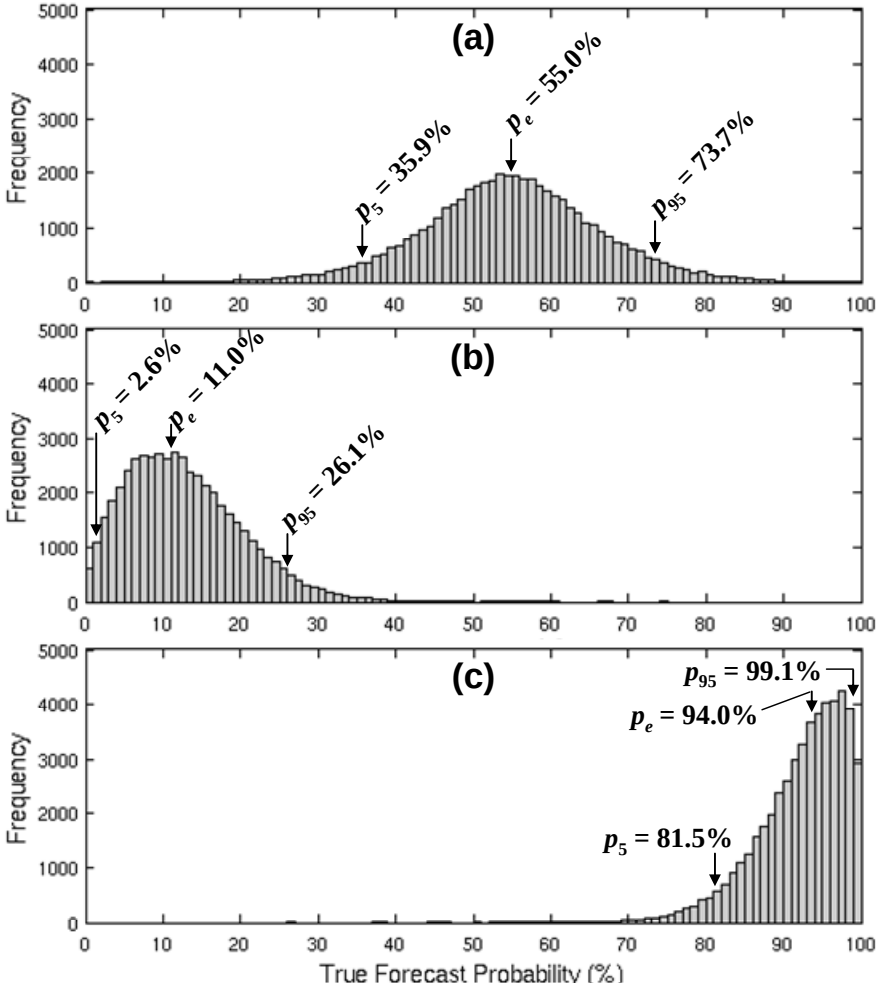



Figure 31. Histogram of 50 000 sample values of true forecast probability for calibrated ensemble forecast probability of (a) 55.0%, (b) 11.0%, and (c) 94.0% generated from random samples from the PDFs in Figure 27. Each histogram is centered on the $p_e$ value from which it was generated since the ensemble forecast PDFs were calibrated. The 5th and 95th percentile values of true forecast probability (for use in Figure 32) are indicated by $p_5$ and $p_{95}$ [From Eckel and Allen 2009].

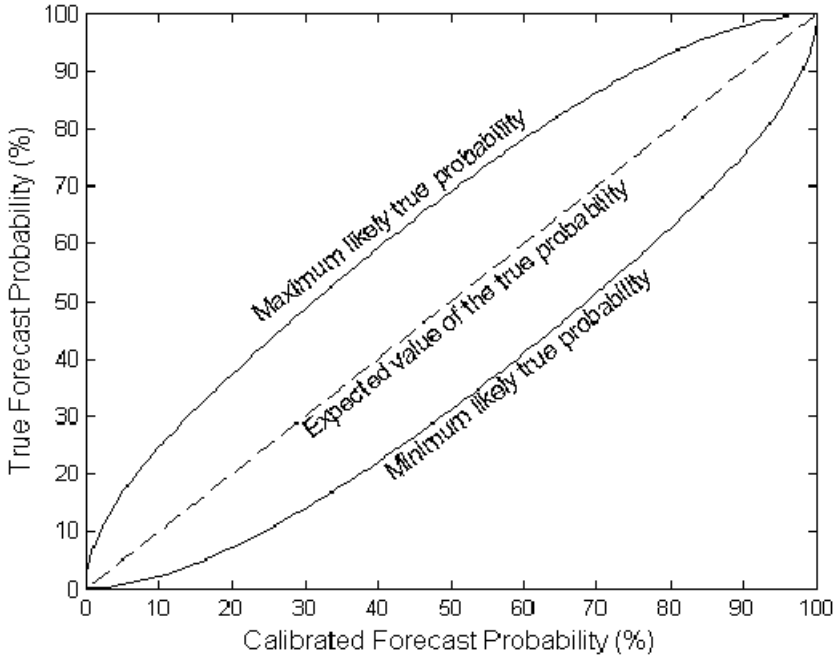



Figure 32. CES ambiguity for all calibrated forecast probability values. After repeated sampling, the 5th and the 95th percentiles of the possible true forecast probability values ( $p_5$ and $p_{95}$) represent ambiguity as a 90% CI about the expected true value (dashed line) for calibrated $p_e$ [After Eckel and Allen 2009].

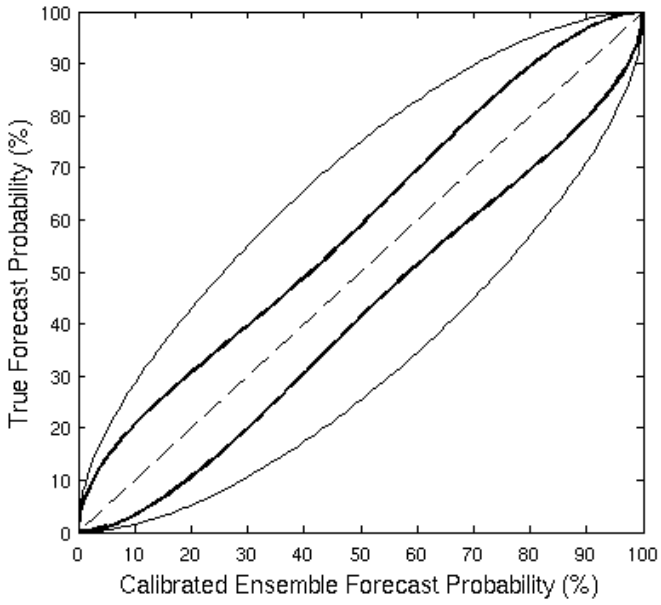



Figure 33. CES ambiguity for all calibrated forecast probability values using a set ensemble spread. Similar to Figure 32 but for specific values of ensemble spread rather than all possible values, but still based on the error distributions in Figure 27 (a) and (b). The thin (thick) curves show the ambiguity for an ensemble spread of 2.0°C (6.0°C) [From Eckel and Allen 2009].

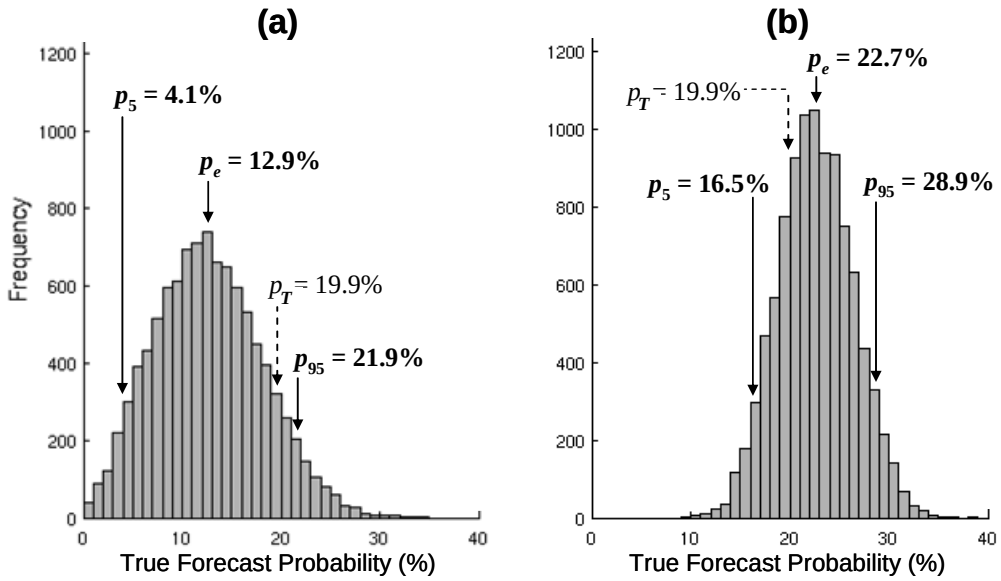



Figure 34. Ambiguity distributions produced by bootstrap resampling of simulated ensemble forecast data (not shown) for (a) An example, perfect 30-member forecast, simulated by 30 random draws from the true PDF in Figure 25 and (b) An example, perfect 80-member forecast simulated using the same true PDF as in (a). The original forecast probability ( $p_e$ ), $p_5$ and $p_{95}$ ($5^{th}$ and $95^{th}$ percentiles that define total ambiguity), and $p_T$ (true forecast probability) are labeled. Total ambiguity values are 17.8% for (a) and 12.4% for (b). Notice that $p_e$ ends up as the distribution's central value [After Eckel and Allen 2009].

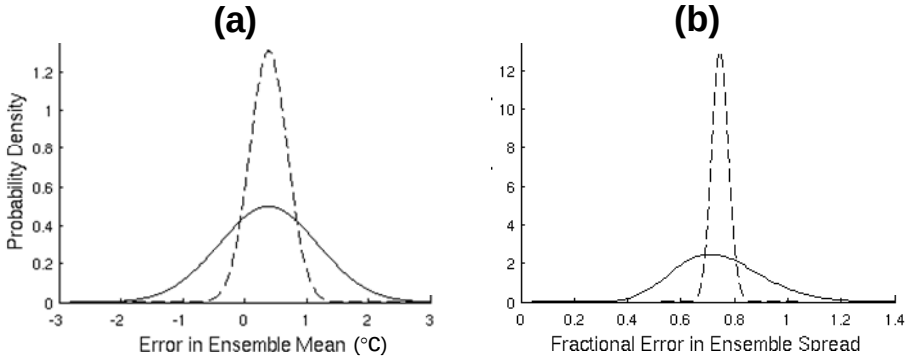



Figure 35. Error distributions of (a) mean error in the ensemble mean and (b) fractional error in ensemble spread. The solid lines are the original, uncalibrated error distributions for the JM 2-m 5-day temperature forecasts. The dashed lines give the reduced error distributions, where the error variance associated with finite sampling (for 51-members) has been removed. The reduced error distributions are used to draw random calibration coefficients during RCR

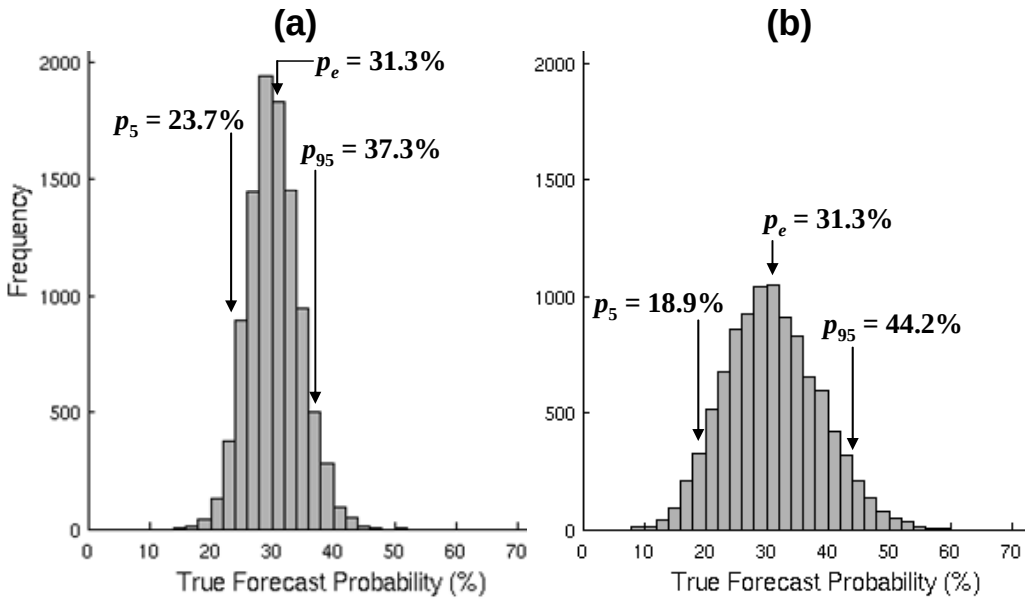



Figure 36. Example RCR ambiguity distributions using (a) fixed, bulk calibration on each resample and (b) random calibration on each resample for the JM 5-day 2-m temperature forecast for a single grid point and date. Note that the random calibration produces a wider ambiguity distribution [After Eckel and Allen 2009].

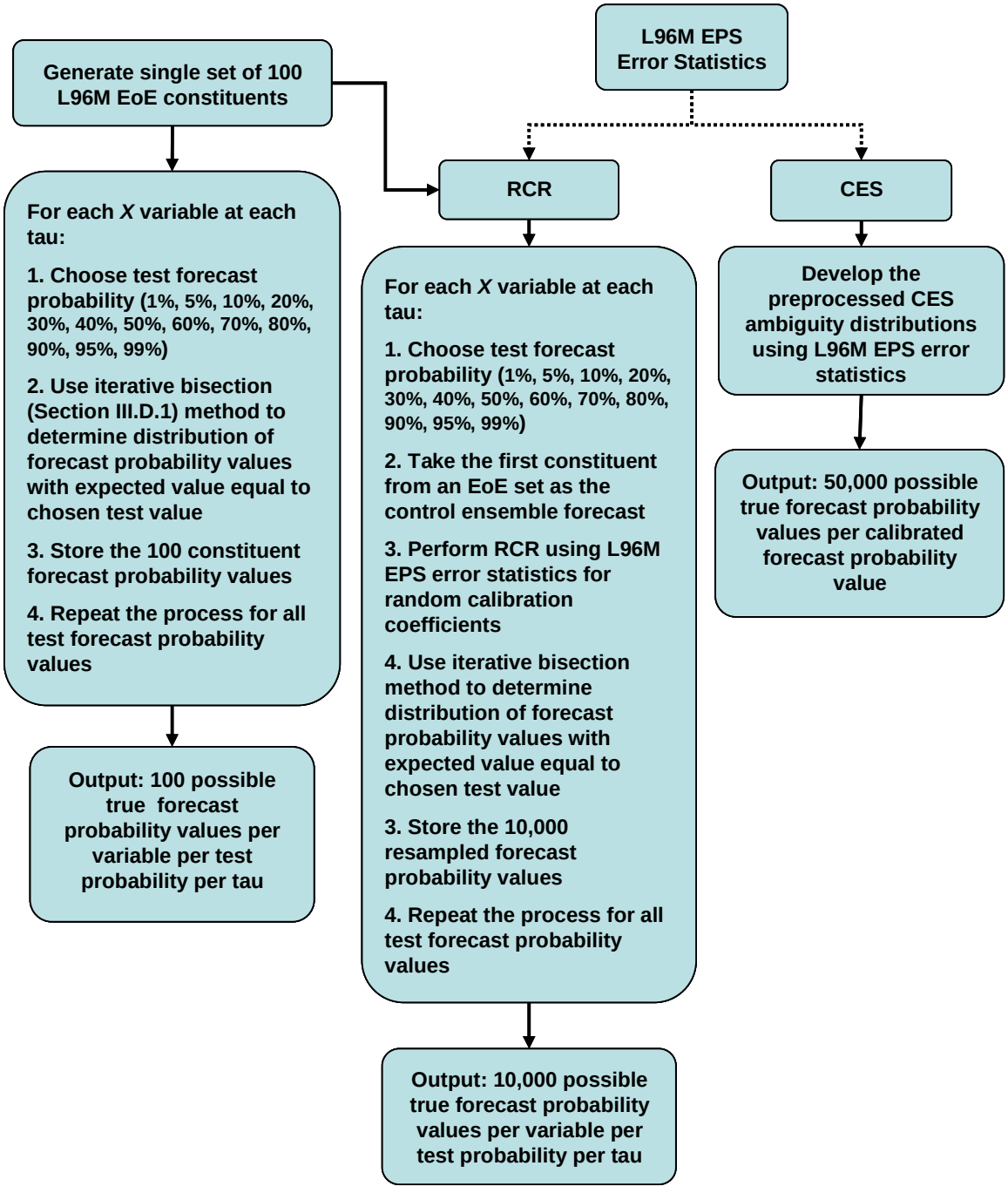
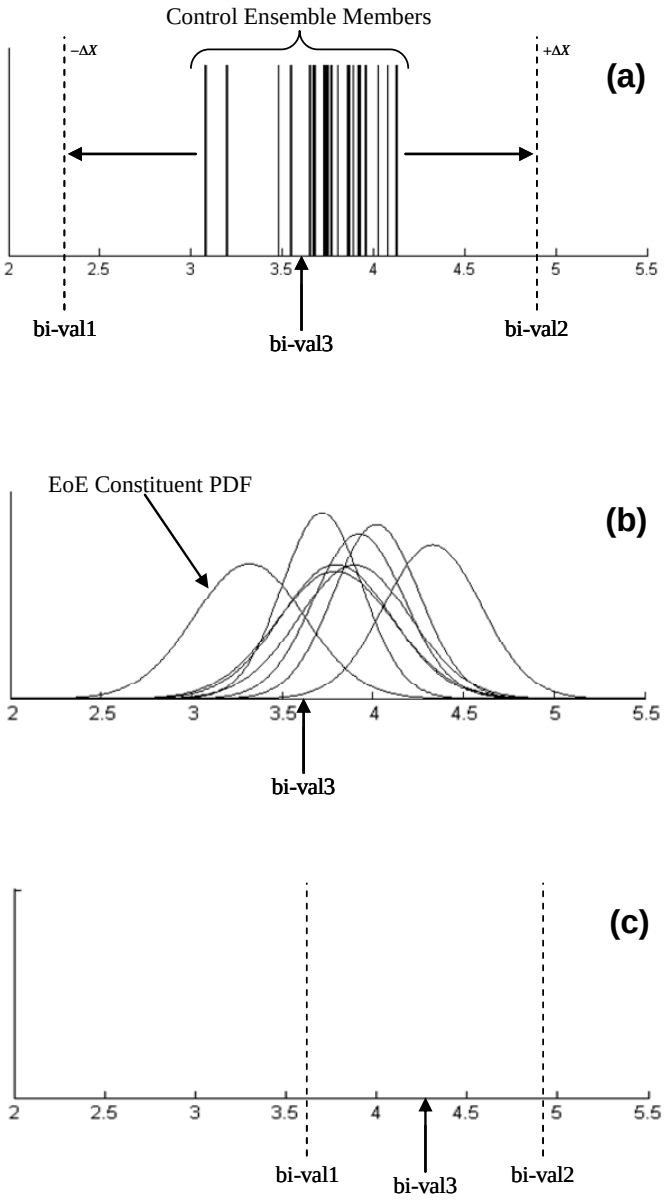



Figure 37. Post-processing steps for ambiguity data for the three estimation techniques.




Bisection Algorithm:

1. Using the control ensemble forecast, find the max and min value of the ensemble members. Add -/+ $\Delta X$ to the min/max values to find bi-val1 and bi-val2. **(a)** Expanding the ensemble range helps the algorithm converge on extreme probability values.

2. Find the mean of bi-val1 and bi-val2 → bi-val3. **(a)**

3. For all EoE constituent forecasts, find $P(X \geq \text{bi-val3})$, and find the expected value of the constituent's probabilities ($E[\,p_e\,]$). **(b)**

4. Compare $E[\,p_e\,]$ to the desired probability value ($p_e^*$):

$$error = p_e^* - E[p_e]$$

5. If the $|error| > \varepsilon = 0.0001$ and:

$error > 0$, set bi-val2 = bi-val3

- or -

$error < 0$, set bi-val1 = bi-val3 **(c)**

Return to step 2 and repeat process using the new bi-val values. **(c)**

6. If $|error| \leq \varepsilon = 0.0001$, the algorithm has converged.

Figure 38. Iterative-bisection method used to converge on the *X*-value giving the expected value of EoE constituent or RCR resampled $\hat{p}_e$ values equal to some desired $p_e^*$ value.

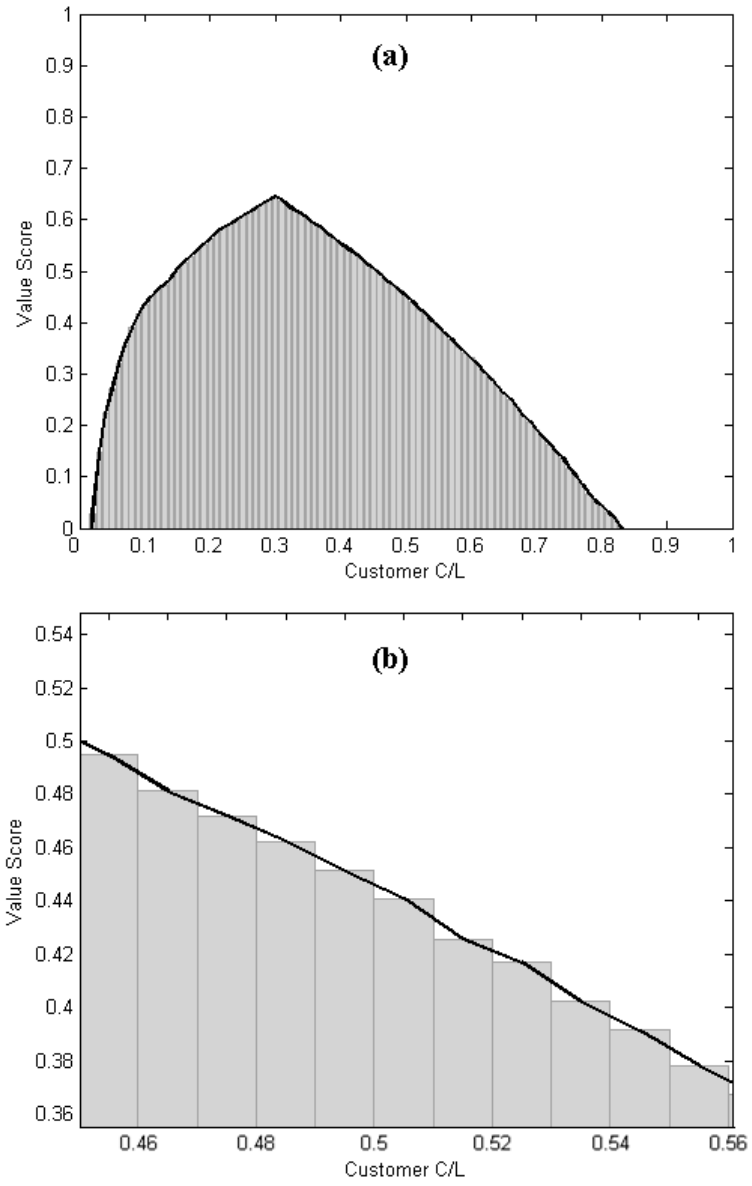



Figure 39. Integrated optimal *VS* (*IOVS*) example for the control forecasts at a single forecast lead time. (a) The optimal *VS* is computed using the 800 control forecast probability values at $\tau = 2.6$. The positive area under the curve is computed using Equation (27) by summing the area of intervals (gray regions) from *C/L* 0-1 using a $\Delta x$ of 0.01. (b) The $\Delta y$ of each interval's area is the optimal *VS* at the center of the interval (e.g., for the interval 0.51-0.52, $\Delta y$ is the optimal *VS* at *C/L* = 0.515). An interval's area is taken as zero if the optimal $VS \leq 0$.

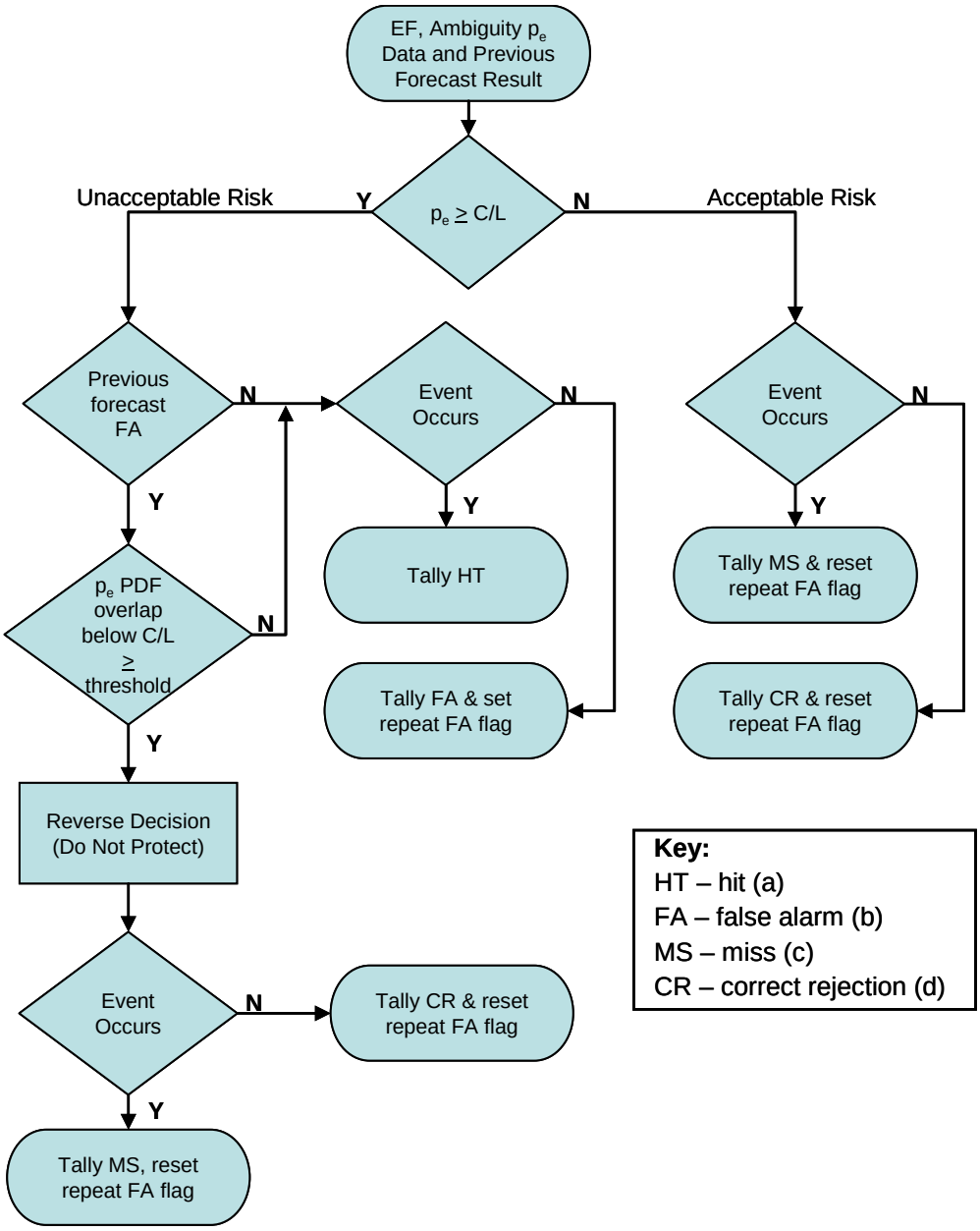



Figure 40. Flowchart of decision process for the repeat false alarms secondary criteria scenario using the ambiguity distribution overlap. Tallying indicates filling in the contingency table (Table 2, page 31) for the current decision rule (*C/L*). The setting of the repeat false alarm flag determines the outcome of the "Previous forecast FA" decision point, where a set flag equals **Y**.

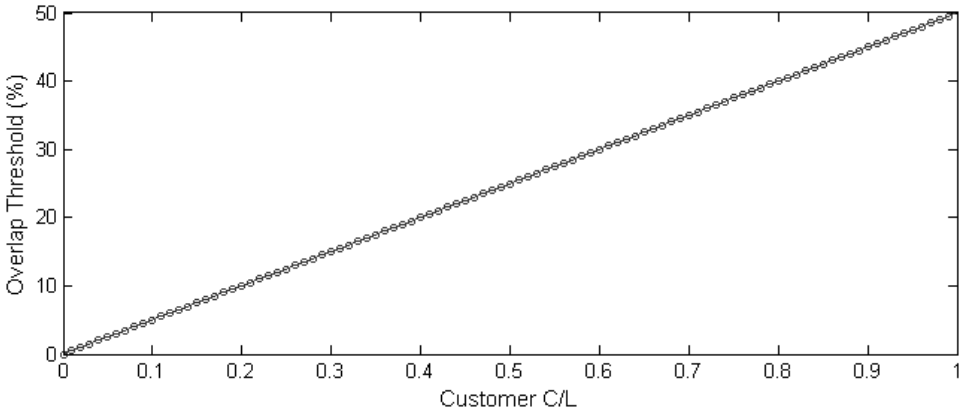



Figure 41. Overlap threshold conceptual model as a function of *C/L* for the repeat false alarm secondary criteria value testing scenario.

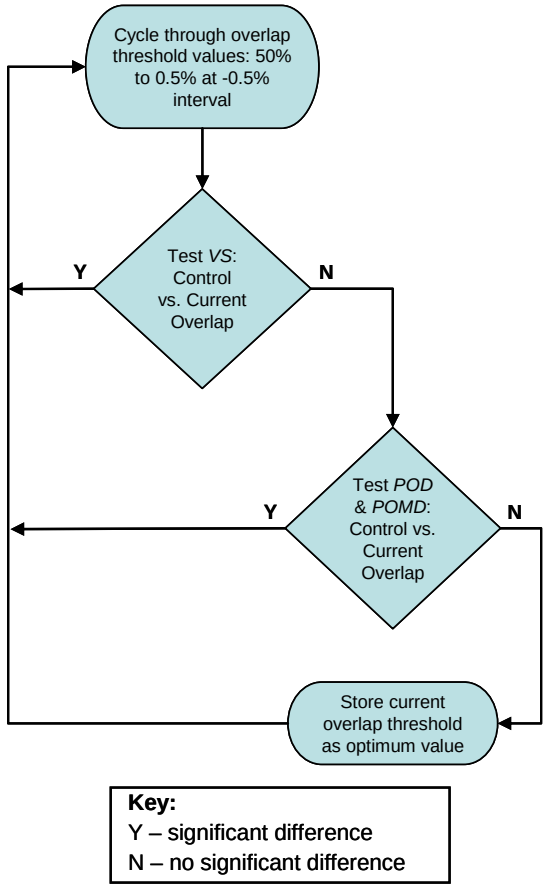



Figure 42. Flowchart for determining empirical secondary criteria overlap threshold value. Performed for each *C/L*, testing compares the metrics derived using the control forecast probability versus using the current overlap threshold.

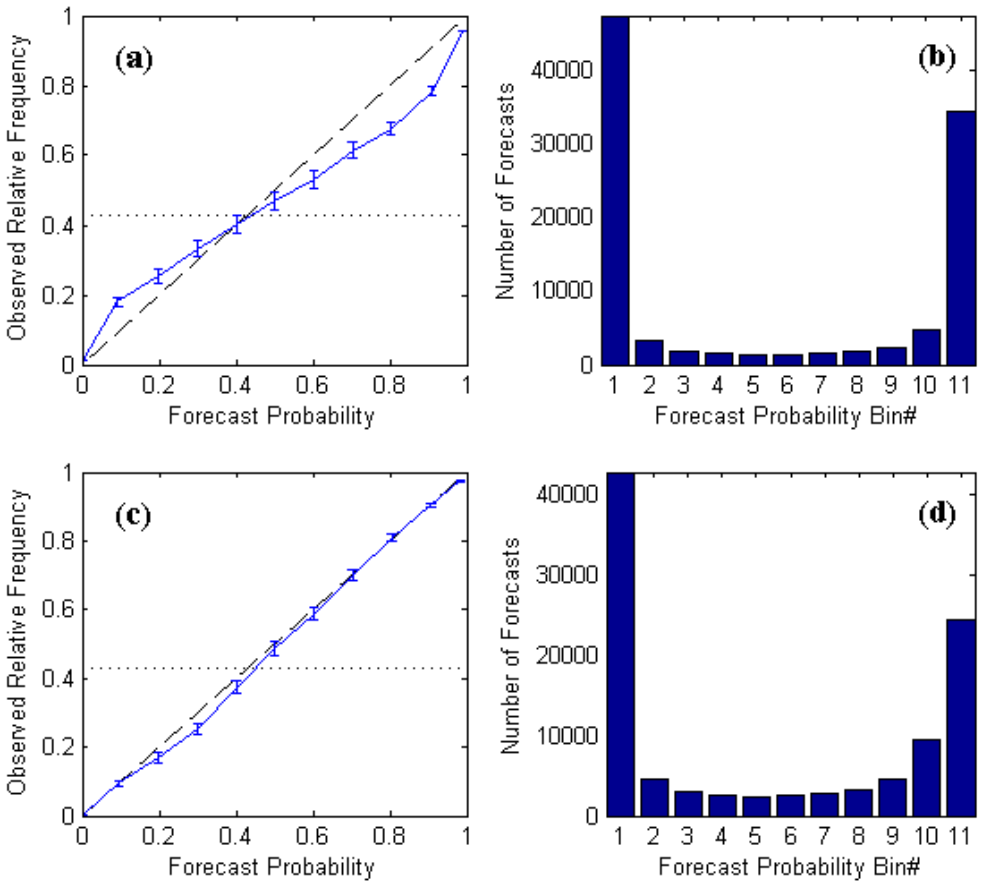



Figure 43. Reliability diagrams for raw and calibrated NCEP GEFS forecasts based on the training dataset with 102,060 forecast-observation pairs. The reliability diagrams for the (a) raw and (c) calibrated data used 11 forecast probability bins (0-0.05, 0.05-0.15, 0.15-0.25,…, 0.95-1.0) where the average forecast probability with each bin is used as the bin's representative value. Error bars represent the 95% binomial CI (Wilks, 2006). The dashed line indicates perfect reliability, while the dotted line shows the sample climatology. The bin usage histograms for the (b) raw and (d) calibrated data give the number of forecast probabilities falling in each of the 11 bins.

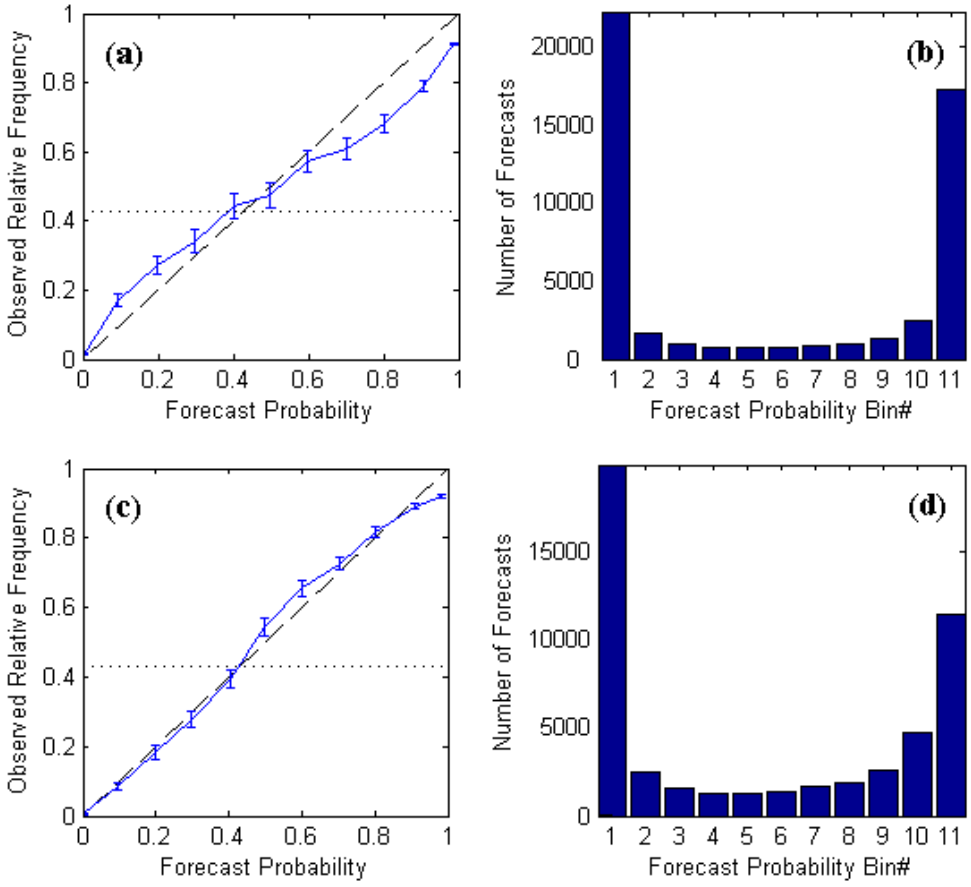



Figure 44. Reliability diagram for raw and calibrated NCEP GEFS forecasts based on the independent application dataset with 50,220 forecast-observation pairs. Same as Figure 43.

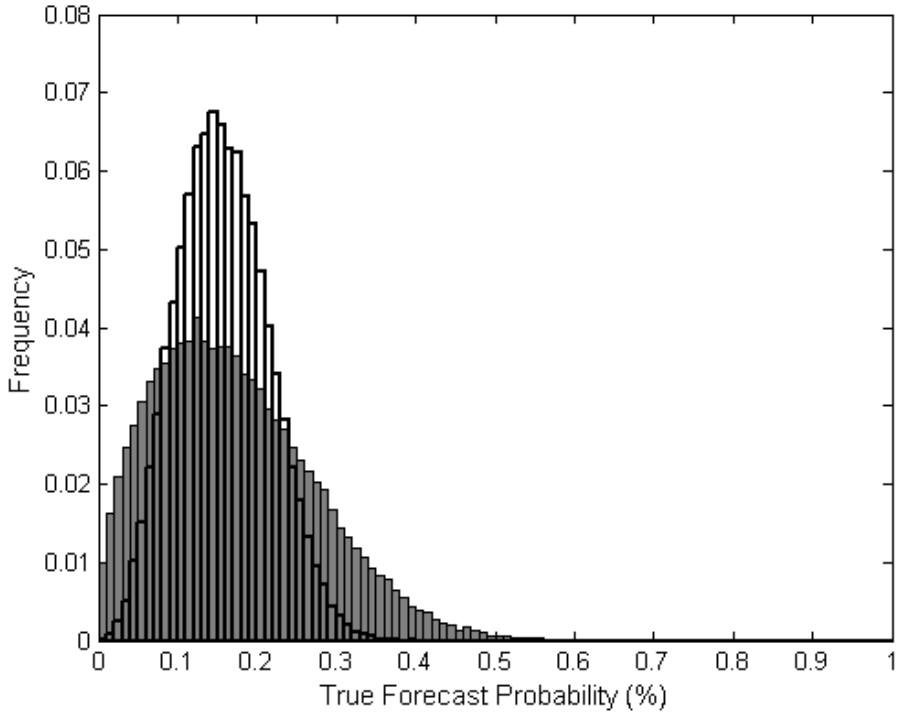



Figure 45. Sample $\text{CES}_{\text{L}}$ NCEP GEFS 21-member EPS ambiguity distributions created using error statistics in Table 7. The histograms show the relative frequency of $\hat{p}_T$ values for $p_e^* = 15\%$ with $\sigma_e = 2\ ^\circ C$ (gray) and $\sigma_e = 8\ ^\circ C$ (transparent).

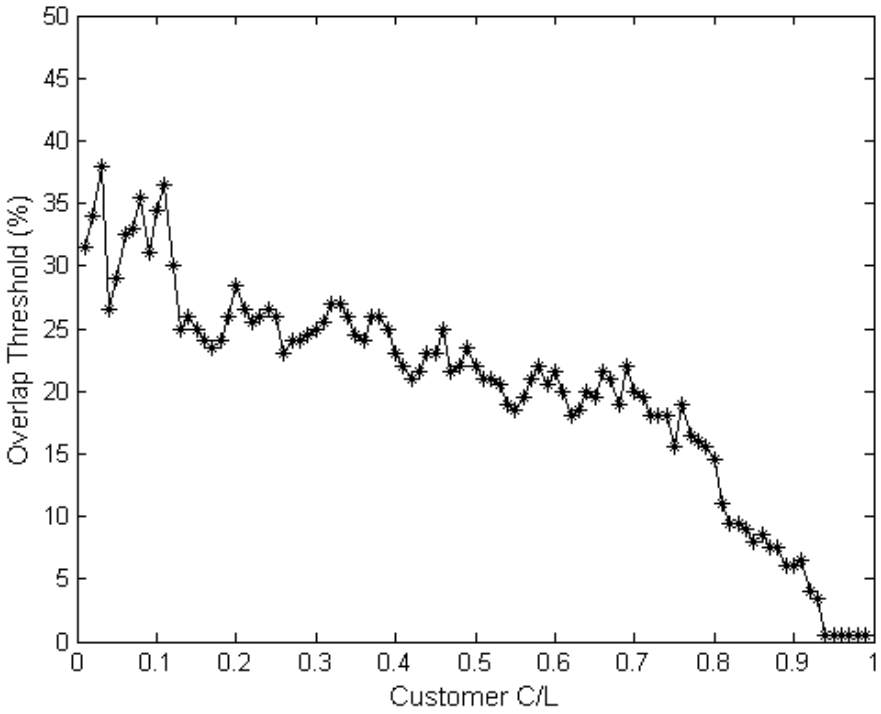



Figure 46. Empirical optimal overlap threshold for reducing repeat false alarms for the event 2-m temperature $\leq 0^{\circ}C$ using the NCEP GEFS training dataset. The optimal overlap threshold is computed at each *C/L* from 0.01-0.99 at an increment of 0.01 (solid line).

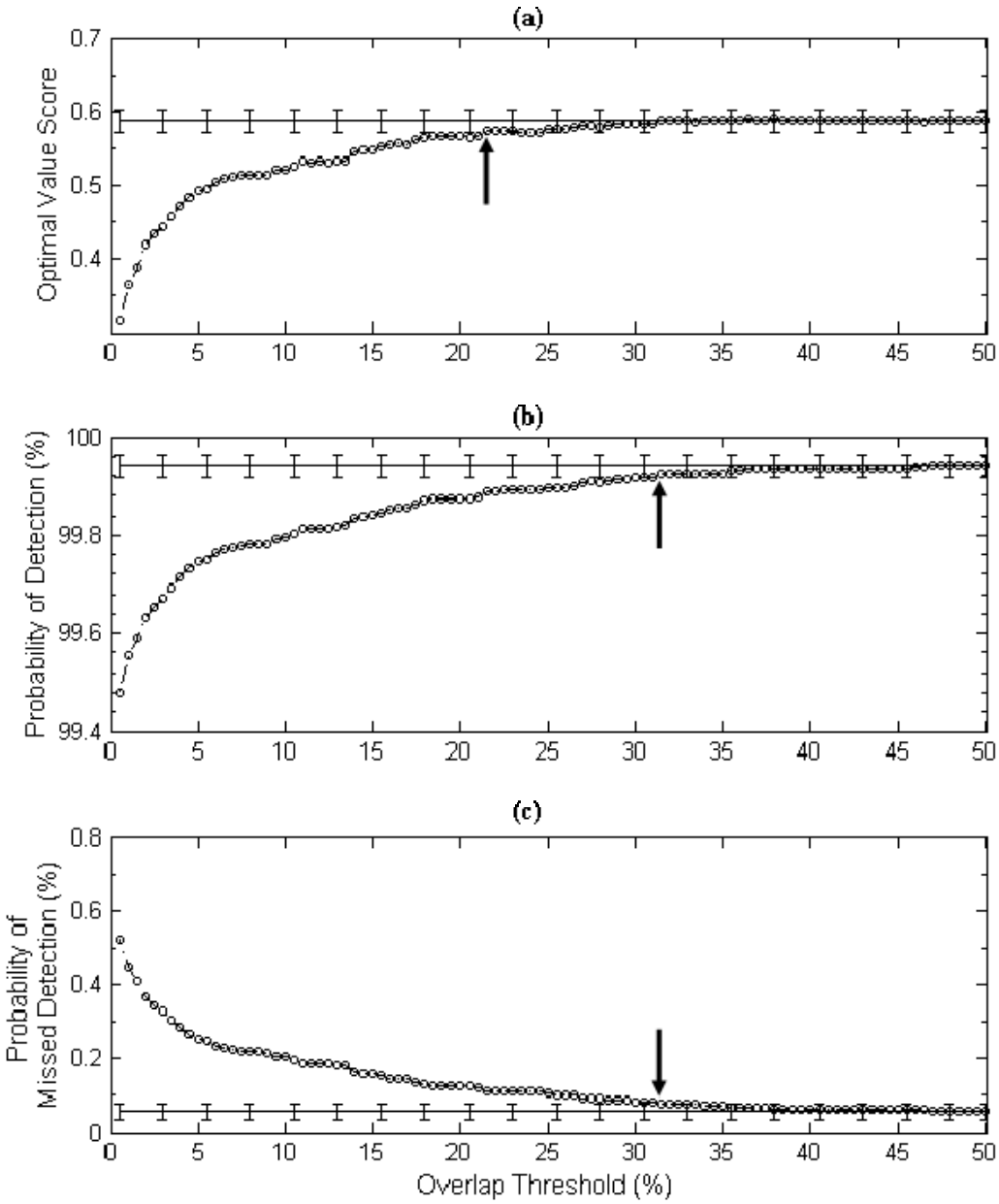



Figure 47. Comparison of primary value metrics (a) optimal *VS*, (b) *POD* and (c) *POMD* used to find the optimal overlap threshold for *C/L* 0.01. Control scores in all three panels are shown by the solid line with error bars representing the 95% CI. The expected value of metrics using overlap threshold values from 0.5% to 50% at a 0.5% increment are shown by the dot-dashed line with a circle at each overlap threshold value. Arrows indicate the first point where expected value of each metric falls within the 95% CI of the control. The optimal overlap threshold is the lowest threshold value where the expected values of all three metrics fall within the 95% CI of the control. In this case, the optimal overlap threshold is 31.5%.

Table 4. Climatological Data for L96 System and Model. The 95% CI about the expected value for each statistic is taken as ± values in parenthesis

| | System | Model (Det.) | Model (Stoch.) |
|---|---|---|---|
| $X_{\max}$ | 14.76 (0.003) | 14.05 (0.004) | 14.78 (0.014) |
| $X_{\min}$ | -8.31 (0.011) | -7.01 (0.012) | -7.99 (0.011) |
| $\mu_X$ | 3.69 (0.004) | 3.76 (0.004) | 3.73 (0.004) |
| $\sigma_X$ | 4.54 (0.002) | 4.42 (0.002) | 4.43 (0.002) |
| $Y_{\max}$ | 2.46 (0.003) | – | – |
| $Y_{\min}$ | -1.8 (0.007) | – | – |
| $\mu_Y$ | 0.12 (0.001) | – | – |
| $\sigma_Y$ | 0.30 (0.001) | – | – |

Table 5. L96M EPS Error Statistics (bulk and variance) at each forecast lead time.

| tau | $ME_{\bar{e}}$ | | $\overline{\sigma^2_{\tilde{e}}}$ | | $MSE_{\bar{\tilde{e}}}$ | | $\sigma'$ | |
|---|---|---|---|---|---|---|---|---|
| | Bulk | Variance | Bulk | Variance | Bulk | Variance | Bulk | Variance |
| 0 | 0.0029 | 0.0036 | 0.0335 | 2.0E-05 | 0.0343 | 0.0004 | 0.9776 | 0.0045 |
| 0.2 | 0.0326 | 0.0122 | 0.1343 | 0.0010 | 0.1405 | 0.0104 | 0.9558 | 0.0119 |
| 0.4 | 0.0476 | 0.0274 | 0.3523 | 0.0228 | 0.3931 | 0.1391 | 0.8963 | 0.0324 |
| 0.6 | 0.0182 | 0.0559 | 0.7412 | 0.2273 | 0.9257 | 0.9761 | 0.8007 | 0.0542 |
| 0.8 | 0.0116 | 0.0962 | 1.3365 | 0.9026 | 1.7989 | 3.3177 | 0.7430 | 0.0588 |
| 1 | 0.0321 | 0.1532 | 2.0735 | 2.1645 | 2.9757 | 7.9045 | 0.6968 | 0.0545 |
| 1.2 | 0.0265 | 0.1957 | 2.8735 | 3.8459 | 4.3806 | 15.8611 | 0.6560 | 0.0459 |
| 1.4 | 0.0237 | 0.2388 | 3.6443 | 4.7197 | 5.4344 | 19.8098 | 0.6706 | 0.0379 |
| 1.6 | 0.0195 | 0.2644 | 4.3504 | 5.2949 | 6.4032 | 25.1928 | 0.6794 | 0.0314 |
| 1.8 | 0.0314 | 0.2855 | 4.9484 | 5.3256 | 7.2733 | 28.7206 | 0.6804 | 0.0249 |
| 2 | 0.0406 | 0.2944 | 5.4895 | 5.1879 | 8.0469 | 34.6424 | 0.6822 | 0.0200 |
| 2.2 | 0.0513 | 0.3064 | 6.0126 | 4.9067 | 8.7885 | 40.7473 | 0.6841 | 0.0159 |
| 2.4 | 0.0366 | 0.3167 | 6.4537 | 4.7358 | 9.3284 | 44.7655 | 0.6918 | 0.0136 |
| 2.6 | 0.0416 | 0.3128 | 6.8608 | 4.5559 | 9.8513 | 47.1874 | 0.6964 | 0.0115 |
| 2.8 | 0.0400 | 0.3197 | 7.2593 | 4.8418 | 10.5452 | 49.3660 | 0.6884 | 0.0105 |
| 3 | 0.0573 | 0.3164 | 7.6156 | 4.5110 | 10.9219 | 53.1869 | 0.6973 | 0.0091 |
| 3.2 | 0.0570 | 0.3272 | 7.9683 | 4.8239 | 11.4755 | 57.6767 | 0.6944 | 0.0089 |
| 3.4 | 0.0503 | 0.3319 | 8.2575 | 4.7032 | 11.6946 | 58.7195 | 0.7061 | 0.0084 |
| 3.6 | 0.0422 | 0.3167 | 8.5592 | 4.8423 | 12.1858 | 62.2317 | 0.7024 | 0.0080 |
| 3.8 | 0.0434 | 0.3189 | 8.8287 | 5.1559 | 12.6769 | 64.4172 | 0.6964 | 0.0078 |
| 4 | 0.0555 | 0.3177 | 9.0636 | 5.0917 | 12.9123 | 65.4157 | 0.7019 | 0.0074 |
| 4.2 | 0.0627 | 0.3266 | 9.2992 | 5.2461 | 13.3640 | 70.6218 | 0.6958 | 0.0071 |
| 4.4 | 0.0502 | 0.3306 | 9.5024 | 5.2769 | 13.4321 | 73.4724 | 0.7074 | 0.0071 |
| 4.6 | 0.0412 | 0.3218 | 9.7396 | 5.4897 | 13.8778 | 75.5714 | 0.7018 | 0.0069 |
| 4.8 | 0.0404 | 0.3116 | 9.9273 | 5.6479 | 14.3054 | 79.2468 | 0.6940 | 0.0067 |
| 5 | 0.0570 | 0.3210 | 10.1377 | 5.7393 | 14.5251 | 79.9435 | 0.6979 | 0.0066 |

Table 6. Decision rules tested for secondary criteria value. With the exception of the "Control," these decision rules are only applicable following a forecast false alarm when the current decision input advises taking protective action

| Name | Decision Rule |
|---|---|
| Control | User always follows the decision based on the control forecast probability, regardless of the consequences from the weather forecast. |
| Always | User will always reverse the decision. |
| Random | User losses confidence and decides randomly (fair coin toss) whether or not to reverse the decision. |
| Brash | User understands the concept of ambiguity. Instead of using an objective method to apply ambiguity to the forecast, the user applies their own 'rough estimate' of the ambiguity to the control forecast probability to avoid repeat false alarms. The 'rough estimate' used here is 5%, thus the user reverses the decision when $p_e^* - 5\% <$ C/L. |
| Overlap Conceptual Model | User employs the estimated ambiguity distribution to determine the overlap. The overlap value is compared to an overlap threshold determined from the conceptual model (discussed in the text). Overlap values greater than the threshold result in the user reversing the decision. |
| Optimal Overlap | User employs the full estimated ambiguity distribution to determine the overlap. The overlap value is compared to an empirically determined overlap threshold (discussed in the text). Overlap values greater than the threshold result in the user reversing the decision. |

Table 7. NCEP GEFS 21-member EPS error statistics used to determine calibration coefficients and $CES_L$ ambiguity distributions

| | $ME_{\bar{e}}$ | | $\sigma'$ | | $\overline{\sigma_{\tilde{e}}}$ | |
|---|---|---|---|---|---|---|
| | Average | Standard Deviation | Average | Standard Deviation | Average | Standard Deviation |
| Training, raw | -0.0319 | 0.767 | 0.596 | 0.139 | 1.78 | 0.392 |
| Training, calibrated | 0.0 | 0.767 | 1.0 | 0.228 | 2.92 | 0.641 |
| Application, raw | -0.156 | 0.915 | 0.620 | 0.175 | 2.00 | 0.569 |
| Application, calibrated | -0.124 | 0.915 | 1.015 | 0.287 | 3.27 | 0.931 |

Table 8. Partial ambiguity distributions from the NCEP GEFS 21-member EPS $CES_L$ tables for $p_e^* = 15\%$ using three different ensemble spread values. The table contains the relative frequency of sample $\hat{p}_T$ values within a 1% bin from 0% to 55%, where the upper bound of each bin is provided.

| Bin Maximum $\hat{p}_T$ | $\sigma_e = 2\ ^\circ C$ | $\sigma_e = 4\ ^\circ C$ | $\sigma_e = 8\ ^\circ C$ |
|---|---|---|---|
| 0.01 | 0.0099 | 0.0007 | 0.0001 |
| 0.02 | 0.0162 | 0.0032 | 0.0009 |
| 0.03 | 0.0210 | 0.0069 | 0.0024 |
| 0.04 | 0.0247 | 0.0112 | 0.0050 |
| 0.05 | 0.0274 | 0.0163 | 0.0101 |
| 0.06 | 0.0305 | 0.0233 | 0.0150 |
| 0.07 | 0.0330 | 0.0255 | 0.0221 |
| 0.08 | 0.0346 | 0.0326 | 0.0289 |
| 0.09 | 0.0354 | 0.0370 | 0.0373 |
| 0.1 | 0.0374 | 0.0425 | 0.0432 |
| 0.11 | 0.0379 | 0.0467 | 0.0502 |
| 0.12 | 0.0381 | 0.0492 | 0.0571 |
| 0.13 | 0.0413 | 0.0535 | 0.0630 |
| 0.14 | 0.0382 | 0.0540 | 0.0646 |
| 0.15 | 0.0373 | 0.0556 | 0.0675 |
| 0.16 | 0.0376 | 0.0539 | 0.0659 |
| 0.17 | 0.0376 | 0.0530 | 0.0629 |
| 0.18 | 0.0364 | 0.0533 | 0.0623 |
| 0.19 | 0.0340 | 0.0507 | 0.0566 |
| 0.2 | 0.0333 | 0.0481 | 0.0532 |
| 0.21 | 0.0322 | 0.0433 | 0.0471 |
| 0.22 | 0.0295 | 0.0401 | 0.0401 |
| 0.23 | 0.0282 | 0.0355 | 0.0340 |
| 0.24 | 0.0269 | 0.0319 | 0.0282 |
| 0.25 | 0.0248 | 0.0278 | 0.0222 |
| 0.26 | 0.0231 | 0.0232 | 0.0180 |
| 0.27 | 0.0216 | 0.0192 | 0.0132 |
| 0.28 | 0.0202 | 0.0161 | 0.0094 |
| 0.29 | 0.0194 | 0.0126 | 0.0072 |
| 0.3 | 0.0167 | 0.0097 | 0.0044 |
| 0.31 | 0.0144 | 0.0070 | 0.0032 |
| 0.32 | 0.0133 | 0.0054 | 0.0021 |
| 0.33 | 0.0118 | 0.0032 | 0.0011 |
| 0.34 | 0.0107 | 0.0028 | 0.0008 |
| 0.35 | 0.0092 | 0.0019 | 0.0005 |
| 0.36 | 0.0083 | 0.0012 | 0.0001 |
| 0.37 | 0.0078 | 0.0007 | 0.0001 |
| 0.38 | 0.0065 | 0.0005 | 0.00004 |
| 0.39 | 0.0054 | 0.0003 | 0 |
| 0.4 | 0.0044 | 0.0002 | 0.00002 |
| 0.41 | 0.0039 | 0.00004 | 0 |
| 0.42 | 0.0037 | 0.00004 | 0 |
| 0.43 | 0.0028 | 0 | 0 |
| 0.44 | 0.0023 | 0.00002 | 0 |
| 0.45 | 0.0019 | 0.00002 | 0 |
| 0.46 | 0.0014 | 0 | 0 |
| 0.47 | 0.0017 | 0 | 0 |
| 0.48 | 0.0012 | 0 | 0 |
| 0.49 | 0.0011 | 0 | 0 |
| 0.5 | 0.0006 | 0 | 0 |
| 0.51 | 0.0005 | 0 | 0 |
| 0.52 | 0.0007 | 0 | 0 |
| 0.53 | 0.0003 | 0 | 0 |
| 0.54 | 0.0003 | 0 | 0 |
| 0.55 | 0.0002 | 0 | 0 |

THIS PAGE INTENTIONALLY LEFT BLANK

# IV. RESULTS

This chapter presents the results obtained during this research regarding the three research objectives outlined in the Introduction.

## A. EVOLUTION OF AMBIGUITY

This section addresses the first research goal of understanding the behavior of ambiguity throughout the forecast. This goal is accomplished using the EoE. The EoE is our best estimate of ambiguity since it directly samples the inherent uncertainties in the IC and model perturbations and their sensitivities to a specific forecast scenario. The constituents' IC and model perturbations span the range of analysis and model errors, giving a distribution of plausible probability forecasts for a given EPS's sensitivity to errors in the ICs and in the model. The evolution studies were performed using 100 EoE forecast cases selected to span the L96M attractor, each with 100 constituents.

Our original hypothesis concerning the behavior of ambiguity regarded the magnitude of the variance of the random errors in the first two moments of the ensemble PDF as the primary influences on ambiguity. Specifically, the mean error of the ensemble mean ($ME_{\bar{e}}$) and the fractional error in ensemble spread ($\sigma'$) were considered. We hypothesized that increases (decreases) in ambiguity are directly related to the increases (decreases) in the variance of the random errors. Errors in the first moment play a larger role in creating errors in forecast probability, thus the variance of the $ME_{\bar{e}}$ dominates.

Using the large dataset of 24,000 ensemble forecasts from the L96M EPS, the variance in $ME_{\bar{e}}$ and $\sigma'$ were diagnosed following bulk calibration of the data to remove systematic error. The evolution of the variance in these two error characteristics is shown in Figure 48 (a) and (b). Early in the forecast (before $\tau = 0.6$), error variance is low as all ensemble members likely exhibit similar high skill. Maximum dispersion in the ensemble forecasts occurs on average between $\tau = 0.6$ and $\tau = 2.0$ (Chapter III.A.4).

During this period, the variance in the $ME_{\bar{e}}$ error distribution increases. As ensemble spread increases, the possible error in the ensemble mean increases since the verification may fall farther from the center of the forecast PDF, thus the variance in the $ME_{\bar{e}}$ also increases. The variance in the fractional error in ensemble spread increases as dispersion ramps up, but quickly peaks and begins a gradual decrease to and below its original level. At early lead times, the skill of all ensemble members is likely high, resulting in consistently low ensemble spread and similar fractional error values between forecasts. As error growth begins to ramp up, some members experience faster error growth than others due to sensitivities to the location in the attractor or deficiencies in the EPS. At this point, fractional error between constituents can vary greatly, and it ascends to its maximum variance. Eventually, high error growth occurs in all members resulting in similarly high ensemble spread for all constituents, which reduces the variation in fractional error values. Following maximum dispersion, the variance in $ME_{\bar{e}}$ levels off but remains high due to the large spread in the ensemble PDF. The variance in fractional error continues to decrease and asymptotes towards zero as the ensemble spread similarly saturates among all constituents.

Employing the initial hypothesis regarding the behavior of ambiguity, we expected the following evolution. Early in the forecast prior to maximum ensemble dispersion, ambiguity should be relatively low since both error variances are low. As the forecast moves into the time of maximum dispersion, ambiguity should rapidly increase to a maximum following the increase in variance of both errors. However, following maximum dispersion, ambiguity was expected to decrease and asymptote to zero as the forecast PDF saturates towards climatology, resulting in no uncertainty in the PDF.

Using the 100 EoE forecast cases, we determined the average total ambiguity [Equation (26), page 61] as a function of forecast lead time. Figure 49 shows the average total ambiguity for the EoE ambiguity distributions for $p_e^*$ values of 5%, 50%, and 95%, for comparison. The behavior of ambiguity shown does not follow our initial hypothesis. Rather than peaking during the height of error growth, ambiguity maximized early in the forecast period then decreased quickly during peak error growth. Late lead time

ambiguity did behave as hypothesized except it asymptoted to a minimum value and not zero. Evidently, the initial hypothesis is in need of revision.

To explore this behavior in detail, it was beneficial to look at the evolution of ambiguity for a single EoE forecast case. We plotted each of the 100 constituents' forecast PDFs for an arbitrary $X_k$ variable using a normal fit to the 21 members in each constituent's ensemble forecast for forecast lead times $\tau = 0.2$ through $\tau = 5.0$ at an interval of 0.2 time units (Appendix). Each figure also displays a histogram of the EoE $\hat{p}_T$ values at each lead time, where the expected value of each distribution is 50%. These figures display a time sequence where ambiguity starts out high and decreases with increasing lead time with some fluctuation around maximum ensemble dispersion.

For a deeper understanding, we looked more closely at $\tau = 0.2$ (i.e., a high ambiguity time) and $\tau = 4.8$ (i.e., a low ambiguity time) in Figure 50 (a) and (b). For analysis of forecast probability in Figure 50(a), the event threshold resulting in $E\left(\hat{p}_T\right) = 50\%$ is $X = -1.72$. A wide range of forecast probability values are possible using this threshold with each constituent individually. The calculated range of values spans from 1% to 98%, with total ambiguity from 7% to 92% (85%). The total ambiguity compares well with the average value shown in Figure 49 for $p_e^* = 50\%$, although the width is slightly larger than the average for this particular EoE forecast case and variable. Looking at the later lead time in Figure 50(b) and using a different event threshold ($X = 2$) that again gives $E\left(\hat{p}_T\right) = 50\%$, the range of constituent forecast probability values is much smaller, spanning 26% to 78% with total ambiguity of 34% (33% to 67%). Again, this is consistent with the evolution shown in Figure 49, where ambiguity decreases for later lead times.

From this analysis, the primary influences on the size of the EoE ambiguity distributions appear to be how much variation is present between the locations of the constituents' PDFs and the uncertainty (i.e., ensemble spread) of the constituents' PDFs. Early in the forecast [Figure 50(a)], the typical spread of each constituent is still quite low with an average spread of 0.371. The standard deviation of the constituents' means is

equal to 0.224, which is comparable in size. For a centrally located decision threshold like the one chosen here, the constituent PDFs will be dispersed on either side of the threshold, but since the variation in PDF location is large or comparable to the average spread of the constituent forecasts, the constituent PDFs will cross the threshold to varying degrees giving a wide range of forecast probability values.

Playing the same game with the constituents at the later lead time [Figure 50(b)], we see that the standard deviation of the constituents' means is now equal to 0.737, which has increased. As expected, the average constituent spread has increased and now equals 3.81, which proportionally is a much greater change than was seen in the increase in variation of constituent locations (~200% increase versus ~900% increase, respectively). For a given event threshold, the percentile location of the threshold within each constituent PDF is now much more alike leading to similar, albeit slightly different, forecast probability values from each constituent. Thus as the typical spread of the constituent PDFs increased without a proportional increase in the variation in PDF location, the ambiguity associated with the forecast decreased.

Figure 51 illustrates the sensitivity of forecast probability to PDF spread and shifts in PDF location, where a low spread (thick solid) and high spread (dot-dash) PDF are shifted from a mean position of 0.75 to -0.25 while maintaining the same spread. In Figure 51(a), the probability of preceding the event threshold (thin solid) for the low and high spread PDFs is 15.9% and 35.4%, respectively. Following the shift in location in Figure 51(b), the low spread probability is 63.1%, which is a change of 47.2%. The high spread probability is 55%, giving a change of 19.6%. The location shift resulted in a larger displacement of probability density relative to the event threshold for the low spread PDF. Thus shifts (or errors) in location are likely to produce a wider ambiguity distribution when ensemble spread is low.

The same concept applies to event thresholds that are not centrally located. In this case though, the forecast probability values for many of the constituent PDFs will become more certain (i.e., closer to 0% or 100%) leading to a relatively tighter more skewed ambiguity distribution. For example, an event threshold of $X = -4$ in Figure 50(a), leads to virtually no ambiguity as all constituent PDFs fall above the threshold.

For this research, we made comparisons of ambiguity distributions using event thresholds that gave the $E\left(\hat{p}_T\right)$ of the constituents equal to specific $p_e^*$ values being tested (Figure 37, page 108), thus event thresholds used always fell amongst the constituent PDFs.

The relationship between the variability in constituent PDF location and the PDF variance is a major influence on ambiguity. To explain this, we found the variance of the constituents' means and the average variance of the constituents for each of the 100 EoE forecast cases. These measures are combined over the datasets in order to compare the average variance between constituent means and the average constituent variance at each lead time, shown in Figure 52. When ambiguity is high, the average variance of the constituent means is comparable in magnitude to the average constituent variance. As forecast lead time increases, there is an increase in both metrics, but the rate of increase in each is not proportional. The average constituent variance increases at a much faster rate leading to a decrease in ambiguity with increasing lead time.

Using Figure 53, we compare the variance information found in Figure 52, the ratio of these two variance values, and the average changes in the EoE total ambiguity for different $p_e^*$ values (same as Figure 49). The variance ratio is computed as the average variance in constituent location over the average constituent variance. At the beginning of the forecast period, the variance ratio is high indicating that the variation in the location of the constituent PDFs is nearly as large as the typical constituent spread. During maximum dispersion, there is a rapid increase in the average constituent variance (0.137 to 7.28 for an over 5000% increase), while the variance of constituents' means increases much less (0.0505 to 0.397 for an increase of less than 700%), resulting in a rapid drop in the variance ratio. This period is accompanied by a 40%-50% decrease in total ambiguity for the $p_e^*$ values shown. Following maximum dispersion, the ratio asymptotes to a minimum value (~0.0370) as the average constituent variance continues to gradually increase. At this time, total ambiguity asymptotes to a minimum value as well, but not to zero. The variance of the constituents' means still results in ambiguity

even though constituent variance saturates towards climatology, since shifts in the constituents' PDF locations play a large role in changing the forecast probability value associated with each constituent.

This analysis furthers our understanding of the relationship between forecast uncertainty and ambiguity, while solidifying the relationship between ambiguity and random errors in the ensemble PDF due to unaccounted for sources of uncertainty. We determined that ambiguity is closely linked to both forecast uncertainty (i.e., first-order uncertainty) and the sensitivity of the EPS to deficient IC and model perturbations, in that the interaction of these two factors controls the magnitude of the total ambiguity associated with a given forecast situation for a particular EPS.

Recall that EoE is an impractical approach to estimating ambiguity due to the large computational expense. Can ambiguity be estimated without the EoE using statistical characteristics of the EPS's ensemble forecasts? Looking at Figure 52 and Figure 11 (page 87), the average variance of the EoE constituents is the same as the average variance of the L96M ensemble forecasts taken over the large forecast dataset. As a proxy for the variance of constituents' means, the variance of the mean error in the ensemble mean ($ME_{\bar{e}}$) found using the L96M ensemble forecasts may be used. This relationship is shown in Figure 54 (similar to Figure 53). The time evolution of average ensemble forecast variance and variance of $ME_{\bar{e}}$ follow the same behavior as seen in Figure 53. The variance ratio (taken as $ME_{\bar{e}}$ variance over average ensemble variance) indicates a similar behavior as well, but note the greatly reduced ratio value early in the forecast when using the EPS error statistics. A comparison of the ratio values from Figure 53(b) and Figure 54(b) is shown in Figure 55. Since the average variance of the ensemble forecasts and the EoE constituents are the same, any difference between the ratio values must be due to the variance in $ME_{\bar{e}}$. In this case, the variance in $ME_{\bar{e}}$ is not large enough to accurately simulate the variation in possible ensemble PDF locations found using the EoE (i.e., possible realizations of the ensemble PDF given limitations in the EPS perturbations). Thus, ambiguity estimates obtained using the EPS error characteristics may be greatly underestimated. After maximum dispersion, the ratio nears

a value of one, indicating ambiguity estimates created using the EPS error characteristics may improve. This problem may be due to the sub-setting used to arrive at the variance in $ME_{\bar{e}}$ (see Chapter III.B.3). Attempts to use the variance of errors in the ensemble mean without sub-setting (i.e., finding the variance of the individual ensemble mean error values without averaging) gave an extreme over-estimate of ambiguity at all lead times, since the variance of possible error in the ensemble mean value is larger than the average variance of the ensemble forecasts at all lead times.

## B. VALIDATION OF AMBIGUITY ESTIMATES

The discussion of $CES_G$ and RCR ambiguity estimate validation in this section refers primarily to the series of comparisons shown in Figure 56 and Figure 57. Each panel in Figure 56 shows comparisons across all forecast lead times for a specific $p_e^*$ value. Alternately, each panel in Figure 57 provides comparisons across all tested $p_e^*$ values for a certain forecast lead time. In both figures, the set $p_e^*$ value or forecast lead time used to create each individual panel is displayed at the top of the panel. All comparisons show the difference in total ambiguity [Equation (26), page 61] of both the $CES_G$ and RCR ambiguity distributions compared to the EoE ambiguity distribution, where a negative difference indicates the $CES_G$ or RCR ambiguity distribution is too narrow compared to EoE. This validation strategy gave us a look at how well the practical ambiguity estimation techniques simulate the variance of our best estimate of the ambiguity distribution.

From Figure 56 and Figure 57, we see that the ambiguity distributions from the practical estimation techniques appeared to perform very poorly at early in the forecast with total ambiguity differences near 30%, but each showed improvement with increased forecast lead time. Although this feature may appear to be tied to forecast lead time, it is actually tied to the ensemble variance, which plays a significant role in the production of ambiguity.

Figure 58 shows that the $CES_G$ and RCR ambiguity distributions generally followed the same evolution as the EoE estimate, where all of the estimates had relatively large ambiguity at early times that decreased with time. The exception may be RCR, where the total ambiguity began to increase again following a period of decrease. From Figure 11 (page 87), as expected, we see that ensemble variance increased on average with increasing lead time. This can also be seen in Figure 59 (a) and (b), where the number of uncertain forecasts (i.e., forecasts with $p_e^*$ between 0.1% and 99.9%) increased with time, indicating that fewer forecasts existed where the event threshold fell outside of the forecast PDF. Even though, the ratio in Figure 55 shows that the variance of the $ME_{\bar{e}}$ error distribution was highly underdone early in the forecast (by a factor of six), the $CES_G$ and RCR distributions still exhibited maximum ambiguity early in the forecast (Figure 58), as a result of the generally low ensemble variance.

This analysis suggests that ambiguity will evolve from high values to low values on average as a result of the typical increase in ensemble spread with time. Of course, it is possible on a case-by-case basis for an ensemble forecast to exhibit small spread at any lead time resulting in large ambiguity associated with the forecast probability. Thus total ambiguity is not necessarily a function of forecast lead time, but rather depends strongly on the spread of the current ensemble forecast at the lead time in question.

From the panels in Figure 57, we see that the largest differences in total ambiguity occurred with mid-range forecast probability values. In the first panel ($\tau = 0.2$), the difference in total ambiguity for $p_e^* = 50\%$ was almost 30%, while the differences for both $p_e^* = 1\%$ and $p_e^* = 99\%$ were between 4%-7%. Thus it may appear the $CES_G$ and RCR estimates performed better for extreme forecast probability values. The apparent disparity in performance is simply a result of the lower and upper bounds (i.e., 0% and 100%, respectively) confining the range of possible forecast probability values. In general, we expect to see tighter ambiguity distributions for the extreme forecast probability values. Consider a single set of 100 EoE constituents. An event threshold that results in an expected value of 1% for the EoE ambiguity distribution (using probability of exceeding) will likely fall above (i.e., to the right) of many of the

constituent's PDFs resulting in forecast probabilities very close to or equal to 0% for those constituents. In this case, the ambiguity distribution is tighter since it is bounded on the low end, where many near 0% forecast probabilities accumulate. An event threshold giving an expected value of 50% for the ambiguity distribution for the same EoE forecast case produces a wider ambiguity distribution since the placement of the threshold is centrally located among the constituent PDFs allowing forecast probability values to spread evenly on either side of 50%.

Therefore, problems with the variance of $CES_G$ or RCR ambiguity distributions were seen on both sides of the distributions as shown in Figure 60. The figure shows EoE and $CES_G$ ambiguity distributions computed from a single EoE forecast case for the same variable, both centered on $p_e^*$ = 50% (in accordance with the validation method) at $\tau = 5$. The $CES_G$ distribution was too narrow (20% versus 32% total ambiguity), and the total ambiguity difference when compared to the EoE distribution was equivalent on either side at 6%. In contrast, Figure 61 shows the EoE and $CES_G$ ambiguity distributions for $p_e^*$ = 5% using the same EoE forecast case and variable at the same forecast time. Here, the differences between the distributions were chiefly present in the direction of higher forecast probability values. Both of the distributions are bounded by 0% on the low side, which resulted in a similar value (approximately 2%) for the lower bound of the 90% CI for each estimate. Thus the difference in total ambiguity was essentially one-sided, where the upper bounds are 9% and 12% for CES and EoE, respectively. Although we may be encouraged by the results for extreme forecast probability values, it is important to understand that this improvement is in part an artificial result.

We found the $CES_G$ total ambiguity to be too narrow in relation to the aggregated EoE ambiguity distributions regardless of forecast lead time or forecast probability value tested. A leading contributor to this problem was the creation of $CES_G$ ambiguity distributions using random draws from the distribution of average ensemble variance, thus ensemble variance was independent of the forecast situation. Therefore, the typical

ensemble variance used to compute the $\hat{p}_T$ values was near the average, which in many cases would likely be too high compared to the flow-dependent variance. Since larger ensemble variance produces a more narrow ambiguity distribution, the configuration of $\mathrm{CES_G}$ will likely result in a consistent underestimation of the total ambiguity. It is likely that the flow-dependent $\mathrm{CES_L}$ estimates would alleviate much of this problem, but recall that this technique was unavailable when the validation study was performed.

Additionally, from Figure 55, the variance of the $ME_{\bar{e}}$ error distribution used to develop the $\mathrm{CES_G}$ ambiguity distributions was not wide enough to adequately simulate the variance in forecast PDF location typically found using the EoE constituent forecasts. This deficiency was particularly severe at the early forecast lead times prior to maximum dispersion, where the variance of the $ME_{\bar{e}}$ distribution was as much as six times lower. So, even if the ensemble variance was correctly simulated, the $\mathrm{CES_G}$ sample forecast distributions would not be sufficiently separated to produce a wide enough ambiguity distribution. Thus early in the forecast, the combined problems of using forecast-independent ensemble variance and largely underdone $ME_{\bar{e}}$ variance resulted in large differences in total ambiguity, where the deficiency in the $ME_{\bar{e}}$ variance was likely the dominant factor.

$ME_{\bar{e}}$ variance improved to less than a factor of two difference later in the forecast following maximum dispersion, performing best towards the end of the period of maximum dispersion (ratio value was approximately 1.35). At this point, we found the best performance in $\mathrm{CES_G}$ total ambiguity, but the total ambiguity was still too small, likely because the $ME_{\bar{e}}$ variance was slightly too low and because we did not account for flow-dependent ensemble variance. Following maximum dispersion, the ratio value began to increase slightly indicating that the $ME_{\bar{e}}$ variance was performing worse, but the slow increase did not continue beyond five time units, and the ratio value never increased past 1.7. The slow deterioration of the $ME_{\bar{e}}$ variance and the continued

increase in average ensemble variance over this period resulted in narrowing of the $CES_G$ ambiguity distributions and a slow increase in total ambiguity difference, which tapered off near five time units.

We attempted to use the ratio value from Figure 55 to improve $CES_G$ total ambiguity by increasing (i.e., correcting) the $ME_{\bar{e}}$ variance at each lead time by its respective ratio value. Results showed improved total ambiguity at all lead times, but the correction factor caused overcorrection early in the forecast and was still too small later on (example shown in Figure 62 for $p_e^* = 50\%$). The $CES_G$ estimates were likely still degraded due to the lack flow-dependence, thus this line of research was not pursued further.

From Figure 56 and Figure 57, the RCR total ambiguity was too narrow during the early forecast lead times, but then transitioned to become slightly too wide later in the forecast for most of the $p_e^*$ values tested. Since the RCR distributions are flow-dependent, we find more evidence that the highly deficient variance of the $ME_{\bar{e}}$ error distribution early in the forecast played a significant role in degrading the $CES_G$ and RCR ambiguity distributions. During this timeframe, the RCR PDFs could not adequately separate to generate sufficient ambiguity compared to the EoE because of the poor $ME_{\bar{e}}$ variance.

As the performance of the $ME_{\bar{e}}$ error distribution began to recover, the total ambiguity difference for RCR, like $CES_G$, improved as well. Unlike $CES_G$, the RCR estimates showed continued improvement beyond maximum dispersion, eventually becoming too wide, but by no more than 3% compared to EoE. As forecast error growth increased, the average variance of each of the EoE constituents followed, thus decreasing the width of the EoE ambiguity distributions. The RCR ambiguity estimate used only the first constituent of a given EoE forecast case, where the variance of the constituent's forecast PDF was varied for each resample based on random draws from the $\sigma'$ error distribution. In general, the variance of any resampled PDF would be similar to the

average constituent variance, but over 10,000 resamples, there were likely many fractional error draws resulting in a relatively more narrow PDF (compared to the average EoE constituent variance), which inevitably produced a wider ambiguity distribution. Thus the total ambiguity difference between RCR and EoE switched from negative to positive values for later forecast lead times.

In the previous discussion of the $CES_G$ and RCR ambiguity distributions, we have not yet made judgments about the validity of the estimates. In order to make a judgment, we first consider the use of EoE ambiguity distributions as the standard. EoE provides a flow-dependent ambiguity estimate that accounts for finite ensemble size and samples the sensitivity of the probability forecast to deficient analysis and model perturbations in the EPS. Analogous to the single ensemble forecast providing the best guess for uncertainty in the deterministic forecast, the EoE gives us our best-guess estimate of the uncertainty in the ensemble forecast. However, EoE suffers from the same basic limitations as an EPS. Limited sampling due to the finite number of constituents results in random error in the EoE ambiguity distribution. Also, any incomplete perturbations (simulating EPS deficiencies) in the EoE design will result in systematic underestimation of ambiguity.

It is obvious that deficiencies exist on average in both $CES_G$ and RCR, especially early in the forecast when ambiguity is the highest (i.e., when ensemble variance is typically low). The total ambiguity estimates from $CES_G$ and RCR improve with time and draw fairly close to the EoE value (generally $< 10\%$ and $< 5\%$ difference for $CES_G$ and RCR, respectively) during the timeframe of highest error growth rate between $\tau = 0.8$ and $\tau = 3.4$.

From basic chaos and ensemble forecasting theory, we understand that nonlinear error growth limits predictability making the ensemble forecast the best source of forecast information in general. However, considering only the deterministic NWP forecast may still be appropriate early in the forecast period while average error is below about 10% of the climatological variance $\left(\sigma_C^2\right)$ (i.e., the deterministic realm), which on average occurs

between $\tau = 0.6$ and $\tau = 1.0$ for the L96M EPS. In Figure 59, prior to maximum error growth, the frequency of uncertain forecasts is low due to the low ensemble variance found in the early forecast period. Therefore, at early forecast lead times when the $CES_G$ and RCR ambiguity estimates are performing at their worst, their deficiencies are not critical since forecast uncertainty is not prevalent (i.e., ambiguity is not or rarely needed).

The rate of error growth is dependent on the scale of the forecasted phenomenon, where faster error growth is generally observed for smaller scale phenomena. This scale-dependency impacts the ambiguity distributions in the same fashion. The EPS error distributions and the ensemble spread statistics are also phenomena-dependent, meaning that the ambiguity distributions are also tied to the forecast error growth. So, regardless of the forecast variable, total ambiguity estimates from $CES_G$ and RCR will evolve from high to low values but on different variable- or scale-dependent time scales, producing reasonably accurate estimates of total ambiguity past the initial deterministic realm. Therefore, we conclude that $CES_G$ and RCR ambiguity distributions are likely good enough to provide valuable information to the decision process.

This conclusion should be tempered to apply to situations where the expected values of the $CES_G$ or RCR ambiguity distributions are equal to or near the expected value of the EoE ambiguity distribution, per our experiment design. In general, the calibrated forecast probability is merely a random sample from the EoE ambiguity distribution, thus it may fall anywhere within the distribution. Since the calibrated forecast probability is also the expected value of the $CES_G$ and RCR ambiguity distributions, the estimated distributions are often not collocated with the EoE ambiguity distribution. This issue is discussed in detail in the next section.

In the following sections, we discuss the results of value studies that incorporated the CES and/or RCR ambiguity information into the decision making process. For these studies, the question of ambiguity estimate validity becomes a question of whether or not the ambiguity information adds value. For example, even if we show that the difference in total ambiguity between RCR and EoE is large for some forecast situation, the RCR

ambiguity information may still positively influence the decision making process over the long-term and add value, while on a case-by-case basis the results will vary due to deficiencies in the estimation process.

## C. VALUE USING UNCERTAINTY-FOLDING

In this section, we assess improvements to value from the uncertainty-folding technique. Recall that we used two separate event thresholds designed to represent a common and a rare event. Uncertainty folding was performed using ambiguity distributions from the EoE, $CES_G$ and RCR estimation techniques to provide $p_a$ values for each method. In addition, a grand ensemble was tested where all constituent members for a single EoE forecast case were combined to form a large ensemble giving a single forecast probability value ( $p_g$ ). These four decision input sources were compared relative to the value provided by basing decisions on the control ensemble forecast probability alone. The control ensemble forecast was taken as the first constituent of each EoE forecast case. Significance of the results in this section was assessed using the 95% CI for the results produced by resampling.

To check if the L96M EPS control forecast was behaving well with respect to value, we first verified that its forecast probability was outperforming the deterministic forecast. If not, the deterministic forecast would be more appropriate to use in decision making, and ambiguity about the forecast probability is irrelevant. If the control forecast probability does add value compared to the deterministic forecast, then our uncertainty-folding results will show if any additional value can be added by incorporating the ambiguity information. For this comparison, we computed the integrated optimal *VS* [*IOVS*, Equation (27), page 72] for both common and rare event thresholds for the deterministic and control ensemble forecasts, displayed in Figure 63 (a) and (b), respectively. The deterministic forecast was taken as the first member of the first constituent in each EoE forecast case. In both figures, we see that the control ensemble forecast provided significantly better value than the deterministic forecast, except at very early forecast lead times when the deterministic skill was still high.

For both events, the relative *IOVS* found using the EoE $p_a$ values and the grand ensemble's $p_g$ values was generally greater than one throughout the forecast as shown in Figure 64 (a) and (b), indicating that these two sources provided additional value compared to the control forecast. We found the improvement to be significant past $\tau = 1.4$ for the common event. For the rare event, the improvement was only significant at sporadic lead times.

At most lead times, the grand ensemble appeared to provide slightly better value than the EoE data. The scores for these two methods started close to one and then increased during the time of maximum dispersion. At the beginning of the forecast, the skill associated with the control ensemble forecast was still quite high, thus it was difficult for the grand ensemble or EoE to improve on the value attained by the control. As the forecast dispersion and error growth ramped up, the skill of the control ensemble decreased, and the grand ensemble and EoE were able to provide greater value due to the additional information available in each method.

Each grand ensemble was a collection of 2,100 ensemble members where IC and model perturbations were varied within the range of uncertainty. Thus the grand ensemble accounted for deficiencies in the modeling system much more thoroughly than a single 21-member ensemble forecast. The EoE ambiguity distribution was able to provide additional value for the same reason, since it incorporated each constituent's simulation of uncertainty in the EPS perturbations. The grand ensemble appeared to marginally outperform the EoE (although not significantly) since information may have been lost during the conversion of each EoE constituent to a single $\hat{p}_T$ value. Prior to computing $\hat{p}_T$, each constituent ensemble forecast contained information regarding the current first-order uncertainty (i.e. spread), as well as higher-order moments of the forecast PDF. This information was lost when a single $\hat{p}_T$ value was used to estimate the event uncertainty, and then combined with the other 99 estimates. The grand ensemble on the other hand retained all information when making its single estimation of the event uncertainty.

The scores found using uncertainty-folding with the $\text{CES}_\text{G}$ and RCR $p_a$ values were generally not significantly different than one throughout the forecast for both the common and rare event, indicating that they performed on par with the control forecast. In the validation section, we saw that both techniques did a reasonable job of estimating ambiguity. Their lack of value here can be explained by considering how their ambiguity distributions were produced. Both techniques' ambiguity distributions were centered on the control ensemble forecast's (i.e., the first constituent from an EoE forecast case) $p_e^*$ value. Based on the uncertainty-folding computation, both the $\text{CES}_\text{G}$ and RCR $p_a$ values should remain close to $p_e^*$. Thus the value attained using the practical ambiguity estimates is unlikely to be significantly different from that of the control ensemble forecast.

Additionally, $p_e^*$ is a random sample from the EoE ambiguity distribution for a certain forecast case, thus it could fall anywhere within the EoE ambiguity distribution. We performed validation by artificially locating event thresholds where the expected value of the EoE ambiguity distribution was equal to $p_e^*$, thus collocating the $\text{CES}_\text{G}$ and RCR ambiguity distributions with the EoE ambiguity distribution. Therefore, validation only provided a measure of how well the estimation techniques matched with respect to the variance of their respective ambiguity distributions.

Figure 65 shows a situation where $p_e^*$ was collocated with the expected value of the EoE distribution using a single EoE forecast case at $\tau = 4$, where the RCR ambiguity distribution was shown to provide a reasonably good ambiguity estimate. The 100 EoE constituent $\hat{p}_T$ values were histogrammed using class interval of 1%. For clarity in the figure, the 10,000 RCR $\hat{p}_T$ values were fit using a beta distribution. Although the beta-fit does not always provide a quality fit to the $\hat{p}_T$ data, it was sufficient for the pedagogical purpose here. For this forecast case, the total ambiguity of the EoE and RCR distributions appeared to match well as was expected (90% CI widths for EoE and RCR are 26% and 31%, respectively). Since the RCR distribution was collocated with the EoE

distribution, both estimations gave similar $p_a$ values (78.2% and 78.8%) when used with uncertainty-folding. In Figure 66, we show a different case at the same forecast lead time where the RCR $p_e^*$ value occurred in the upper tail of the EoE ambiguity distribution. The total ambiguity of the two distributions was still relatively close (27% and 34%), but there is a large difference between the $p_a$ values (19.5% and 36.4%).

From this analysis, we see that while the practical estimation techniques were fairly effective at simulating the variance of the ambiguity distribution, differences should typically exist between the EoE and the $CES_G$ and RCR $p_a$ values since $p_e^*$ is a random sample within the EoE ambiguity distribution. These differences produce errors when using the estimates to compute a single decision input combining the first- and second-order uncertainty, reducing the value of the decision input in normative decision making. On the other hand, the theoretical and impractical EoE ambiguity estimate was able to add significant value to the decision making process, since its $p_a$ value is not tied to the control forecast probability. Additionally, while each of the estimation methods produces consistent estimates of the ambiguity, EoE provides a sharper distribution eliminating bogus $\hat{p}_T$ possibilities, resulting in a better $p_a$ value.

## D. VALUE USING SECONDARY CRITERIA

This section describes our experiments using the ambiguity information to add value to the decision making process when considering the secondary criteria of repeat false alarms. Our goal was to use the ambiguity information to significantly reduce the number of repeat false alarms while maintaining the primary value (measured by optimal *VS*, as well as *POD* and *POMD*) associated with normative decision making within the *C/L* scenario. To alter the secondary criteria (i.e., reduce repeat false alarms), a user was allowed to reverse the current decision of taking protective action if and only if a false alarm had just occurred at the same location. We compared the primary value and secondary criteria results for various possible user decision rules (Table 6, page 120) to evaluate the effectiveness of each. The experiment was performed using real-world

forecast data as described in Chapter III.F. We used the 95% CI found through resampling to assess the significance of results among user.

Prior to exploring the secondary criteria, we first evaluated the performance of GEFS in relation to the deterministic forecast (member #1). As described previously, forecast error growth was still quite low and barely out of the deterministic realm at 120-hours (Chapter III.F.4), thus we needed to determine if GEFS was adding value compared to the deterministic forecast at this time. From Figure 67, the deterministic forecast provided value for a large range of *C/L* (10% to 85%), but GEFS added significant value over the deterministic forecast, plus, it provided value over a greater range of *C/L* (1% to 91%). Thus it made sense to use the ensemble forecasts since we were at a lead time where the ensemble was adding significant value over the deterministic forecast. We were primarily concerned with the value in secondary criteria that could be added to users with low *C/L* since they experience frequent false alarms. At low *C/L*, the opportunities for false alarms are numerous since there are many forecasts directing the user to protect (have low $p_e^*$ and result in a non-occurrence of the event). Alternately, high *C/L* users generally see fewer false alarms so may be less concerned with their repeats.

The number of repeat false alarms found following GEFS with each *C/L* is shown in Figure 68. As expected, there were a large number of repeat false alarms for the extremely low *C/L* values, because there were many forecasts that required the user to protect. As the *C/L* increased, fewer false alarm opportunities were available. The fastest rate of decrease in the number of repeat false alarms occurred between the *C/L* 1% and 5%. From Figure 44(b) (page 114), approximately 40% of all forecast probability values from the 50,220 forecasts fell within the 0%-5% bin. Accordingly, once the *C/L* increased beyond 5%, a large portion of the forecast opportunities would direct the user to take not protect, greatly reducing the overall number of false alarm opportunities. The rate of decrease slowed as *C/L* increased, but the number of repeat false alarms never reached zero, even for the highest *C/L* of 99%. Since ensemble spread was still relatively low at the forecast lead time, many of the control probability forecasts were close to 0% and 100% (i.e., forecast *res* was high). Approximately 23% of the forecasts fell within

the 95%-100% bin, as seen in Figure 44(b). Thus due to the large number of high probability forecasts, there were still false alarm and repeat false alarm opportunities, even for the highest *C/L* values.

The *VS* results for the *always* user (Table 6, page 120) and the control user are compared in Figure 69. Obviously, the number of repeat false alarms for the always user was zero at all *C/L* (i.e., 100% reduction), and the change was significant. However, choosing to always avoid repeat false alarms severely degraded the primary *VS*, because many of the reversals resulted in additional misses (e.g., for *C/L* 1%, total misses were increased from 35 to 1767). Trading false alarms for misses can severely degrade the *VS* for low *C/L* users, due to the large change in expense ($C \ll L$). Thus it would take many correct reversals (i.e., false alarm to correct rejection) to account for one incorrect reversal (i.e., hit to miss) (Table 3, page 32). Making an incorrect reversal will have a much smaller effect on the *VS* for high *C/L* users, since $C \approx L$ and the total expense will not be increased greatly. Therefore, changes to the *VS* will typically be insignificant for high *C/L* users following the always decision rule, as seen in Figure 69.

The *VS* for the always user was significantly reduced compared to the control over the *C/L* range 1% to 70%. Beyond *C/L* 70%, the difference in *VS* was not statistically significant, but the change in our other primary value metrics (*POD* and *POMD*) was significant through *C/L* 90% (e.g., *POD* shown in Figure 70). For *C/L* greater than 90%, there was no significant difference between the control user and the always user, but at these *C/L*, false alarms are typically not a concern (as discussed above). We found that this user provided the most significant reduction in our secondary criterion, but also the greatest degradation in primary value.

Results for the *random* decision rule are shown in Figure 71 and Figure 72. This uninformed user who based the decision to reverse his protective action on a coin toss was also able to significantly reduce repeat false alarms for all *C/L*. However, the primary value metrics indicated that the random user's decision strategy was also significantly reducing the primary value. Specifically, the *VS* was significantly lower over the *C/L* range 1% to 59%, while the performance based on *POD* and *POMD* was significantly different through *C/L* 80% (e.g., Figure 73).

The percent reduction in repeat false alarms at each *C/L* was approximately 60% (Figure 74), which was greater than our anticipated amount of 50% (i.e., over many cases the option to change should occur in approximately half of the opportunities). This was an indication that the decision rule was breaking up series of repeat false alarms. For example, consider a specific grid point that had three false alarms in a row, resulting in two repeat false alarms events counted at that point. If the random user reversed the decision for the second false alarm, then both of the repeat false alarm events would be eliminated.

The *brash* user, who applied a fixed decrease (i.e., 5%) to the control forecast probability to mitigate repeat false alarms, was surprisingly able to achieve primary value scores similar to those associated with the control user (Figure 75) at all *C/L*. Furthermore, the brash user significantly reduced the number of repeat false alarms for two *C/L* ranges, 1% to 9% and 95% to 99% (Figure 76). The percent reduction from Figure 74 for the lower *C/L* range decreases from 32% to approximately 11%. The reduction then fluctuated between 5% and 10% for mid-range *C/L* before dramatically increasing once again for *C/L* above 90%. The larger reductions for the very low *C/L* values were mainly due to the large proportion of forecasts (~43%) found between 0% and 10%, which resulted in more chances to reverse the decision. For the second range of *C/L* (95% to 99%), the percent reduction was 100% (Figure 74). Since the brash user always decreased the forecast probability by 5% for repeat false alarm opportunities, there were no repeat false alarms for *C/L* > 95% (i.e., forecast probabilities greater than 95% were always reduced to 95% or less), which mimicked the always decision rule.

If we increased the brash user's arbitrary percent decrease to forecast probability, we would see wider ranges of significantly higher percent reduction at both *C/L* extremes due to the same effects described above. However, the brash user's primary value would be significantly reduced for the extreme low *C/L* if the arbitrary reduction is too large. In other words, increasing the brash user's percent decrease takes him closer to behaving like the always user, who clearly failed to maintain primary value.

We now turn to the decision rules where the estimated $CES_L$ ambiguity distribution was employed to reduce to the number of repeat false alarms. The decision to reverse the protective action for repeat false alarm opportunities used a variable overlap threshold (function of *C/L*). If the overlap exceeded the threshold, the decision was reversed and no protective action was taken. The results for the *conceptual model* user (Figure 77 and Figure 78) reveal a significant decrease in the secondary criteria at all *C/L*, but an inability to maintain all the primary value. The *VS* was significantly lower from the control's *VS* only for *C/L* 1% to 4%, but the *POD* and *POMD* indicated a significant difference through *C/L* 12% (Figure 79).

Recall that the *optimal* overlap threshold was designed to find overlap threshold values that reduced repeat false alarms while maintaining the primary value metrics (Figure 80 and Figure 81). The optimal user was able to match the control user for the *VS, POD* and *POMD* metrics for all *C/L*, while also realizing an impressive improvement in secondary criteria.

The percent reduction in repeat false alarms (Figure 74) for the conceptual model and the optimal user indicated that both decision strategies improved as *C/L* increased (i.e., percent reduction increased). As *C/L* increased, the number of repeat false alarm opportunities decreased, thus any reversal comprised a larger proportion of the available opportunities. The conceptual model had significantly fewer repeat false alarms through *C/L* 12%, but since this decision rule degraded primary value over the same *C/L*, it was not superior over this range of users. The conceptual model employed relatively small overlap thresholds (Figure 41, page 112) compared to the optimal user (Figure 46, page 116) for the low *C/L* values (e.g., for *C/L* 1%, 0.5% versus 31.5%, respectively). Given the large number of false alarm opportunities for the low *C/L*, the conceptual model resulted in many more cases where the decision to protect was reversed, which lead to an increase in expensive misses.

Beyond *C/L* 12%, there was no significant difference in the percent reduction of repeat false alarms between the conceptual model and optimal users. Although the difference was insignificant, there was a crossover point ($C/L = 57\%$) where the expected

value in percent reduction for the optimal user began performing better than the conceptual model's expected value (i.e., fewer repeat false alarms for the optimal user on average). Since the conceptual model overlap threshold increased with increasing *C/L*, it allowed fewer decision reversals than the decreasing optimal overlap threshold.

As described in Chapter IV.C, the $CES_L$ ambiguity estimate used during this experiment unavoidably suffered from errors in the location of the ambiguity distribution as a result of being centered on the control forecast probability. While the $CES_L$ distribution likely provided a robust estimate of the variance of the ambiguity distribution, the range of possible forecast probability values may be shifted (compared to the ambiguity distribution from EoE). The shift in the $CES_L$ ambiguity distribution was random since the control forecast probability is a random sample from the EoE ambiguity distribution. Thus there are random errors in the amount of overlap in cases where the decision is unclear, resulting in sub-optimal application of the ambiguity information. However, even with this deficiency, $CES_L$ clearly added value to the secondary criteria.

The results attained during this study clearly show the value of employing an estimate of the ambiguity associated with the ensemble forecast. The decision rules explored above provided evidence that mere random or arbitrary reversals of the decision for repeat false alarm opportunities were inferior to reversals made by intelligently applying the ambiguity estimate (even if the estimate was flawed) only when the decision was unclear. Moreover, we were able to train our decision process based on past performance to optimally select an overlap threshold at each *C/L* to maintain primary value while significantly adding value to our secondary criteria.

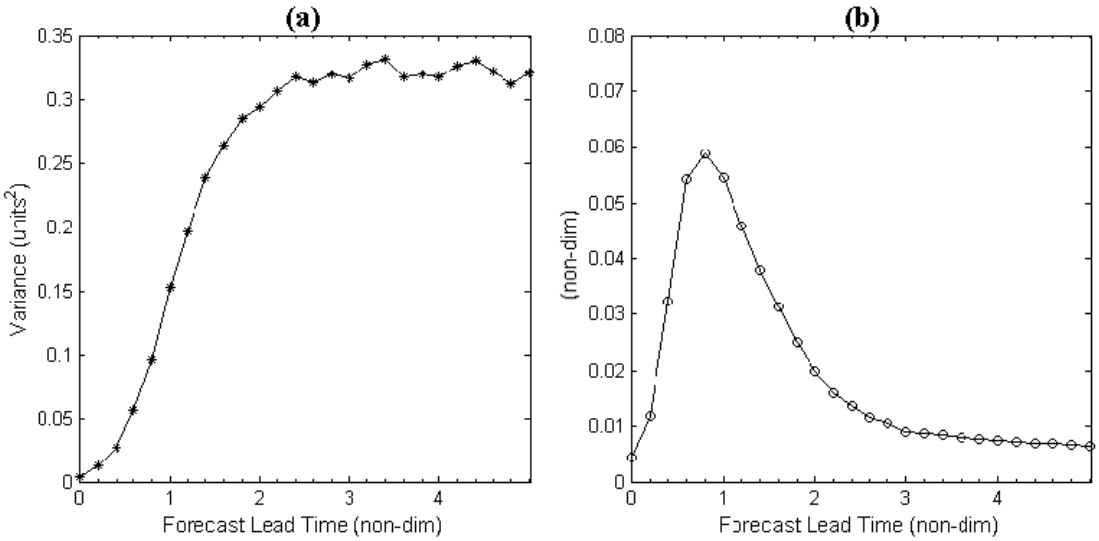



Figure 48. Evolution of L96M EPS error variance for (a) mean error of ensemble mean and (b) fractional error in ensemble spread. The error variances are shown following calibration to remove systematic error.

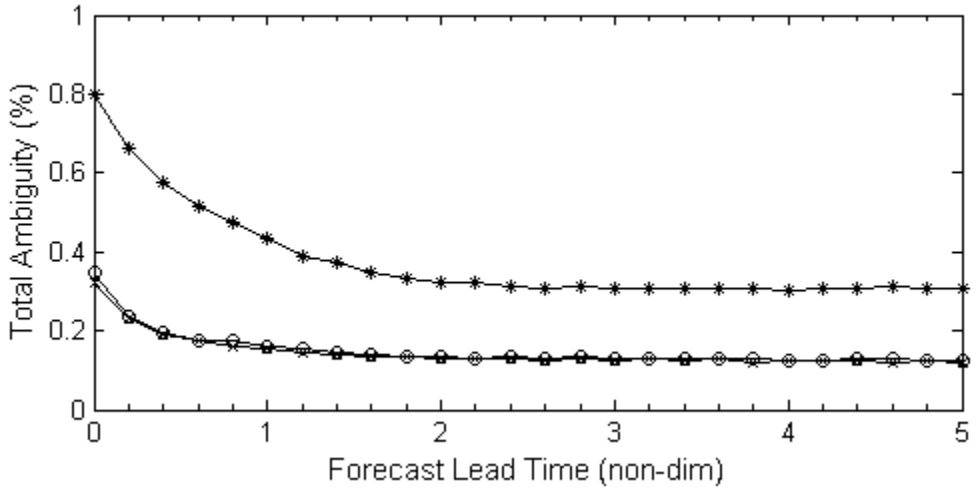



Figure 49. Average total ambiguity of the EoE ambiguity distributions for test forecast probability values 5% (o), 50% (*) and 95% (x).

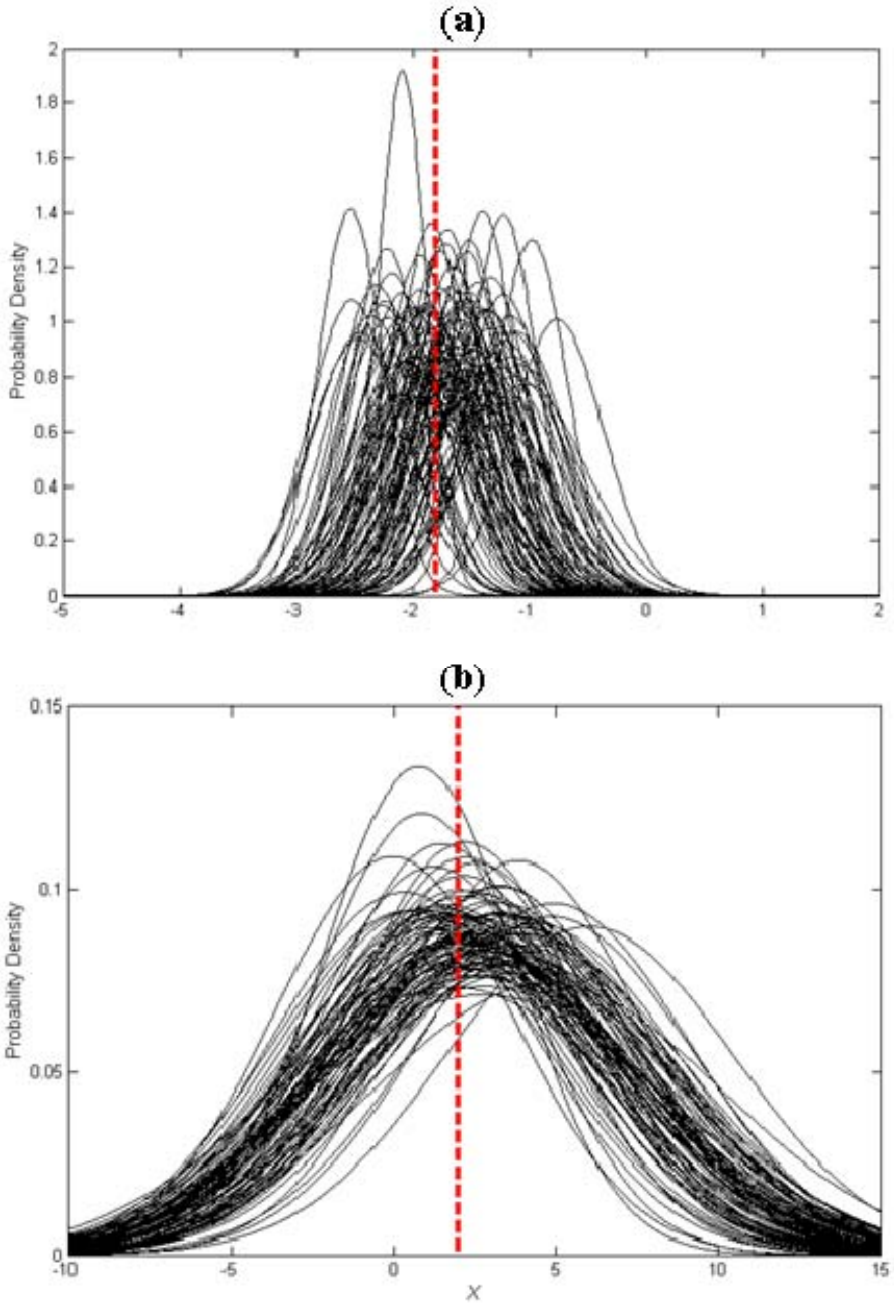



Figure 50. Arrangement of EoE constituents at a (a) high and (b) low ambiguity timeframe. The PDFs for 100 constituents in a single EoE forecast case are displayed using a normal fit (solid lines) for (a) $\tau = 0.2$ and (b) $\tau = 4.8$ time units. An arbitrary event threshold (dashed line) is also shown for analysis of forecast probability values for each constituent. Note that in (b) a different event threshold is used, and abscissa and ordinate scaling has changed.

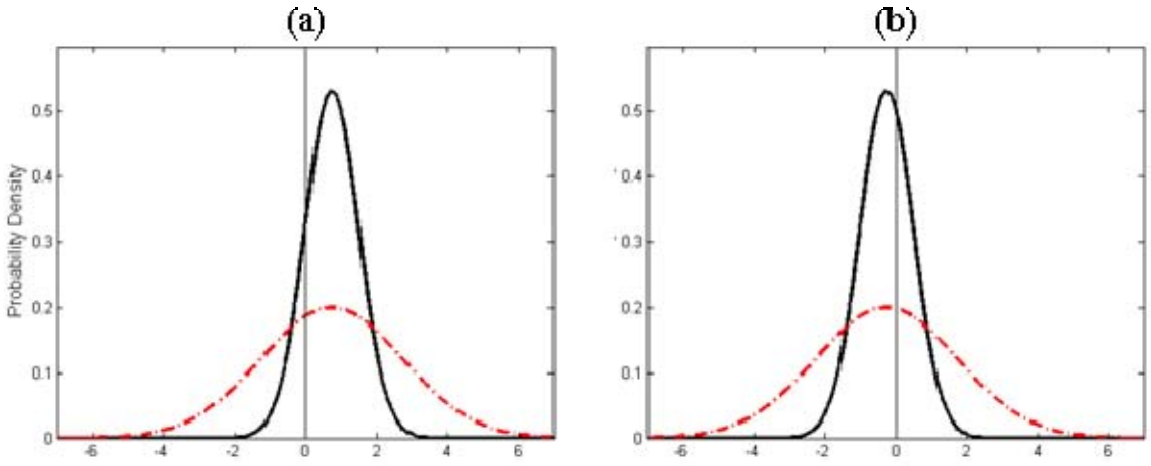



Figure 51. Example of forecast probability sensitivity to PDF spread and shifts in PDF location for low spread (thick solid) and high spread (dot-dash) PDF. In (a), both PDFs are located at 0.75, and the probability of preceding the event threshold (thin solid) is 15.9% and 35.4% for the low and high spread PDFs, respectively. In (b), each PDF is shifted to -0.25 while holding spread constant, giving probability values of 63.1% and 55% for the low and high spread PDFs, respectively. Probability for the low spread PDF changed by 47.2%, while the change was 19.6% for the high spread PDF.

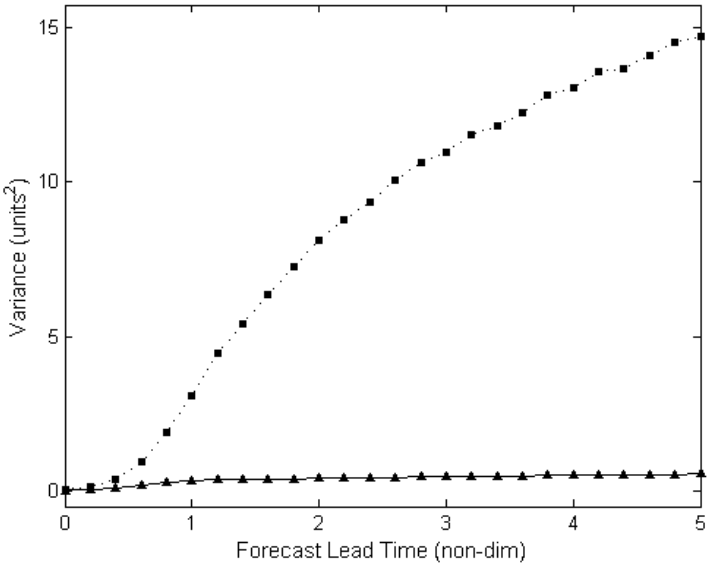



Figure 52. Comparison of average variance between EoE constituent ensemble forecast mean values (▲) and average variance of EoE constituent ensemble forecasts (■) with increasing lead time. The comparison was made using 100 EoE forecast cases each containing 100 constituent ensemble forecasts.

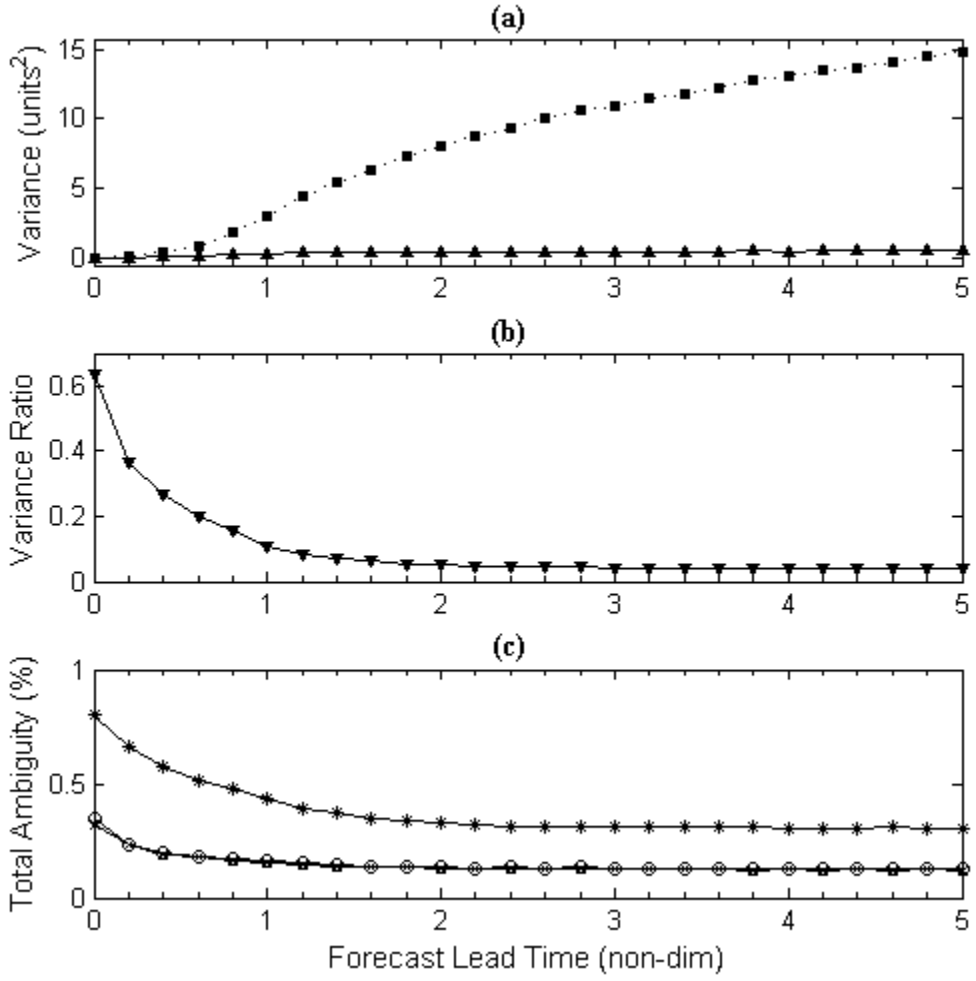



Figure 53. Comparing the average evolution of EoE constituent relationships to the typical EoE ambiguity evolution using (a) same as Figure 52, (b) the ratio of average variance in location of EoE constituent ensemble forecasts' means to average constituent variance and (c) same as Figure 49.

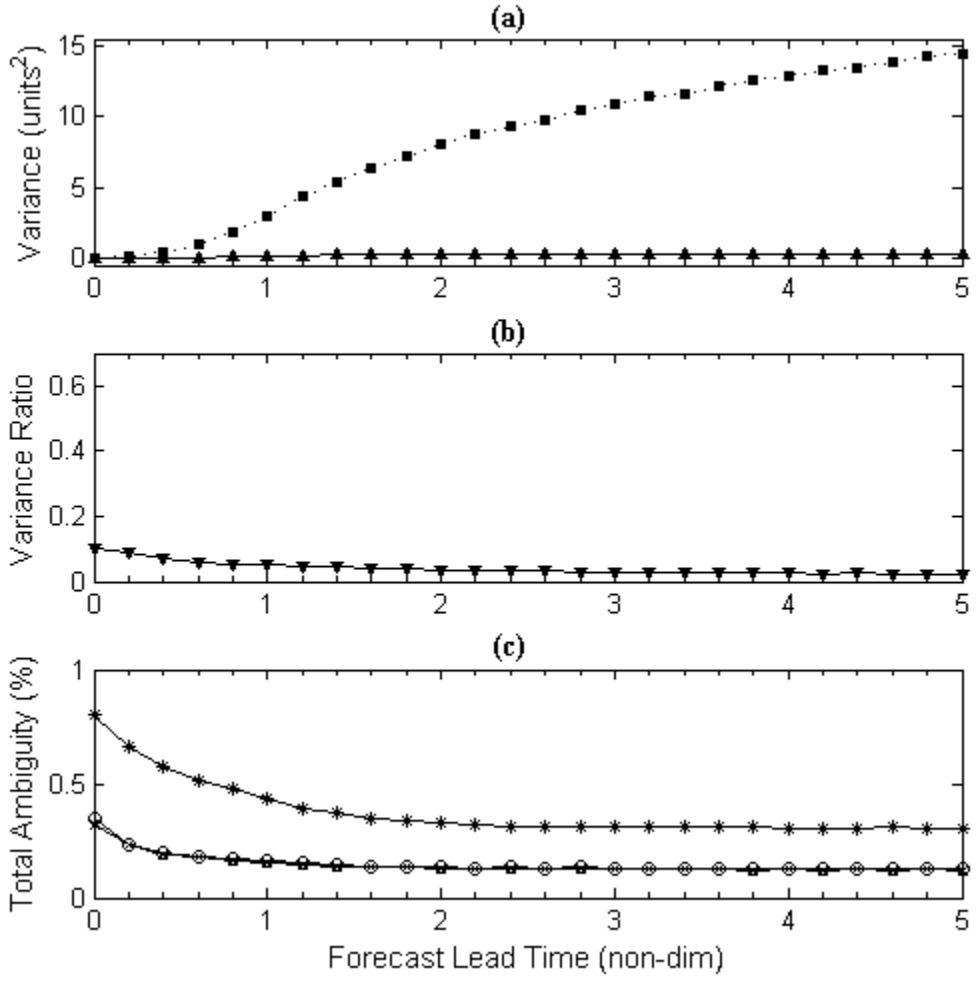



Figure 54. Comparing the evolution of average L96M ensemble forecast statistics to the typical EoE ambiguity evolution using (a) the variance of mean error in the ensemble mean (▲) and average ensemble forecast variance (■) computed from 24,000 L96M forecast cases, (b) the ratio of the variance of the mean error in the ensemble mean to the average ensemble variance in location and (c) same as Figure 49.

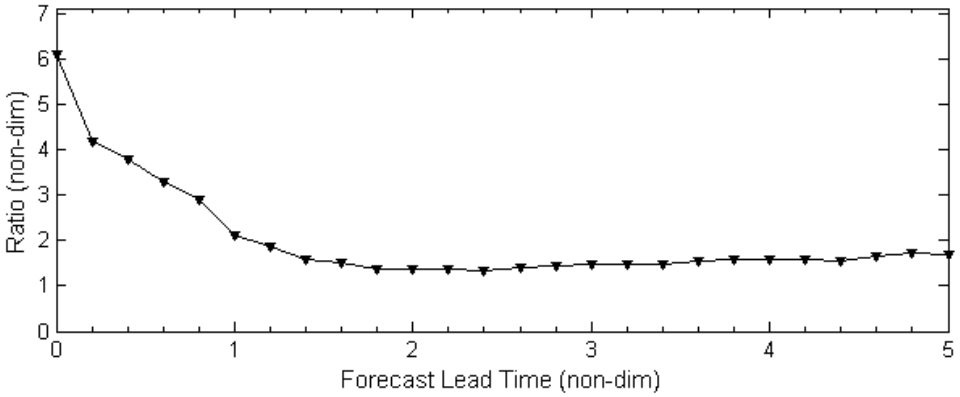



Figure 55. Ratio of average variance of EoE constituent ensemble forecast means to the variance of the mean error in the ensemble forecast mean. The average variance in constituent means is computed using 100 EoE forecast cases each with 100 constituent forecasts. The mean error is computed using 24,000 L96M EPS forecast cases, where the variance in mean error is found by computing the mean error over 3,000 subsets of eight forecasts each and taking the variance.

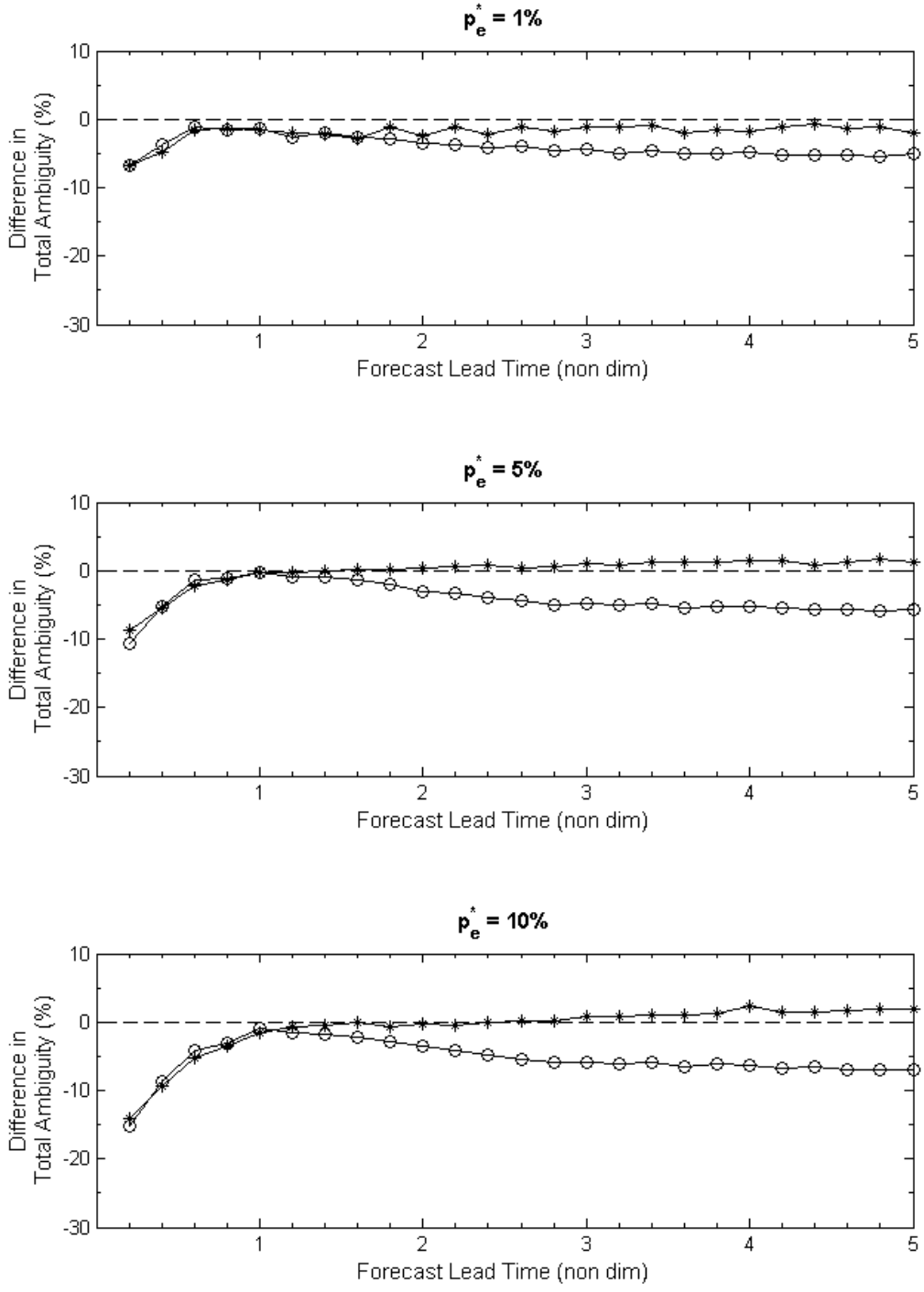
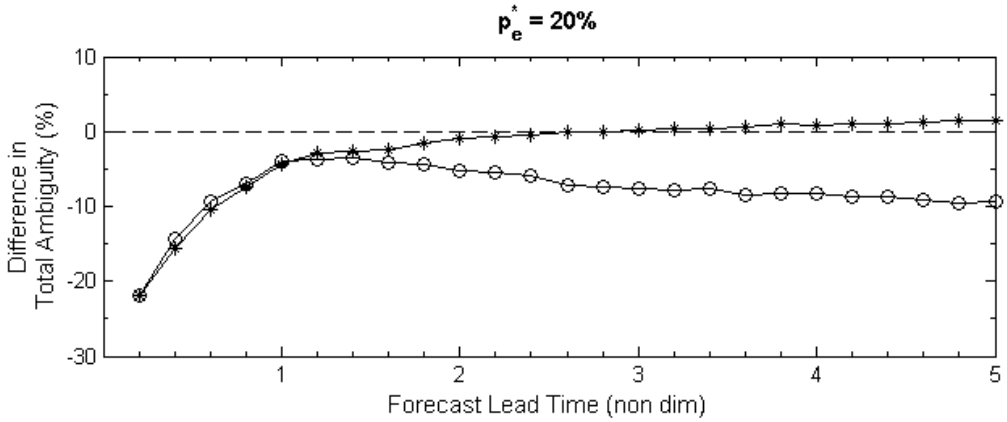
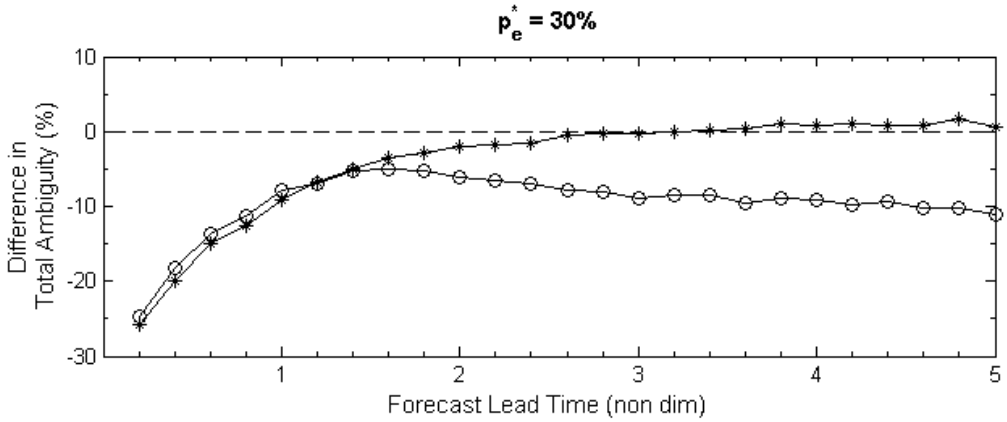
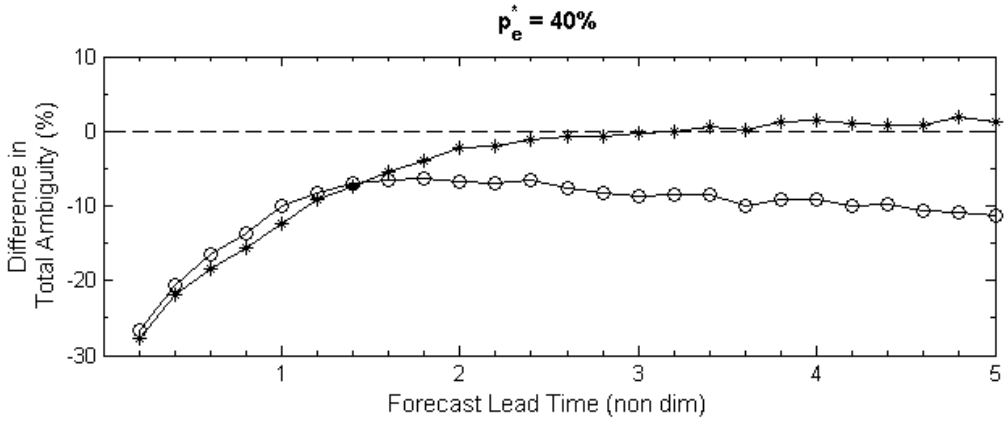


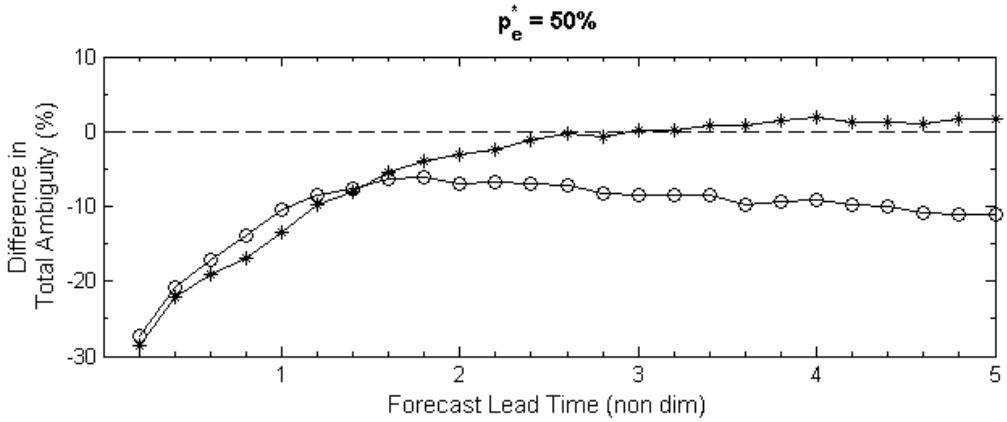

Figure 56. Validation of $CES_G$ (o) and RCR (∗) total ambiguity across all forecast lead times for the specific $p_e^*$ test values (shown in Figure 37, page 108), which are labeled at the top of each panel.

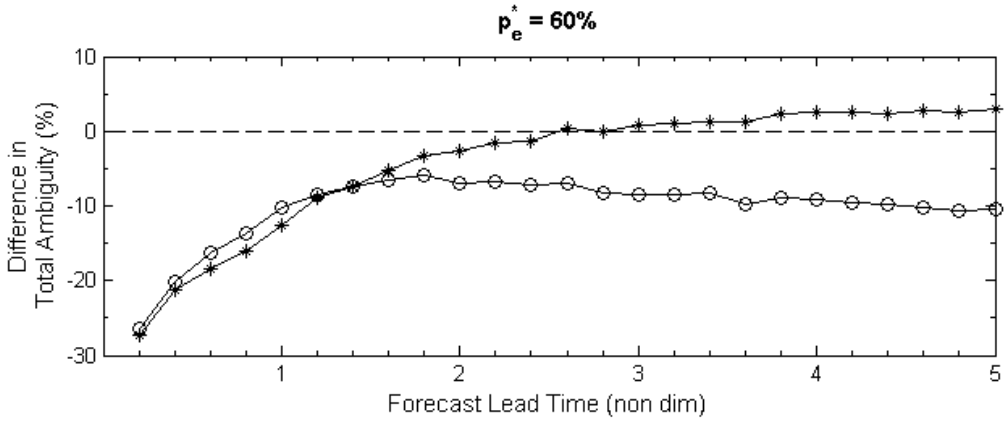


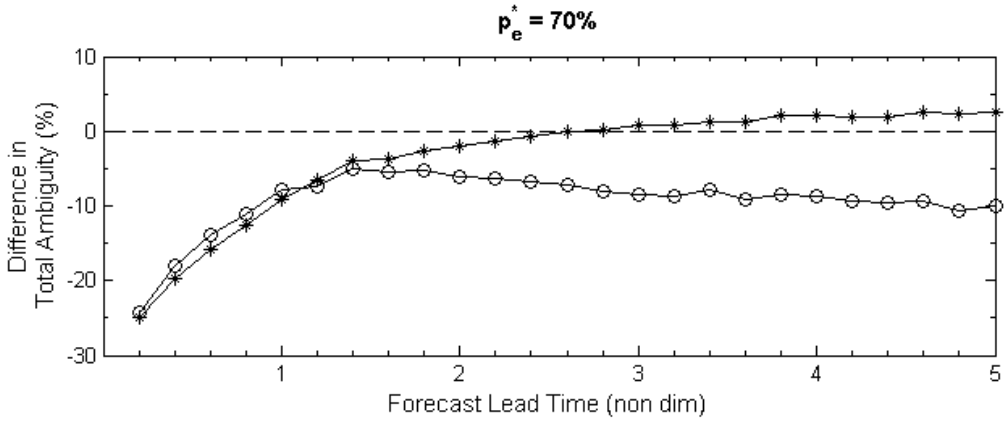



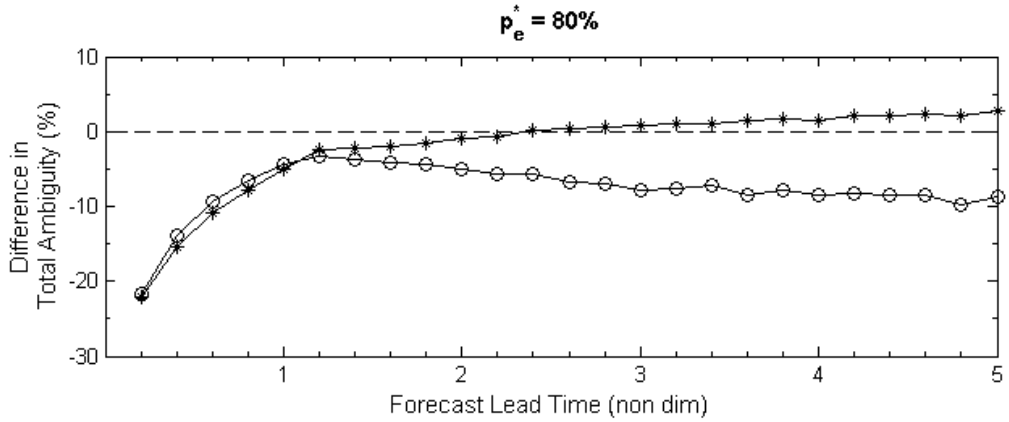

(Figure 56 continued.)

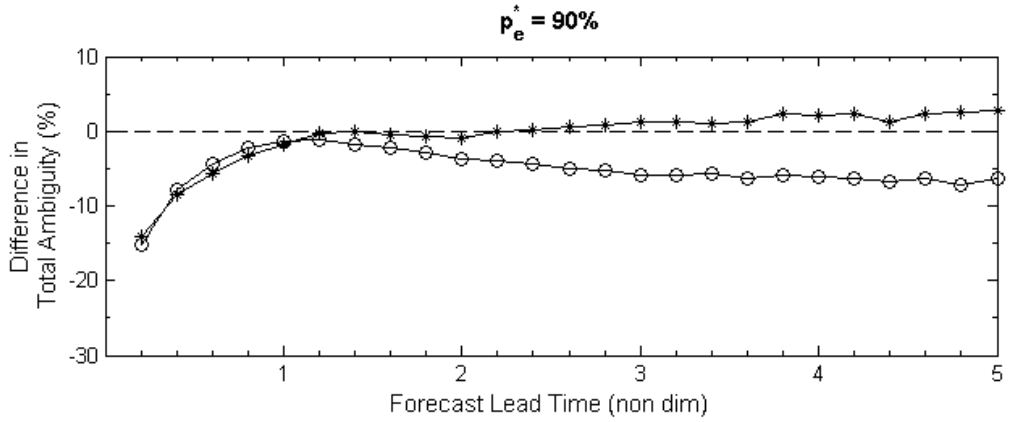


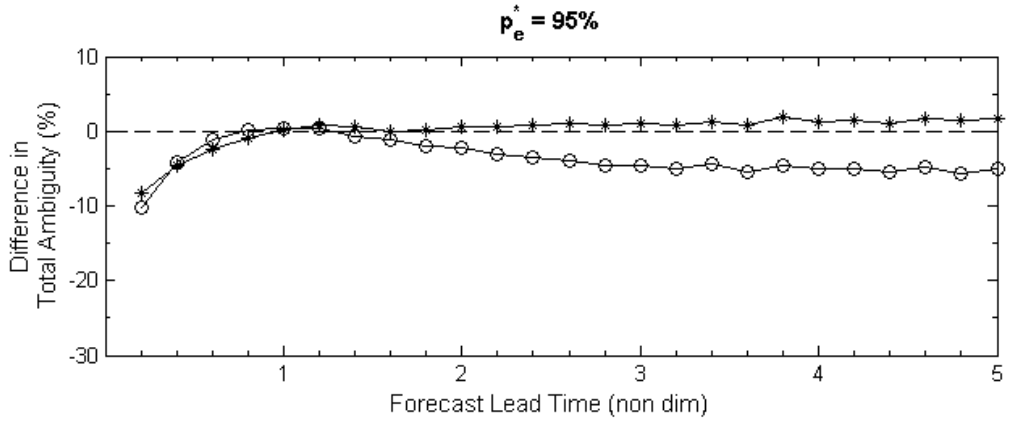



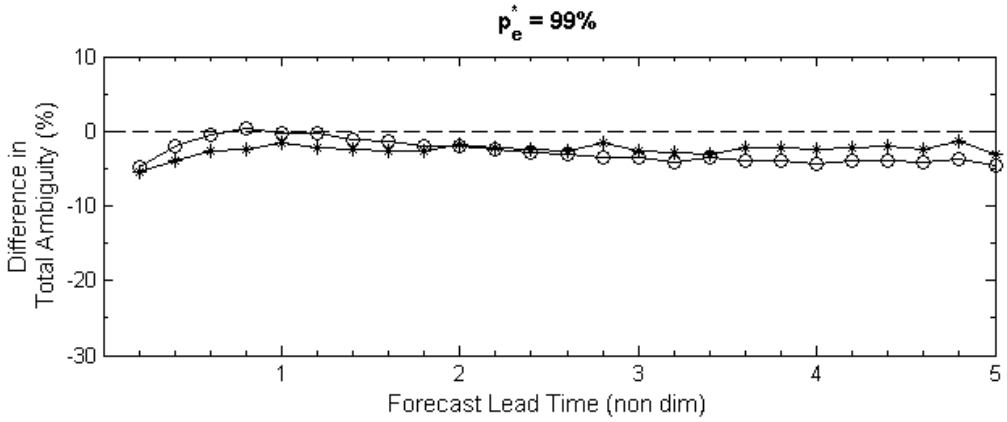

(Figure 56 continued.)




(Figure 56 continued.)




(Figure 56 continued.)

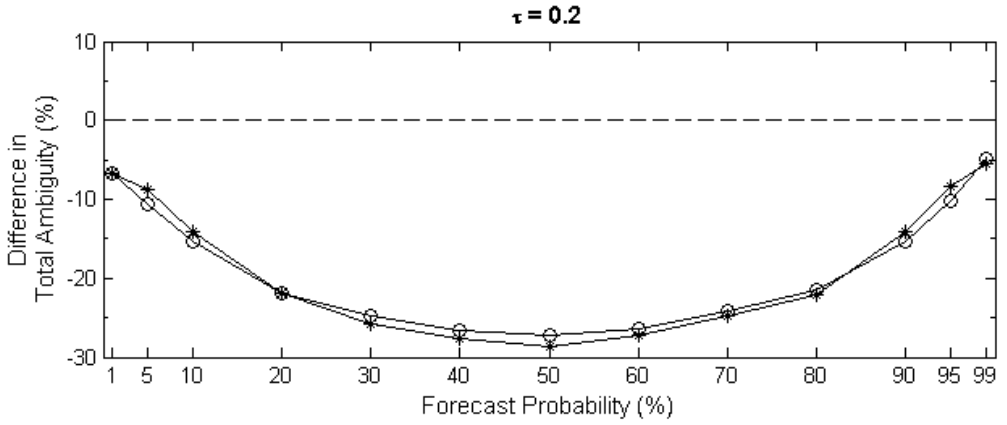
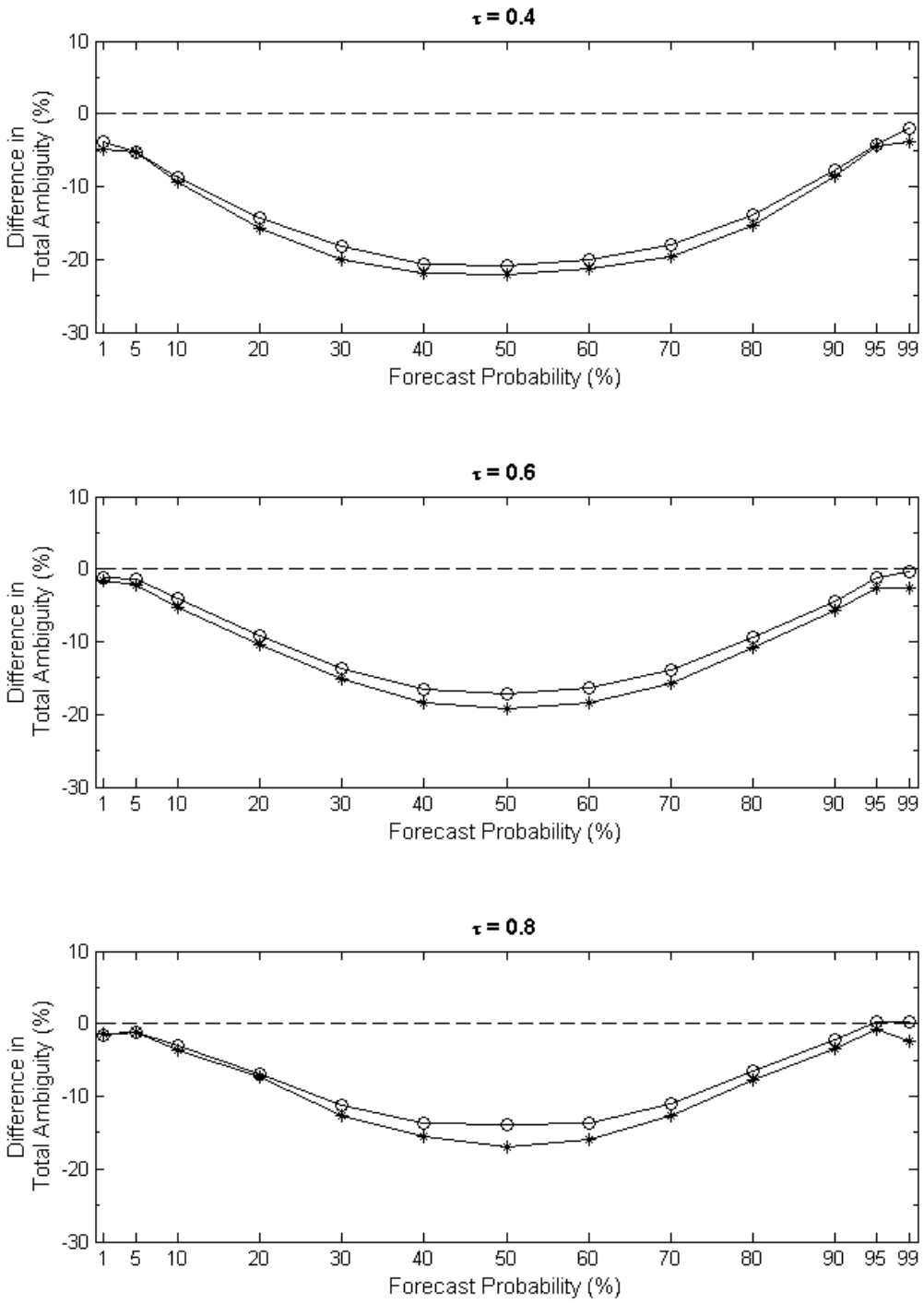
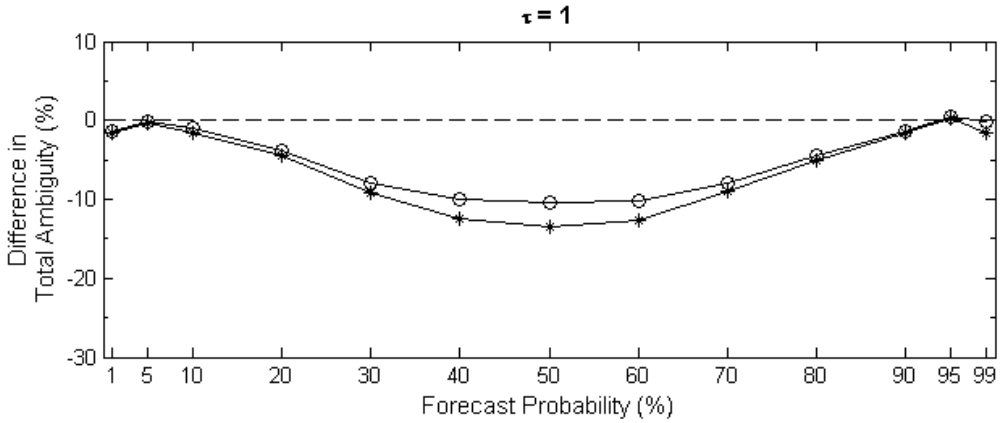
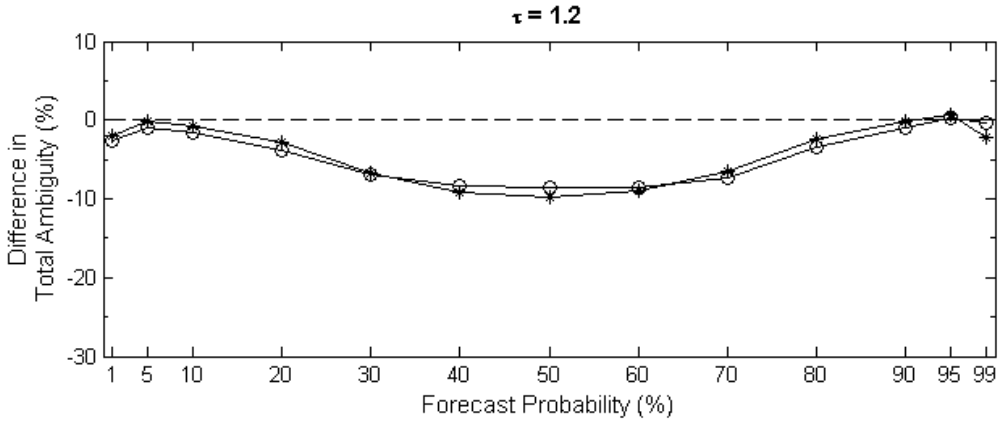


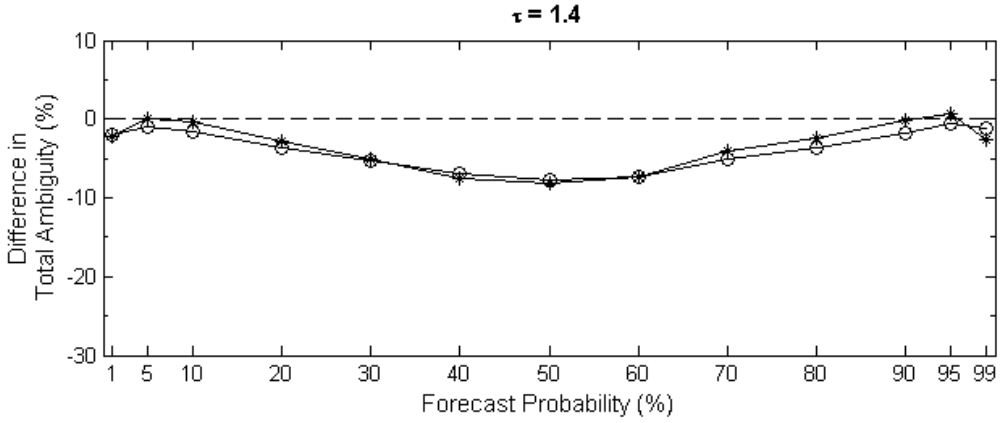

Figure 57. Validation of $CES_G$ (o) and RCR (*) total ambiguity at select calibrated forecast probability values ( $p_e^*$ ) (shown in Figure 37, page 108) for forecast lead times 0.2-5.0 at an increment of 0.2. Lead times ( $\tau$ ) are labeled at the top of each panel.



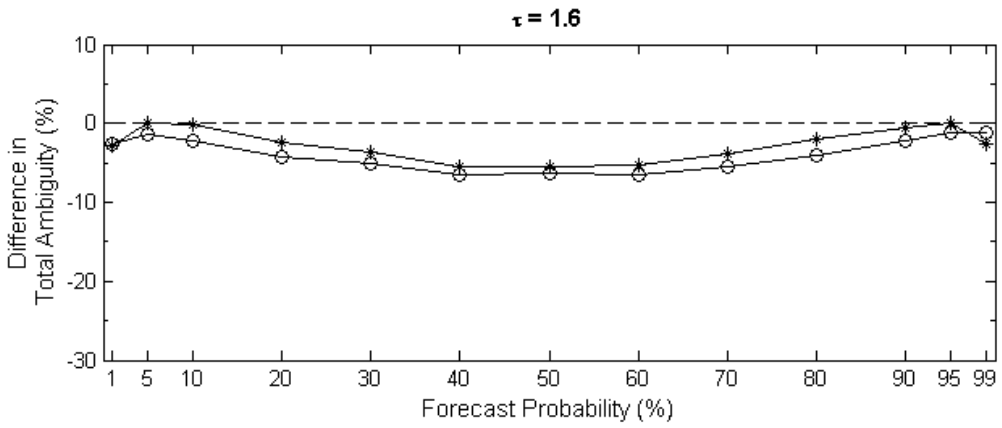

(Figure 57 continued.)

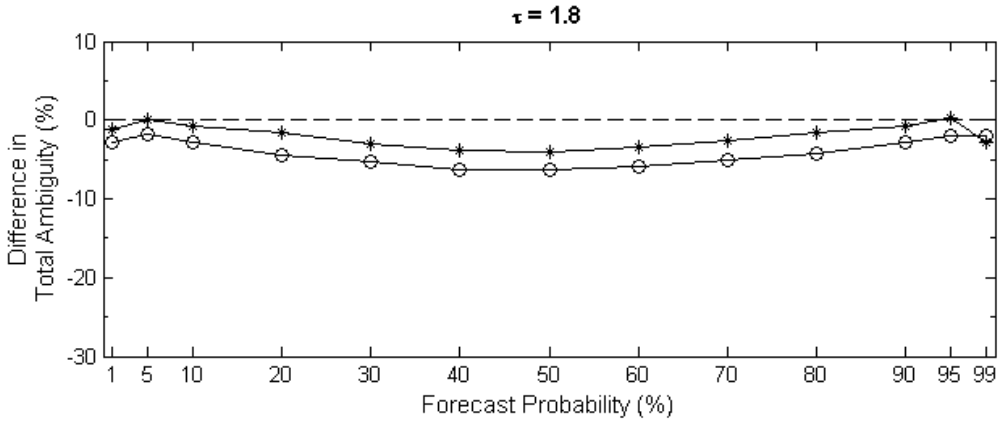


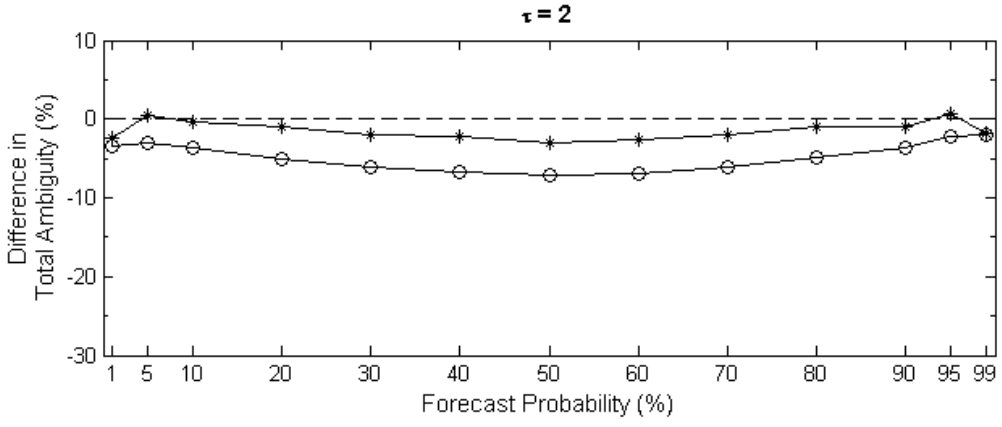



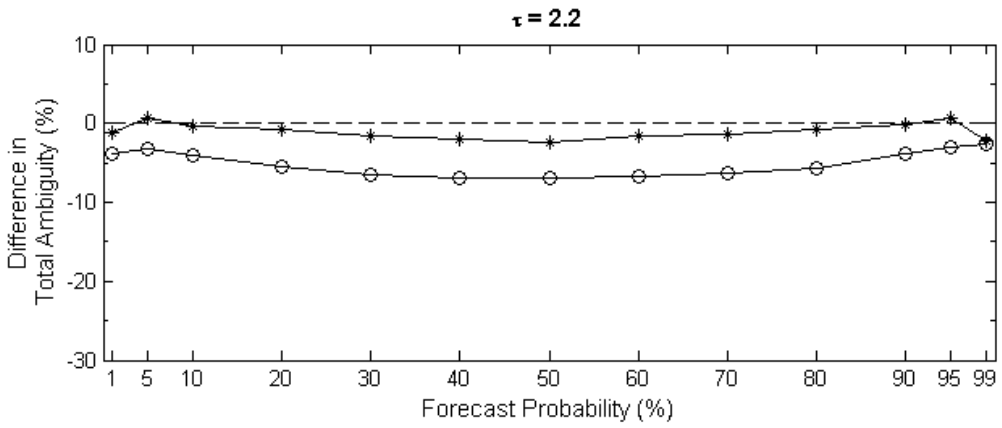

(Figure 57 continued.)

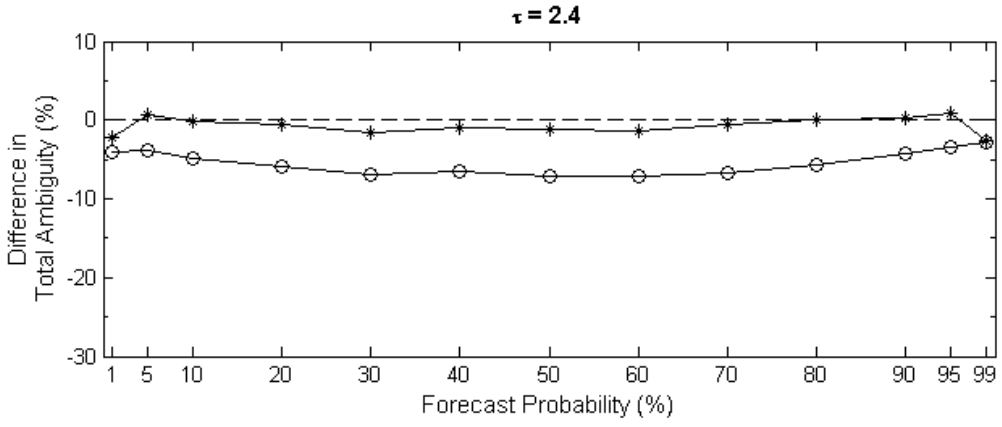


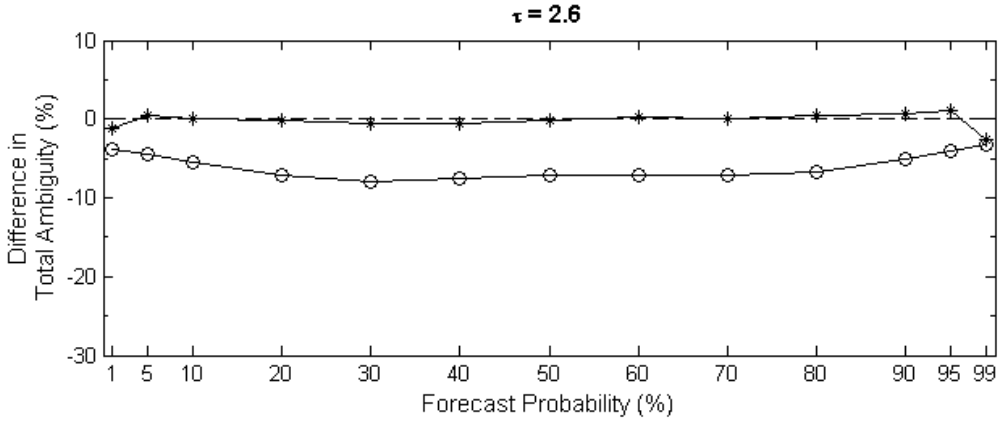



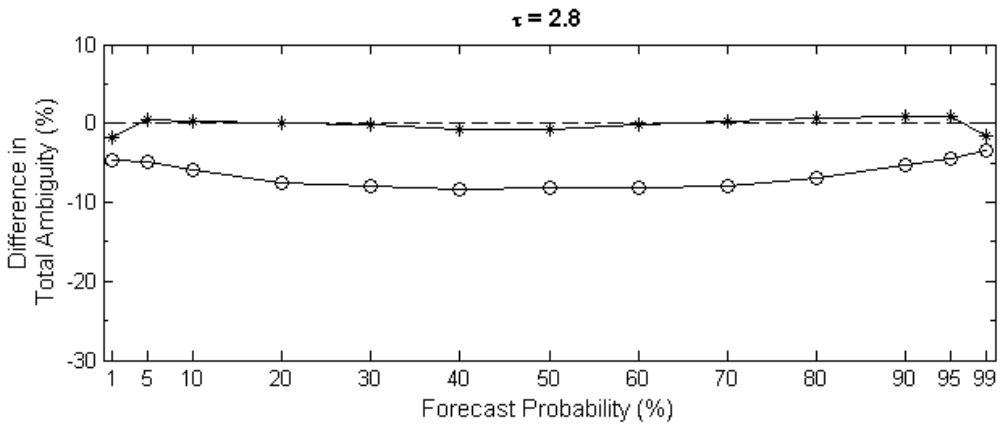

(Figure 57 continued.)

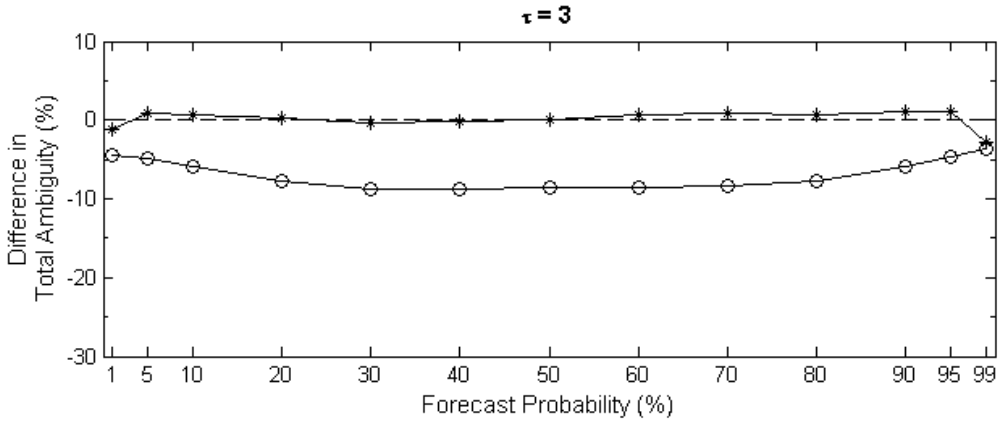


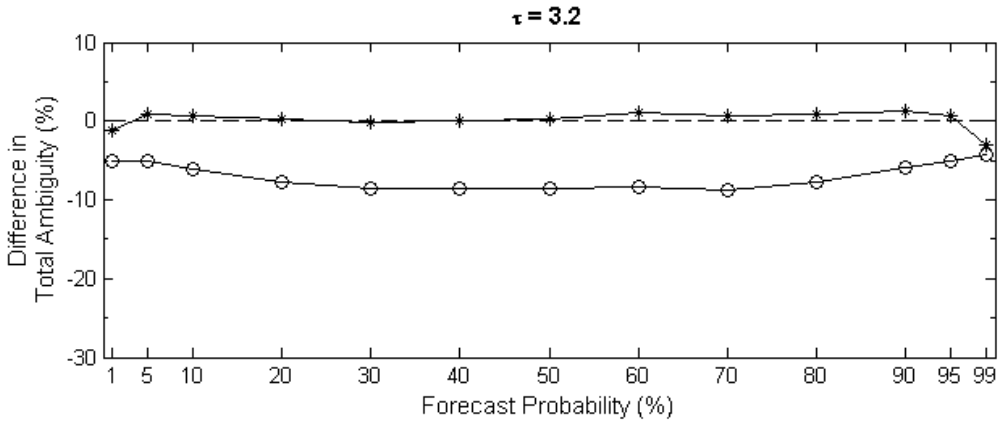



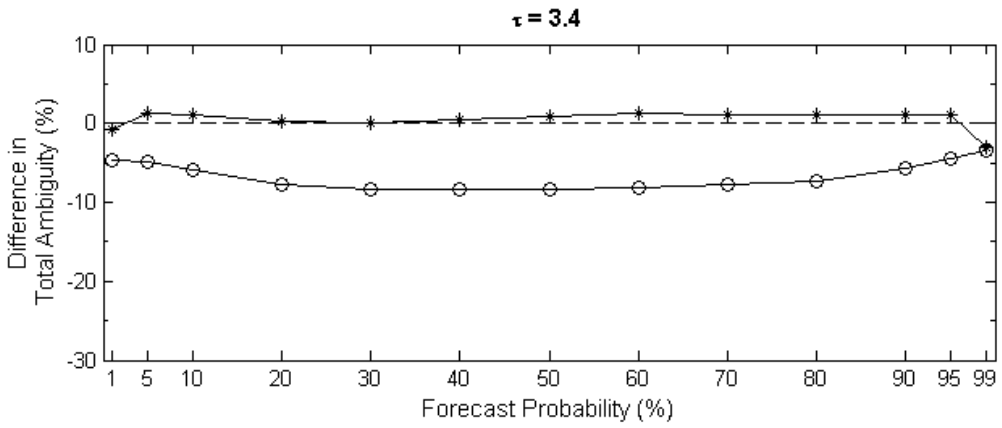

(Figure 57 continued.)

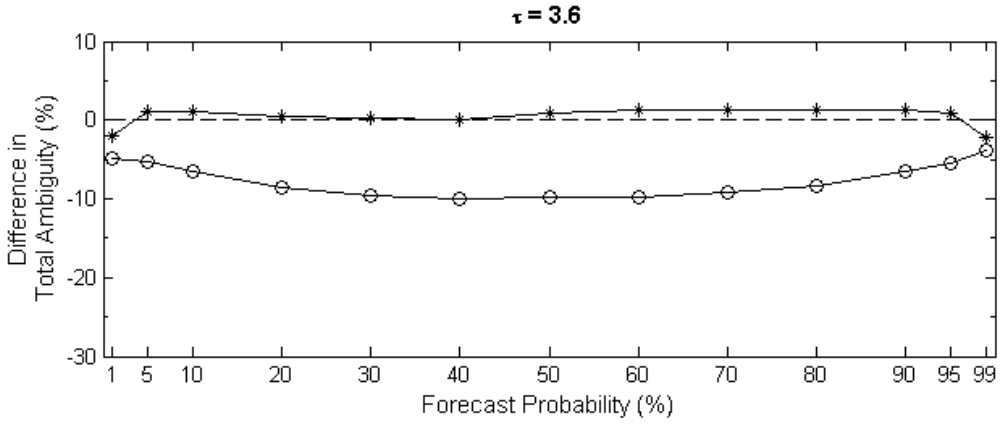


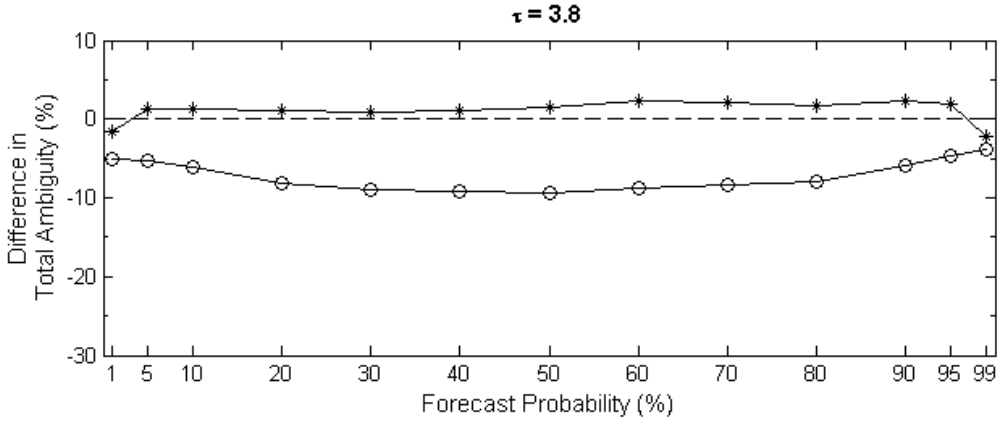



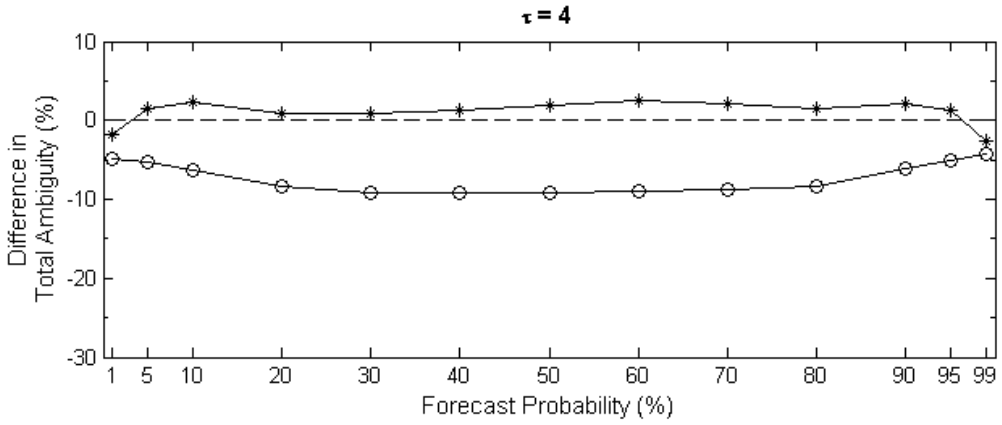

(Figure 57 continued.)

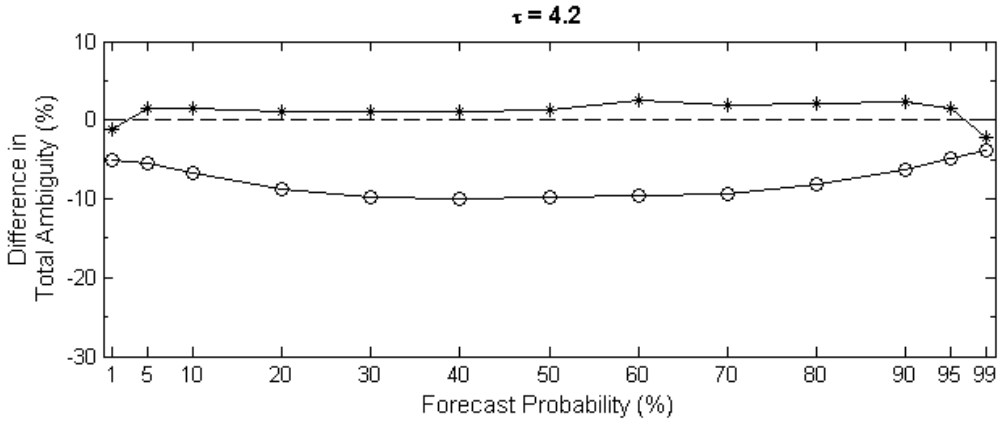


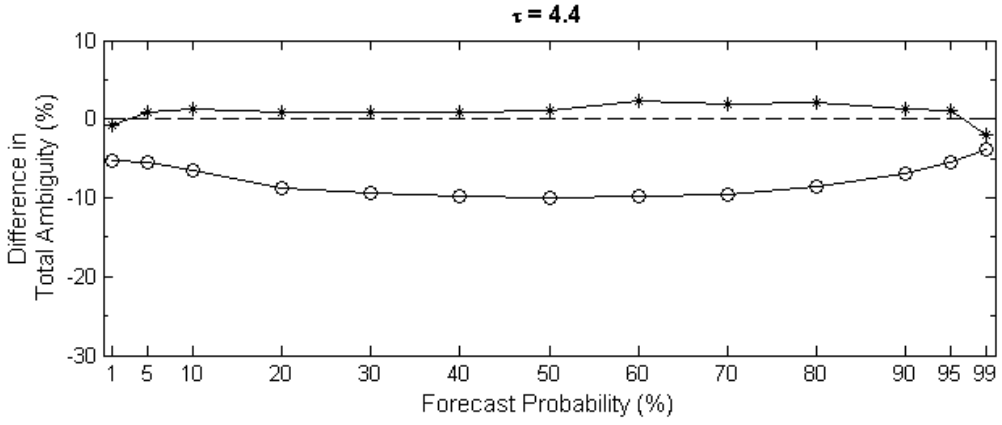



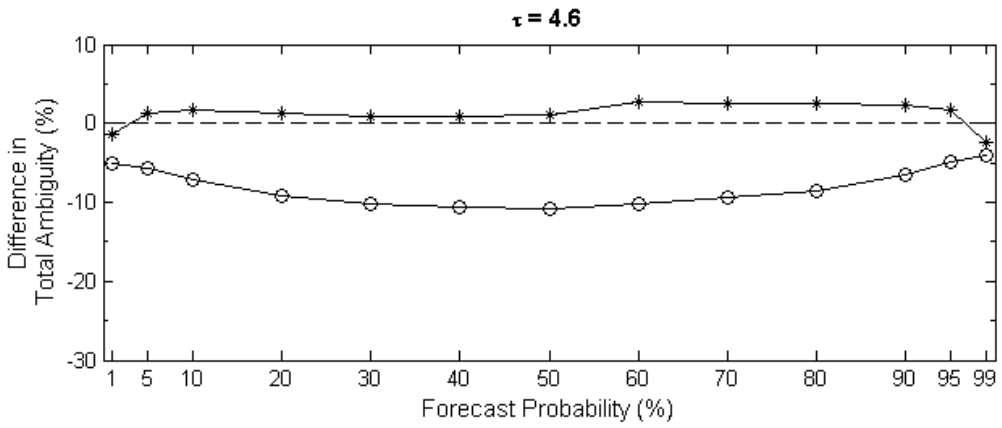

(Figure 57 continued.)

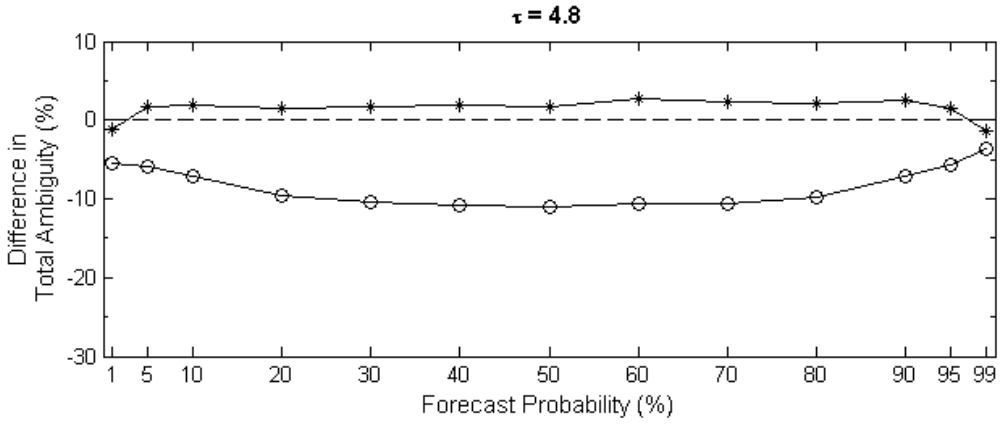


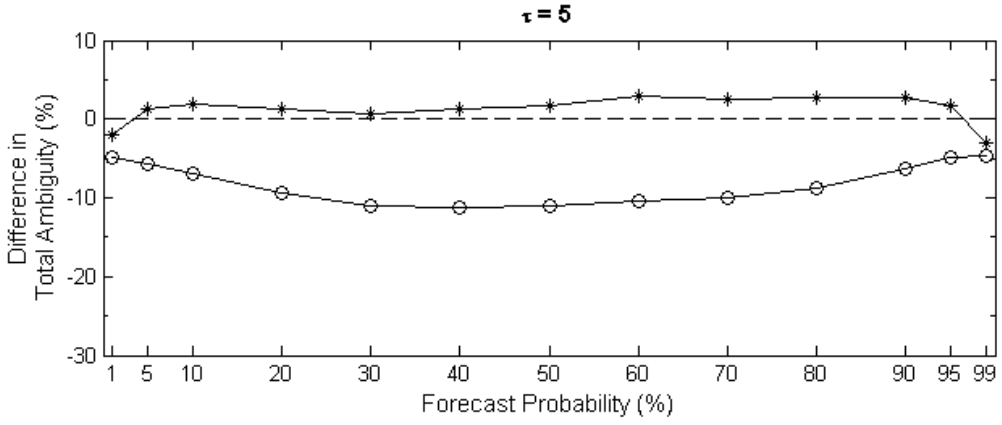




(Figure 57 continued.)




(Figure 57 continued.)

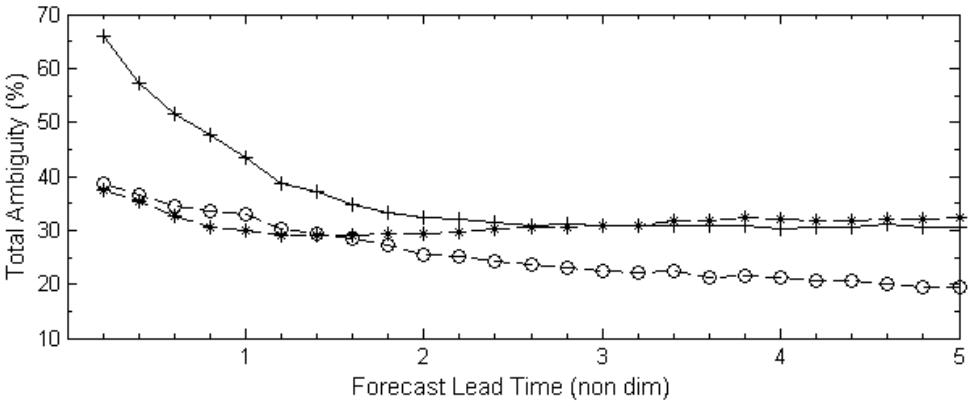



Figure 58. Total ambiguity evolution for EoE (+), CES (o), and RCR (∗) for ambiguity distributions with expected value of 50%.

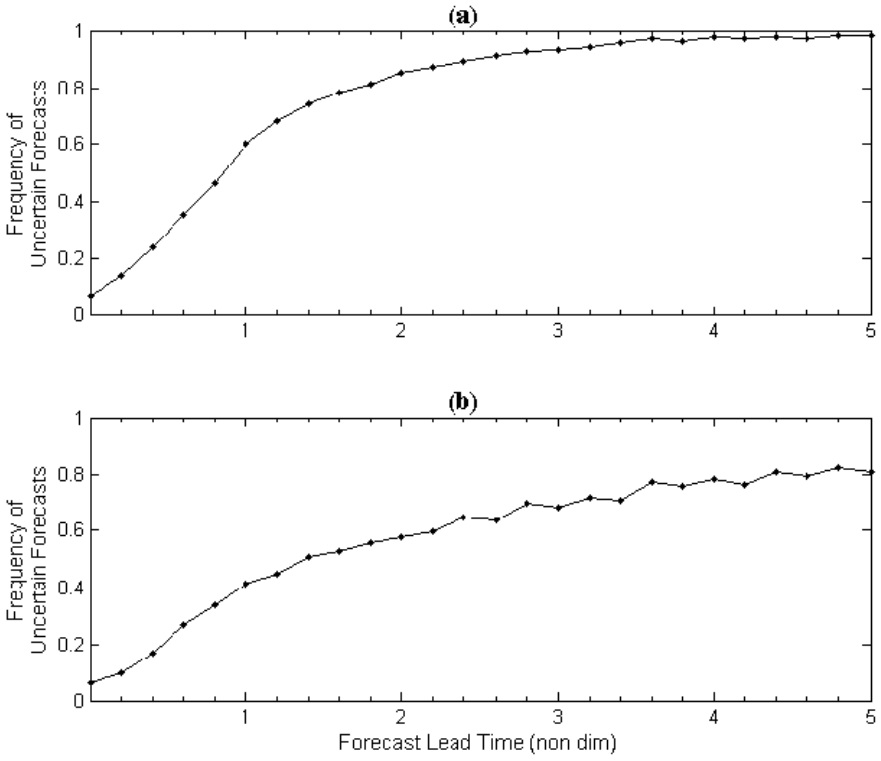



Figure 59. Frequency of uncertain ensemble forecasts (i.e., control ensemble forecasts with $p_e^*$ between 0.1% and 99.9%) for (a) the common event of $X \geq 6.31$ and (b) the rare event of $X \geq 9.98$. The ensemble forecast for each variable from the first constituent of each EoE forecast case was utilized as a control ensemble forecast for a total of 800 forecasts.

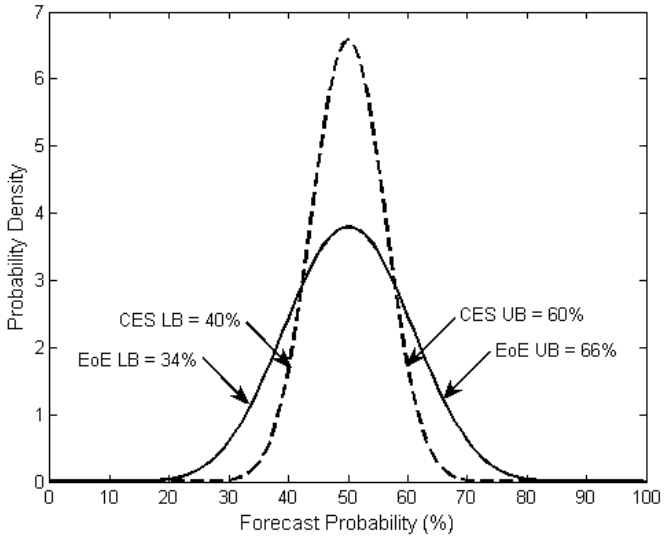



Figure 60. Ambiguity distributions for EoE (solid) and $CES_G$ (dashed) with expected value equal to 50% for a single EoE forecast case at $\tau = 5$ time units for a single $X_k$ variable. The distributions are approximated using a beta-fit to the estimated forecast probability values for each technique. The upper (UB) and low (LB) bounds of each technique's 90% CI (i.e., total ambiguity) are labeled.

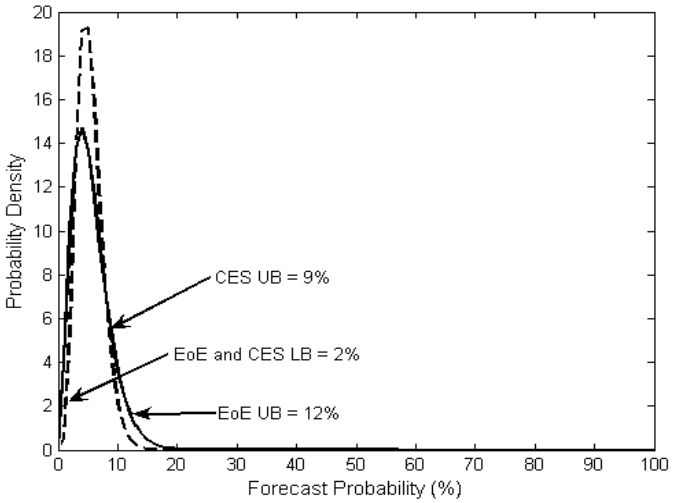



Figure 61. Ambiguity distributions for EoE (solid) and $CES_G$ (dashed) with expected value equal to 5%. Same as Figure 60.

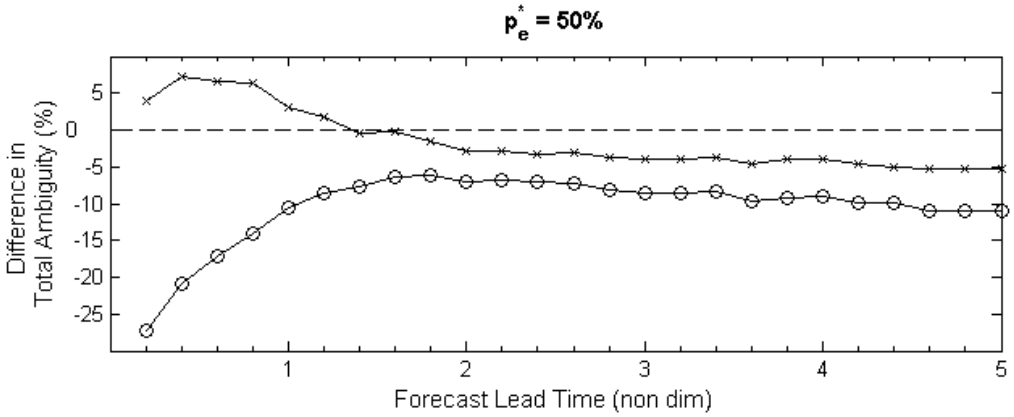



Figure 62. Comparison of validation of $CES_G$ without correction (o) and with correction (x) applied to the variance of the $ME_{\bar{e}}$ distribution. The correction is based on the ratio of variance in EoE constituents' location to variance in $ME_{\bar{e}}$ (Figure 55).

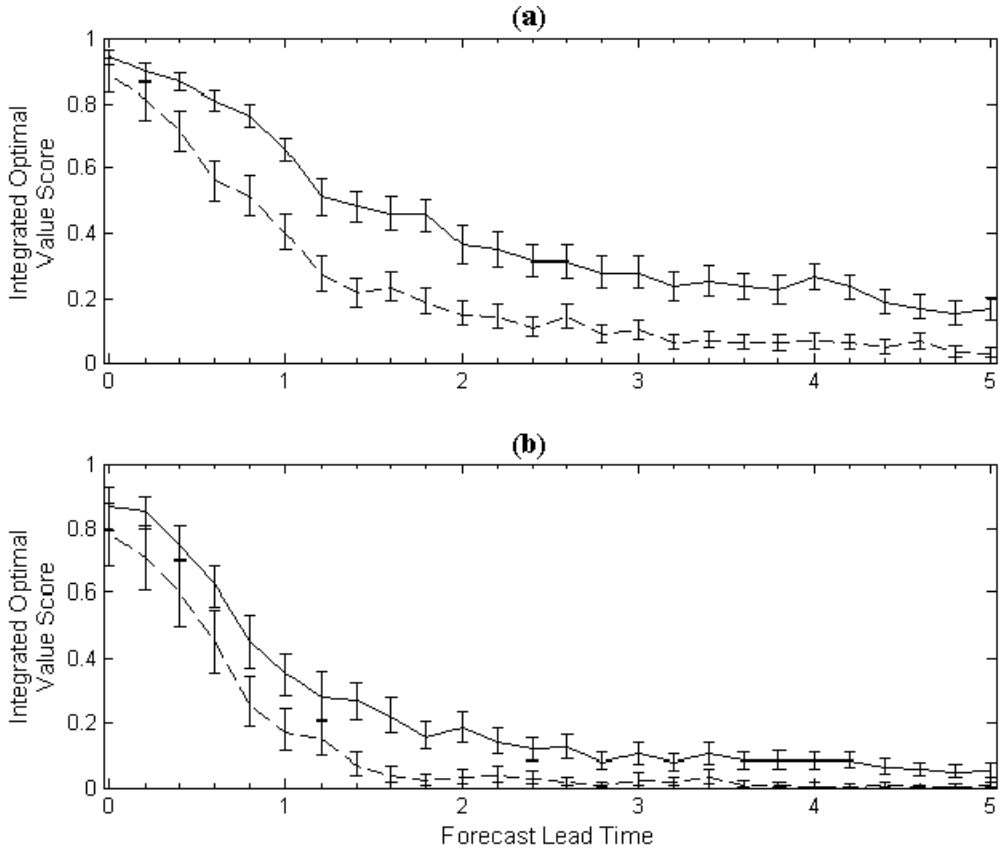



Figure 63. Integrated optimal value score [*IOVS*, Equation (27), page 72] for the calibrated control ensemble forecast (solid) and the deterministic forecast (dashed) for (a) the common event and (b) the rare event.

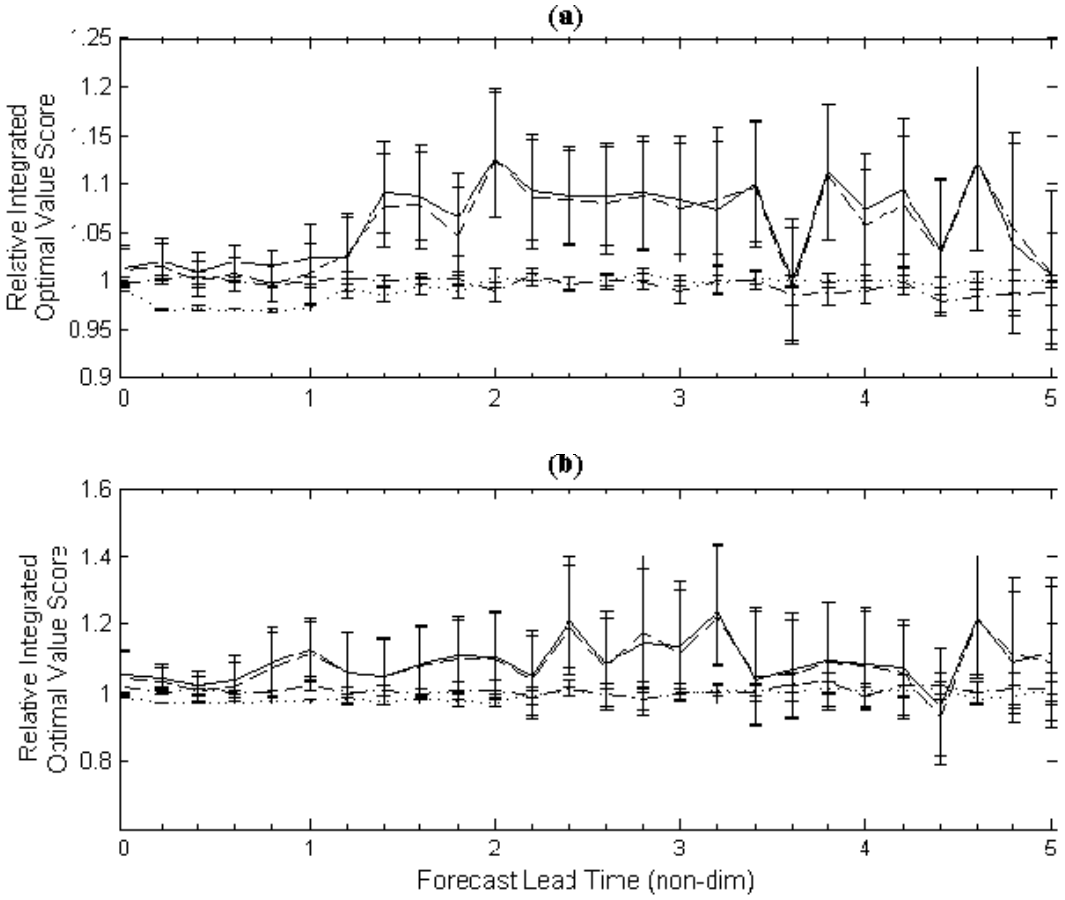



Figure 64. Relative integrated optimal value score [IOVS, Equation (27), page 72] using uncertainty-folding with EoE (dashed), $CES_G$ (dotted) and RCR (dot-dashed) for (a) the common event and (b) the rare event. The score for the grand ensemble (solid) is also shown in both panels. Error bars represent the 95% CI found using resampling. Note the ordinate scale change between (a) and (b).

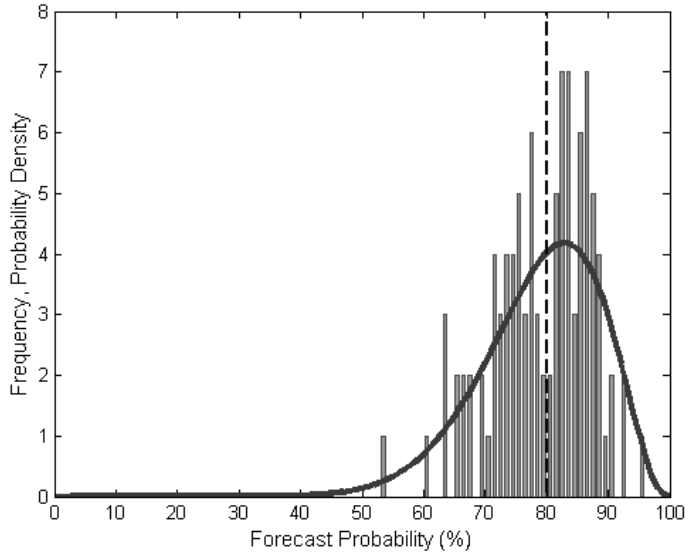



Figure 65. Control forecast probability well located with respect to the expected value of the EoE ambiguity distribution (80%). A histogram of $\hat{p}_T$ values for a single EoE forecast case (100 constituents) is shown with a Beta-fit curve for the RCR ambiguity distribution (solid line) created using the first constituent in the EoE forecast case as the control forecast. The control forecast probability ( $p_e^* = 80\%$ ) is marked by the dashed line.

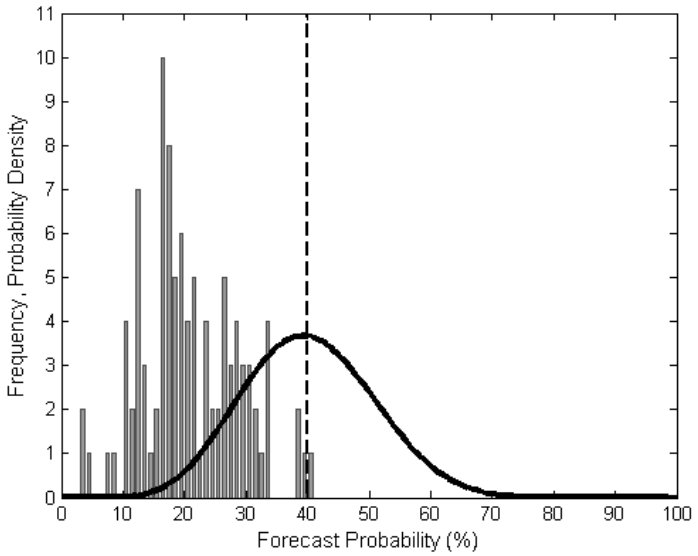



Figure 66. Control forecast probability poorly located with respect to the expected value of EoE ambiguity distribution. Same as Figure 65 with the expected value of the EoE ambiguity distribution at 20% and $p_e^* = 40\%$.

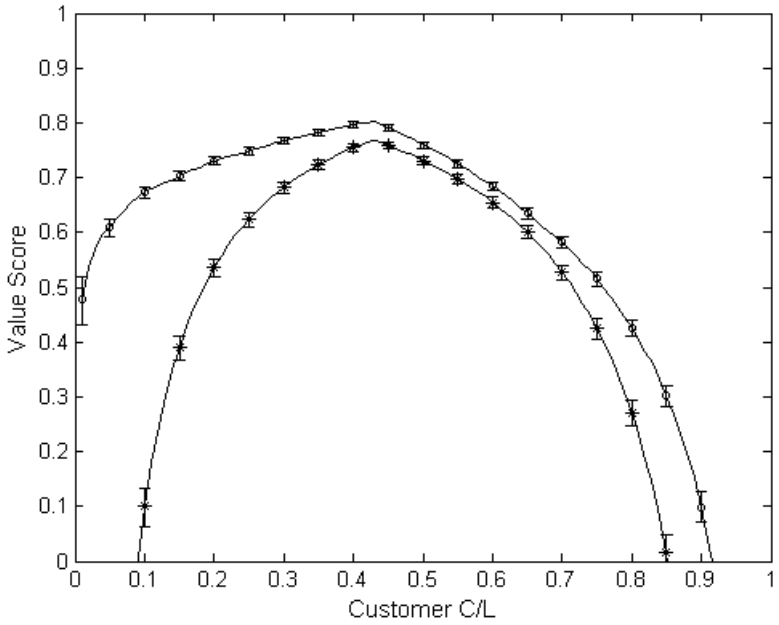



Figure 67. Optimal *VS* comparison for the GFS deterministic forecast (∗) versus the GEFS forecast (o) using the application dataset of 50,220 forecast-observation pairs.

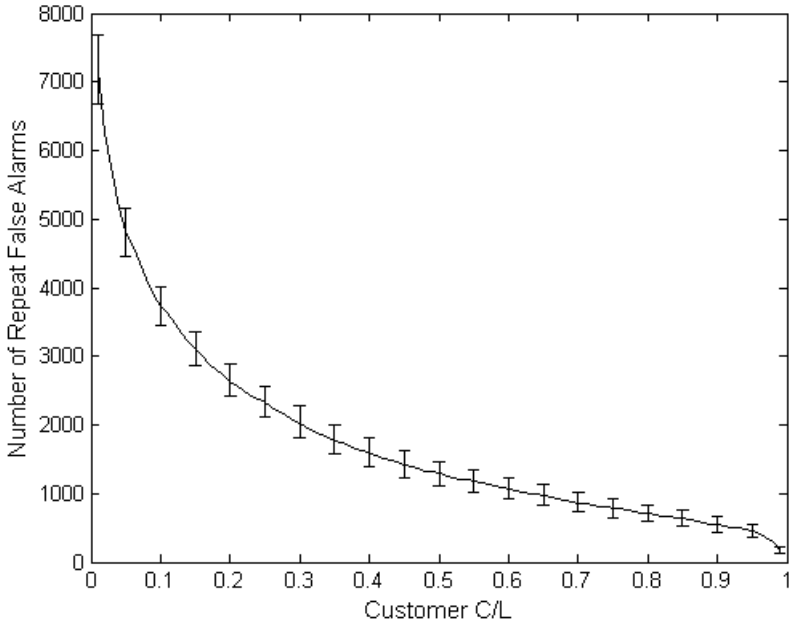



Figure 68. Number of repeat false alarms for the control user at each *C/L* based on the application dataset of 50,220 forecast-observation pairs.

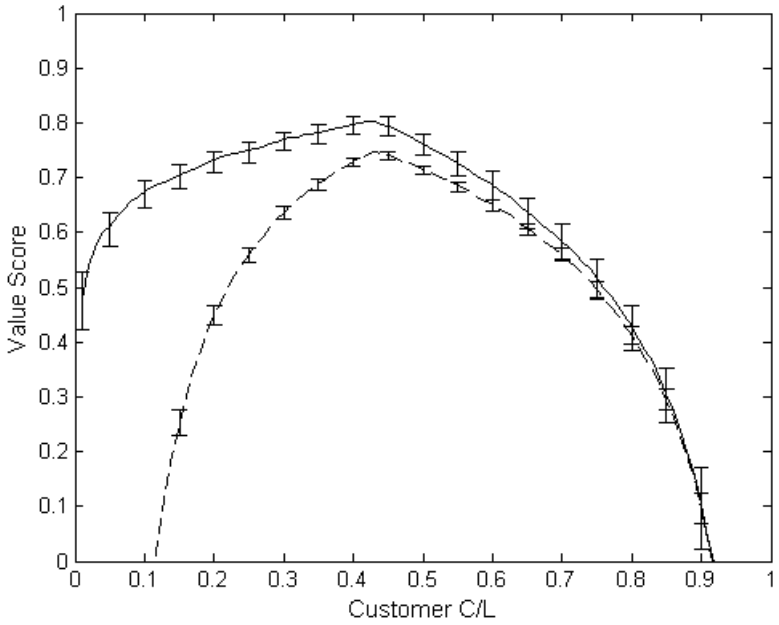



Figure 69. Optimal *VS* comparison for the control user (solid) versus the always user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

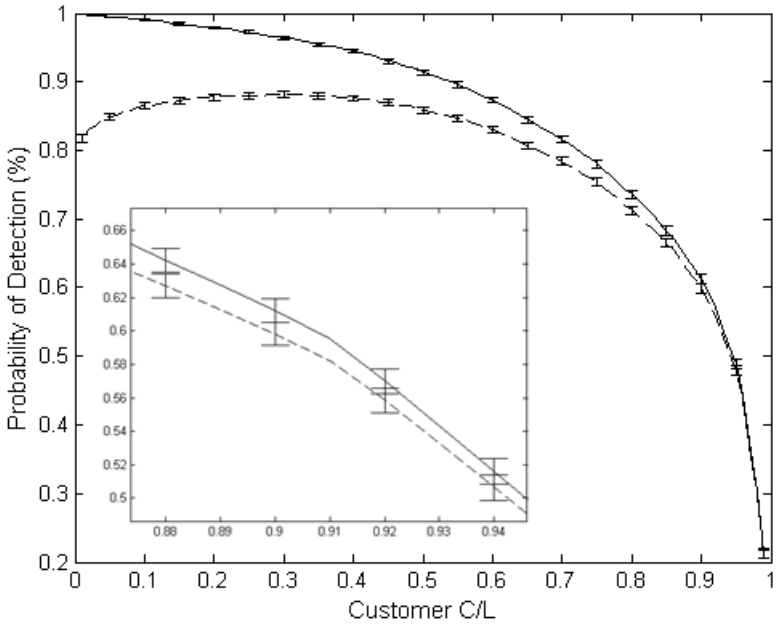



Figure 70. *POD* comparison for the control user (solid) and the always user (dashed) based on the application dataset of 50,220 forecast-observation pairs. The difference between the users becomes insignificant beyond *C/L* 90% (inset).

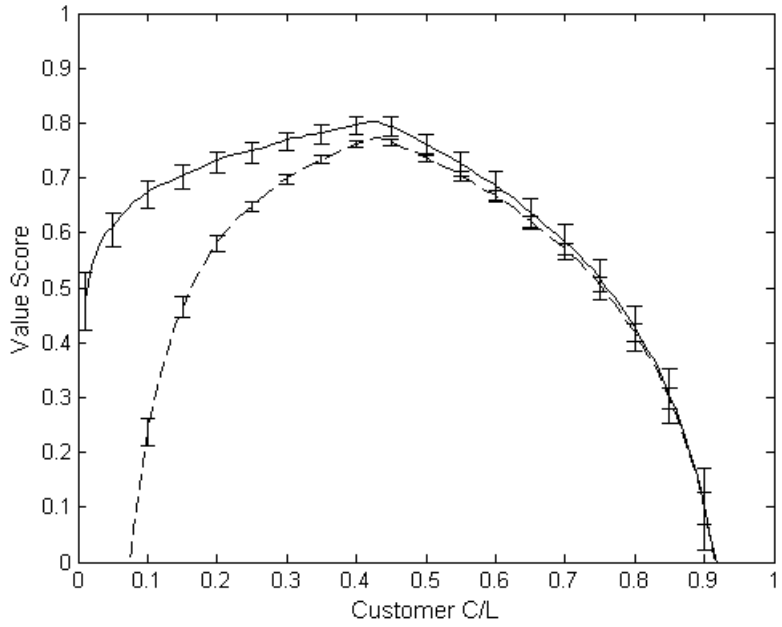



Figure 71. Optimal *VS* comparison for the control user (solid) versus the random user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

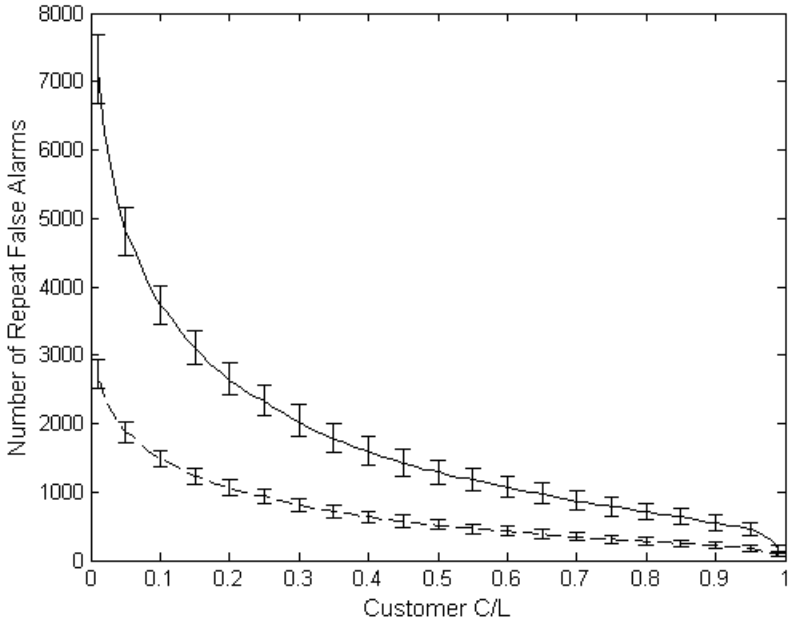



Figure 72. Repeat false alarm comparison for the control user (solid) versus the random user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

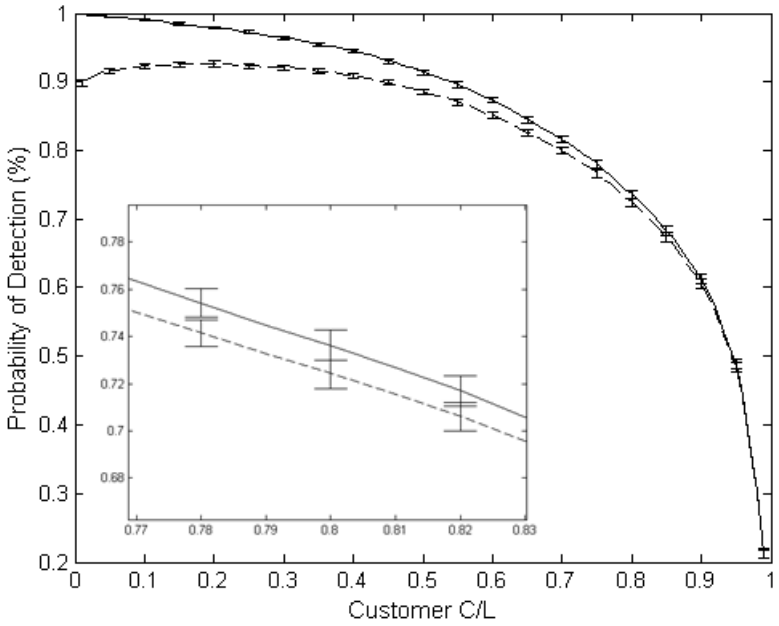



Figure 73. *POD* comparison for the control user (solid) and the random user (dashed) based on the application dataset of 50,220 forecast-observation pairs. The difference between the users becomes insignificant beyond *C/L* 80% (inset).

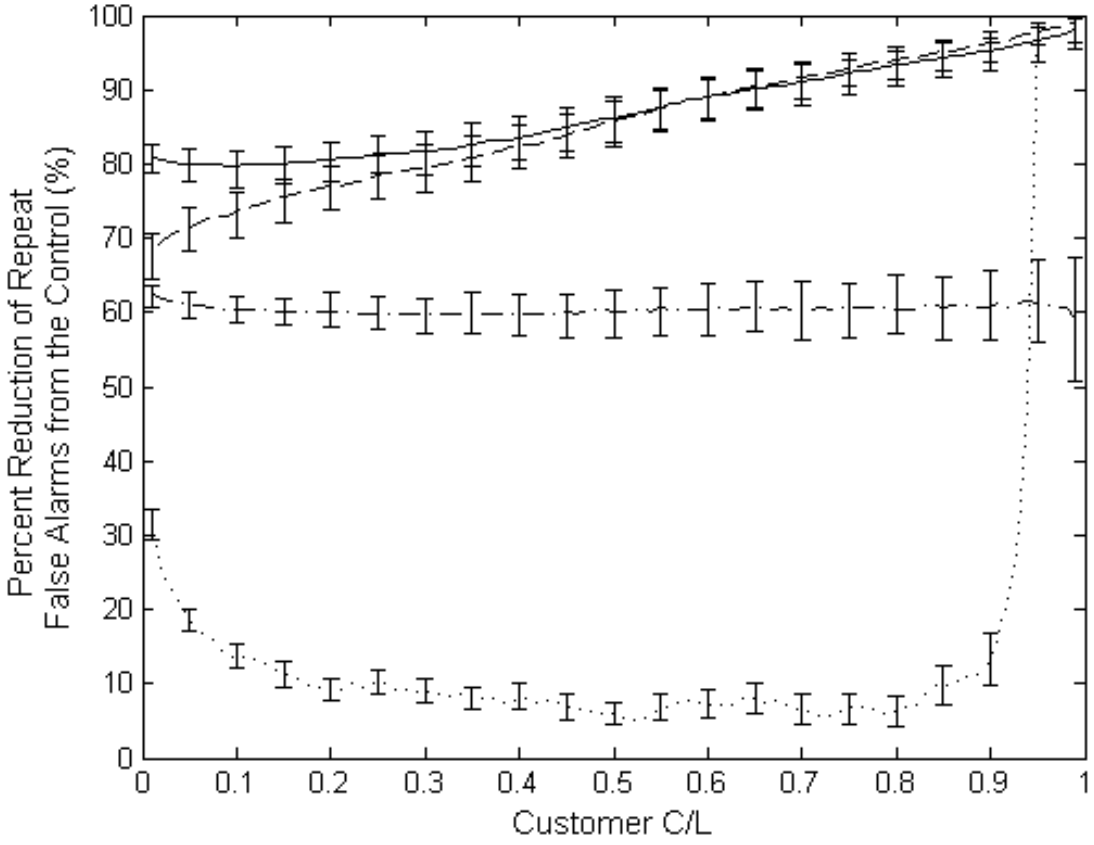



Figure 74. Percent reduction in repeat false alarms from the control user using alternate decision rules in Table 6. Shown are the percent reduction for the optimal (solid), conceptual model (dashed), random (dot-dashed) and brash (dotted) users. The always user provided 100% reduction at all *C/L* and is not displayed. Results are based on the application dataset of 50,220 forecast-observation pairs

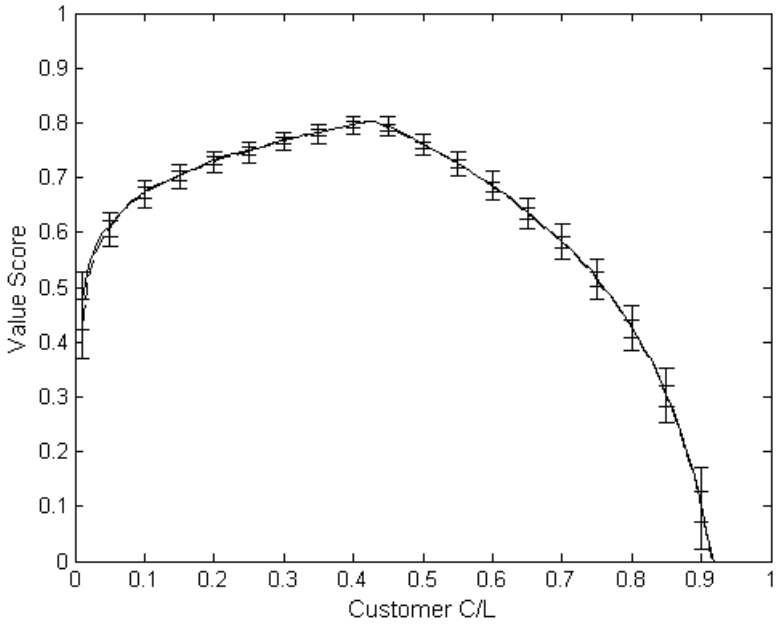



Figure 75. Optimal *VS* comparison for the control user (solid) versus the brash user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

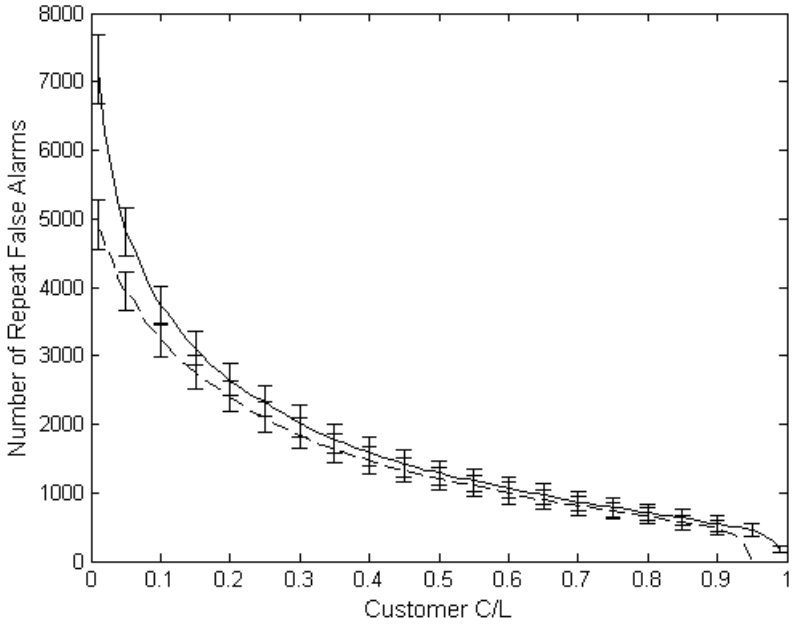



Figure 76. Repeat false alarm comparison for the control user (solid) versus the brash user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

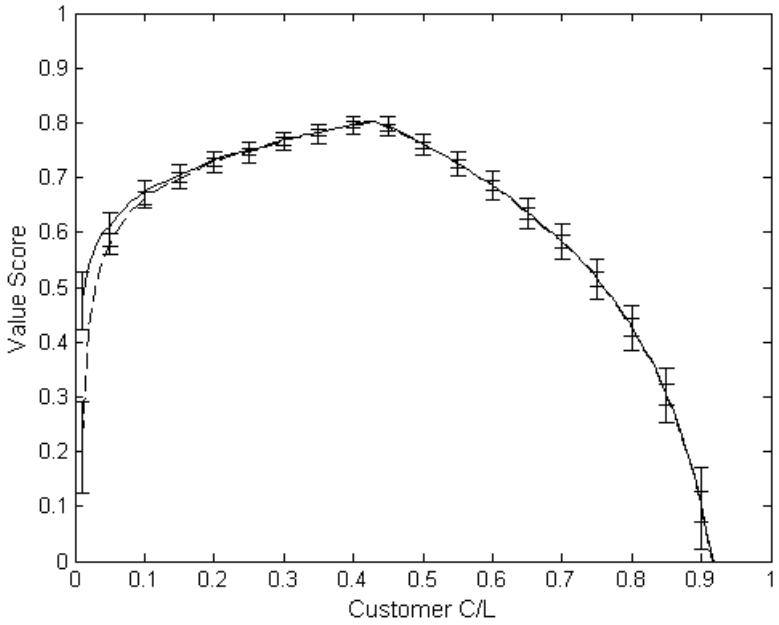



Figure 77. Optimal *VS* comparison for the control user (solid) versus the conceptual model user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

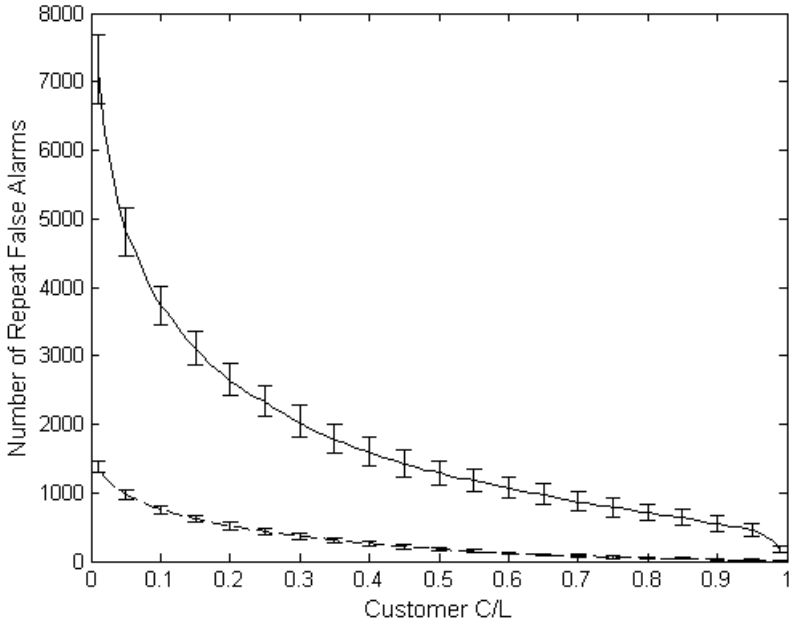



Figure 78. Repeat false alarm comparison for the control user (solid) versus the conceptual model user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

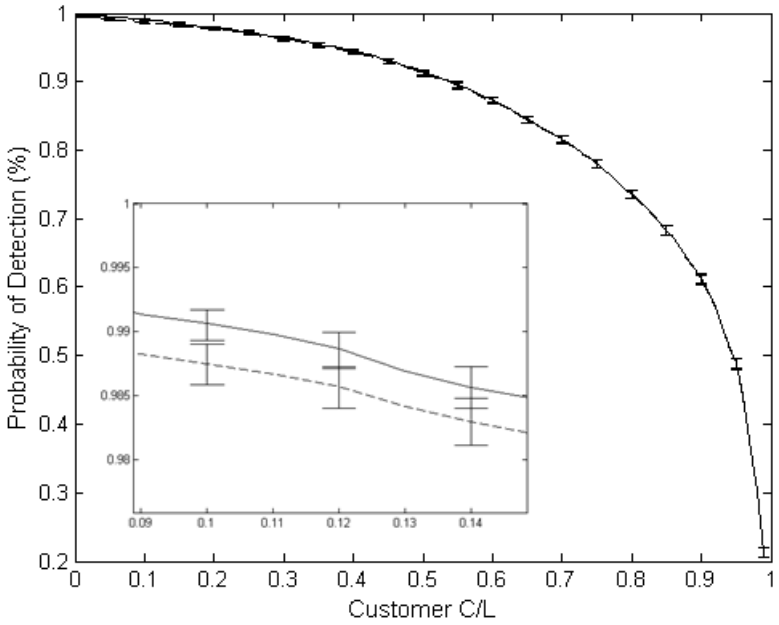



Figure 79. *POD* comparison for the control user (solid) and the conceptual model user (dashed) based on the application dataset of 50,220 forecast-observation pairs. The inset indicates that the difference between the users becomes insignificant beyond *C/L* 12%.

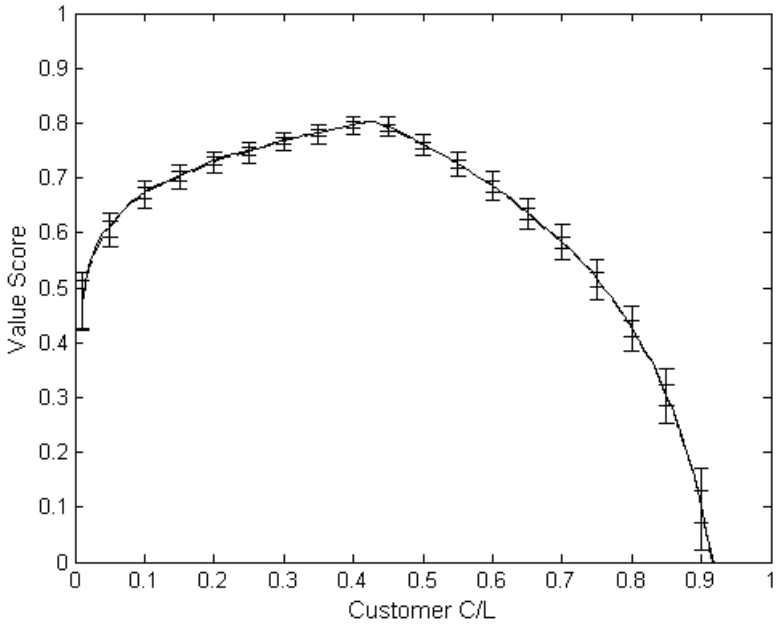



Figure 80. Optimal *VS* comparison for the control user (solid) versus the optimal user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

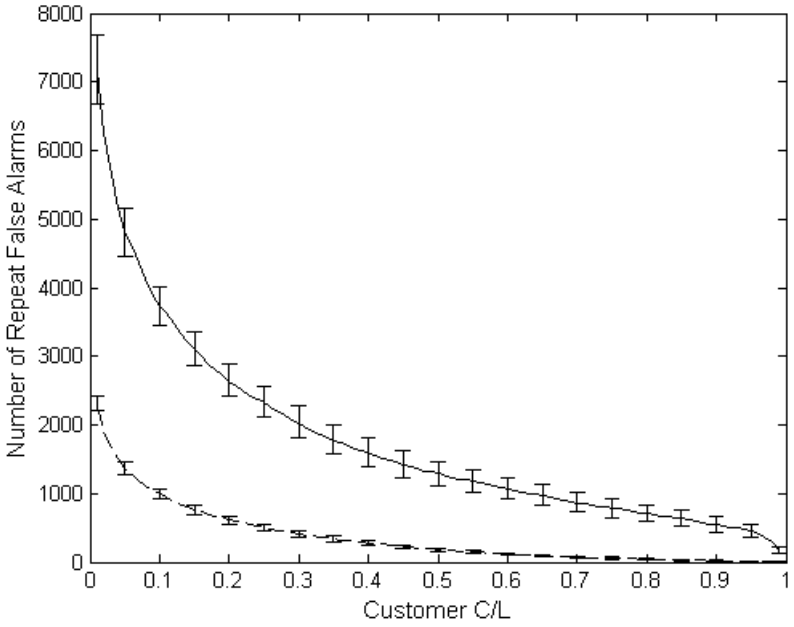



Figure 81. Repeat false alarm comparison for the control user (solid) versus the optimal user (dashed) based on the application dataset of 50,220 forecast-observation pairs.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. CONCLUSIONS

## A. SUMMARY

The primary tool for weather forecasters today is the NWP model, which provides a detailed forecast that unfortunately contains significant uncertainty (i.e., random error) due to analysis and model errors. An ensemble prediction system (EPS) generates a flow-dependent estimate of that uncertainty to provide information critical to optimal decision making. An ideal EPS will account for all sources of uncertainty associated with a particular deterministic modeling system. Today's EPSs use a finite number of ensemble members and inadequate representation of the uncertainty associated with the initial conditions and model design. These deficiencies result in errors in the ensemble forecast PDF, thus measures of forecast uncertainty will be incorrect, including forecast probability specific to an event criterion. Thus, there is uncertainty in the estimation of forecast uncertainty, a phenomenon termed ambiguity, which can negatively impact the ability to optimize decisions. Ambiguity is the uncertainty surrounding the forecast probability, which can be described by a distribution of forecast probability values, referred to as an ambiguity distribution (NRC 2006; Eckel and Allen 2009).

Ensemble forecasts can have high value in the decision making process. Numerous studies have shown the value of using probabilistic decision inputs over using deterministic or climatological information in the cost-loss (*C/L*) decision framework (e.g., Katz and Murphy 1997; Richardson 2000; Palmer 2002; Zhu et al. 2002). However, the possible additional value of using information about ambiguity has not been considered. In situations where the decision input is unclear, (due to ambiguity), an objective estimate of the ambiguity may be valuable to the user.

The three objectives of this research were to: (1) understand the mechanisms behind the evolution of ambiguity associated with an ensemble forecast, (2) validate objective estimates of ambiguity associated with an EPS, and (3) explore methods of applying the ambiguity information to add value in decision making. All three objectives were accomplished using an EPS based on a low-order, chaotic dynamical system, where

our aim was to follow state-of-the-art practices in designing the low-order EPS so findings would reflect the performance of real-world, operational EPSs. Additionally, real-world EPS data was used for exploring value in objective #3.

To explore the research objectives, we used the low-order, chaotic dynamical system first introduced by Lorenz (1996) as a suitable proxy for the atmosphere. The system (L96) describes the evolution of variables on two distinct scales (Lorenz 1996; Wilks 2005). The small-scale variables are unresolved and thus parameterized in a model of the system (L96M) using a stochastic parameterization, providing a model with random error. Data assimilation for the control analysis was accomplished using a perturbed-observation Ensemble Kalman Filter (EnKF) scheme. The L96M EPS used random draws from the EnKF members as its suite of initial conditions (IC). Model deficiencies were simulated in the EPS using the perturbed parameter approach applied through the stochastic parameterization, which randomly varied the parameter value for each member at every time step. For verification, ground truth was the solution from the complete L96 system.

Ambiguity was estimated using three different techniques. The first technique, ensemble-of-ensemble (EoE), consisted of running multiple, parallel EPSs (constituents) for the same forecast case. The IC and model perturbations were varied within each constituent's EPS, resulting in a spectrum of equally plausible ensemble forecast PDFs and a forecast probability PDF (i.e., ambiguity distribution) for any particular event at a given lead time. The EoE dynamically captures the EPS limitations (i.e., limited sampling and inadequate simulation of uncertainty), reflecting the EPS output's sensitivity to the flow-dependent deficiencies in the perturbations associated with different regions in the model attractor. Since EoE is an impractical approach to estimating ambiguity, we also used two practical ambiguity estimation techniques, calibrated error sampling (CES) and randomly calibrated resampling (RCR). These techniques created ambiguity estimates using the long-term, average error characteristics of the first two moments in the ensemble PDF, mean error of the ensemble mean ($ME_{\bar{e}}$) and fractional error in ensemble spread ($\sigma'$), as proxies for the relationships among EoE constituent PDFs.

The CES method took two forms, $CES_G$ (global) and $CES_L$ (local). $CES_G$ used 50,000 sets of random draws from the long-term, average distributions for $ME_{\bar{e}}$, $\sigma'$ and ensemble spread to create a distribution composed from 50,000 possible values of true forecast probability for any value of calibrated forecast probability. $CES_G$ produced a bulk (generic) ambiguity estimate, independent of ensemble spread, that could come from any event since the EPS characteristics are taken as the same across the entire attractor. $CES_L$ provided a somewhat flow-dependent ambiguity estimate by following similar processing as $CES_G$ but for specific values of ensemble spread. Thus the $CES_L$ ambiguity estimate for a certain calibrated forecast probability value is different for different values of ensemble spread. (Note: $CES_L$ was actually developed in response to the evolution and validation discoveries in this research so was omitted from validation but applied in the value studies).

RCR produces a somewhat flow-dependent ambiguity estimate using bootstrap resampling of the ensemble members. A distribution of 10,000 possible values of forecast probability is produced by generating 10,000 different versions of the members at each forecast point by resampling with replacement. This process accounts for limited sampling of the true forecast PDF due to the finite number of members in the EPS, and the ambiguity estimate is dependent on the number of members (i.e., fewer members give higher ambiguity). For RCR, each set of resampled members is calibrated using random coefficients drawn from the distributions for mean and fractional error of the ensemble PDF, which removes systematic error and brings in solutions missed by the original members due to EPS deficiencies. The RCR ambiguity distribution is generally wider than would be found using resampling alone.

The evolution of ambiguity was explored using the EoE, as it produces our best estimate of ambiguity. Ambiguity was found to be highest early in the forecast period and then decrease quickly during peak forecast error growth. We found the primary influences on ambiguity magnitude to be the variability in location of the constituents' PDFs and the uncertainty (i.e., ensemble spread) of the constituents' PDFs. When the ratio of variance in constituents' locations to ensemble variance is large (typically early

in the forecast), large differences in forecast probability may be seen (i.e., high ambiguity) since changes in probability density relative to an event threshold are more sensitive to location changes when spread is low (Figure 51, page 147). Later in the forecast, the disproportionately larger increase in ensemble variance (due to error growth) compared to the separation between constituents' PDFs results in a narrower range of forecast probabilities (i.e., low ambiguity), as probability density shifts amongst the constituents are similar. Since ensemble variance plays a significant role in the production of ambiguity, our results suggest that ambiguity generally evolves from high to low values as a result of the typical increase in ensemble spread with forecast lead time. Of course, it is possible on a case-by-case basis for an ensemble forecast to exhibit small spread at any lead time resulting in large ambiguity. However, the general conclusion is that ambiguity is a function of both ensemble spread and the sensitivity of forecast probability estimates to errors in PDF location. The irony of this finding is that sharper ensemble PDFs are generally considered to reflect better performance, but ambiguity can be greatly increased by the sensitivity to errors in location with sharper forecasts.

Validation was performed using aggregated $CES_G$ and RCR ambiguity distributions built over many locations on the L96M attractor to determine the overall effectiveness of the estimates in comparison to EoE. However, we could not validate the estimation methods' ability to consistently capture the location of the EoE ambiguity distribution since a random error in location generally exists between EoE and the $CES_G$ and RCR distributions. Validation showed how well $CES_G$ and RCR captured the variance of the EoE ambiguity distribution. Comparisons made using the total ambiguity [Equation (26), page 61] of each method's aggregated ambiguity distributions indicated the following trends:

- The ambiguity distributions from the practical estimation techniques appeared to perform very poorly at early forecast lead times with total ambiguity differences near 30%, but each showed improvement with time;

- The largest differences in total ambiguity appear to have occurred with mid-range forecast probability values;
- The $CES_G$ total ambiguity was too narrow in relation to the aggregated EoE ambiguity distributions regardless of forecast lead time or forecast probability value tested;
- The RCR total ambiguity was too narrow during the early forecast lead times, but then transitioned to become slightly too wide later in the forecast for most of the forecast probability values tested.

The apparent disparity in performance of the $CES_G$ and RCR estimates at mid-range and extreme forecast probability values is simply a result of the lower and upper bounds (i.e., 0% and 100%, respectively) confining the range of possible forecast probability values. In general, we expect to see tighter ambiguity distributions for the extreme forecast probability values, thus total ambiguity is naturally smaller. Additionally, as the expected value of the ambiguity distribution approaches either extreme, the total ambiguity difference between the EoE and the CES or RCR estimates used for validation becomes more one-sided reducing the difference. For example, when the expected value approaches 0%, the lower bounds of each estimation method's ambiguity distributions become more similar, thus differences in total ambiguity are found primarily in the upper bounds.

We found a leading factor in the under-spread $CES_G$ ambiguity distributions was the absence of flow-dependent ensemble spread. The typical ensemble variance used when estimating the forecast probability values was near the long-term average, which in many cases would likely be too high compared to the flow-dependent variance, producing a more narrow ambiguity distribution. Additionally, the variance of the $ME_{\bar{e}}$ error distribution used to create the $CES_G$ ambiguity distributions was inadequate, particularly at the early forecast lead times, thus the $CES_G$ sample forecast PDFs were not sufficiently separated to produce a wide enough ambiguity distribution. Therefore, $CES_G$ often underestimates the total ambiguity.

Similar to $CES_G$, the RCR ambiguity distributions were highly under-spread likely due to the low variance of the $ME_{\bar{e}}$ error distribution. Even though RCR considers the flow-dependent ensemble spread, the RCR PDFs could not adequately separate to generate the ambiguity levels provided by EoE. RCR total ambiguity recovered later in the forecast due to the improvement in variance of the $ME_{\bar{e}}$ error distribution as well as its application of flow-dependent ensemble spread. The random calibration was likely the cause for slightly excessive ambiguity estimates later in the forecast.

In general, ambiguity found using $CES_G$ and RCR evolved similarly to that of EoE (i.e., from high to low values), but the magnitude of the ambiguity early in the forecast was notably lower compared to EoE. We concluded that the variance in $ME_{\bar{e}}$ (as used by $CES_G$ and RCR) underestimated the variation in possible ensemble forecast PDF locations found using the EoE (i.e., the constituents' PDFs), thus limiting the variance of the ambiguity distributions, especially early in the forecast. However, we found the practical ambiguity distributions to be reasonably accurate estimates of the total ambiguity once error growth exceeded approximately 10% of the climatological variance. In cases where error growth is below 10%, ambiguity generally increases as the ensemble forecast PDF gets sharper, but for sharper PDFs, ambiguity is less often a factor since any given event is more certain (i.e., forecast probability closer to 0% or 100%). Therefore, we conclude that the $CES_G$ and RCR ambiguity distributions are likely good enough to provide valuable information to the decision process.

This research introduced two approaches for attempting to add value to the decision making process using objective ambiguity estimates. The first approach, uncertainty-folding, combines the first- and second-order uncertainty information to once again give the user a single probabilistic decision input based on the weather information. We performed uncertainty-folding using ambiguity distributions from the EoE, $CES_G$ and RCR estimation techniques. We also tested a grand ensemble where all constituent members for a single EoE forecast case were combined to produce a single forecast probability value. These four decision input sources were compared in relation to the

value provided by basing decisions on the control ensemble forecast probability alone. Results for two event thresholds (representing a common and a rare event) were examined.

For both events, the integrated optimal value score (*VS*) found using the EoE and the grand ensemble showed improvement over the control ensemble forecast, but results were only significant for the common event at lead times beyond maximum error growth. The grand ensemble and EoE generally performed the same, indicating that resources may be better spent reducing ambiguity by running a larger EPS than estimating ambiguity with an impractical approach like EoE. The scores found using uncertainty-folding with the $CES_G$ and RCR were generally not significantly different from the control ensemble. Since the $CES_G$ and RCR ambiguity distributions are centered on the control forecast probability, the probability value computed using uncertainty-folding will not vary greatly from the control value, which prevented significant improvement in value. Additionally, random error in the location of the practical techniques ambiguity distributions produced errors when combining the first- and second-order uncertainty, likely reducing the value of the decision input in normative decision making. Thus uncertainty-folding may not be a useful approach to garner value from ambiguity since it only works well for EoE, the impractical method of ambiguity estimation.

For the second method used to attain value using the ambiguity information, we looked at improving secondary criteria important to the decision-maker beyond the primary value (tied to minimizing total expense). The example secondary criteria considered was repeat false alarms, so the objective was to use the ambiguity information to significantly reduce the number of repeat false alarms while maintaining the primary value (measured by optimal *VS*, as well as probability of detection and probability of missed detection) associated with normative decision making within the *C/L* scenario. Several user decision rules were studied using real-world ensemble forecast data from National Center for Environmental Prediction's Global Ensemble Forecast System, where the different users were allowed to reverse the current decision of taking protective action if and only if a false alarm had just occurred at the same location and their decision criteria was met.

Of all the decision rules studied, the two that considered ambiguity (via an overlap threshold) when reversing decisions outperformed the others at significantly reducing repeat false alarms while maintaining the primary value. The overlap is the proportion of the ambiguity distribution indicating a different decision than the normative input; thus the overlap threshold represents the value of overlap at which the user reverses decisions. We saw the best overall performance by the user who followed the optimal overlap threshold. Developed from a training dataset, the optimal overlap threshold for each *C/L* was the lowest threshold giving the greatest reduction in repeat false alarms that resulted in no significant reduction in primary value. Although the conceptual model had significantly fewer repeat false alarms than the optimal user for low *C/L*, the optimal user faired better in regards to primary value than the conceptual model since it prevented excessive reversals, thus avoiding a large increase in misses (i.e., expense). For mid-range and high *C/L*, there was no significant difference between the optimal and conceptual model users.

The results clearly show that we can attain tremendous improvements to secondary criteria by employing an objective ambiguity estimate in decision making. Moreover, we were able to train our decision process based on past performance to optimally select an overlap threshold at each *C/L*. Using the flow-dependent $CES_L$ estimates for this study (instead of $CES_G$ which inherently underestimated ambiguity), likely played a large role in attaining significant value for the secondary criteria.

## B. FUTURE RESEARCH

The results presented in this research suggest several areas of future research, the first of which is to perform a validation study using the $CES_L$ estimation method. The refinements made to include flow-dependence are likely to improve ambiguity estimation for $CES_L$ compared to $CES_G$, especially at later forecast lead times when the low $ME_{\bar{e}}$ variance played less of a role in degrading the estimates. However, the inclusion of flow-dependent ensemble spread at early times may allow $CES_L$ to produce a wider range of

forecast probabilities, thus improving its estimates compared to EoE. Like RCR, $CES_L$ validation will require aggregates of ambiguity estimates specific to EoE forecast cases.

The next subject for research is continued investigation of the method used to determine the variance of the error distributions, i.e., sub-setting of the long-term verification dataset. The variance of the error distributions is obviously dependent on the size of the subset, and properly determining subset size is non-trivial. For this research, sub-setting was based on complete EPS runs to capture flow-dependent error characteristics, which appears to be inadequate.

A related area of future research involves the implications of the spread-skill relationship in $CES_L$. $CES_L$ ambiguity estimates found using the domain averaged $ME_{\bar{e}}$ variance in this research ignored the spread-skill relationship, but obtained reasonable and ultimately valuable estimates. However, for a well-calibrated EPS, the correlation between ensemble spread and ensemble mean error variance is nearly perfect (as seen in a binned spread-skill plot), so $CES_L$ should perhaps use that information in estimating ambiguity. In that case, ambiguity would be similar regardless of the ensemble spread value since the variability in location error is proportional. Additionally, ambiguity would be much larger and overestimated in most cases given such a large ratio in the variances. Research is thus needed to resolve this contradiction.

Further research should be conducted using the correction to $ME_{\bar{e}}$ variance provided by the ratio in Figure 55 (page 150) (i.e., comparison of variance in constituent location to $ME_{\bar{e}}$ variance) with the $CES_L$ method. Greater improvements are expected than those seen with $CES_G$ in Figure 62 (page 166), due to the flow-dependence of $CES_L$. If corrections to the total ambiguity are nearly perfect at all lead times, further investigations may be performed using a different low-order model to determine if a general relationship (i.e., correction) exists between $ME_{\bar{e}}$ variance and constituent location variance that may be used for higher order models.

Several subjects for future research involve utilization of the ambiguity information in decision making. While using the practical methods with uncertainty-

folding may not have provided significant improvements in value using the optimal integrated *VS* (a combination of all users), it may be beneficial to perform a more thorough evaluation. Uncertainty-folding should be used with $CES_L$ and RCR ambiguity estimates for events at specific lead times, thus allowing analysis of the results for specific users (i.e., *C/L*) instead of integrating all users into the optimal integrated *VS*. Future uncertainty-folding studies should also include real-world EPS data.

We investigated just one of many possible secondary criteria, but future research in this area of value is nearly unlimited. Studying different secondary criteria entails developing methods to measure primary and secondary value, as well as determining methods for optimization of the decision process. The value of some secondary criteria may be hard to assess. For example, mission effectiveness (primary value) may be evaluated through battle damage assessment, but the intangible benefits such as improved morale (secondary value) that come with a successful mission are hard to quantify. In this case, the user may choose an alternate strike location with a greater chance of success to hopefully improve morale. Additionally, we looked at a single secondary criterion in isolation, but it may be equally important to the customer to consider multiple criteria (e.g., repeat false alarms and repeat misses).

# APPENDIX: FIGURE SEQUENCE DISPLAYING THE TIME EVOLUTION OF AMBIGUITY

This appendix includes figures referenced in Chapter IV.A showing the evolution of ambiguity with increasing forecast lead time for a single EoE forecast case of 100 constituents using an arbitrary $X_k$ variable. The ambiguity distributions were determined for each forecast lead time using an *X*-value event threshold that resulted in $E\left(\hat{p}_T\right) = 50\%$, thus the event threshold was different for each forecast lead time. The histograms of constituent forecast probability values were created using a class interval of 1% over the range 0%-100%. Constituent PDFs were generated using a normal fit to the *n* ensemble members in each constituent ensemble forecast. Note that the abscissa range is fixed for all figures in both (a) and (b), while the ordinate range may vary based on the data.

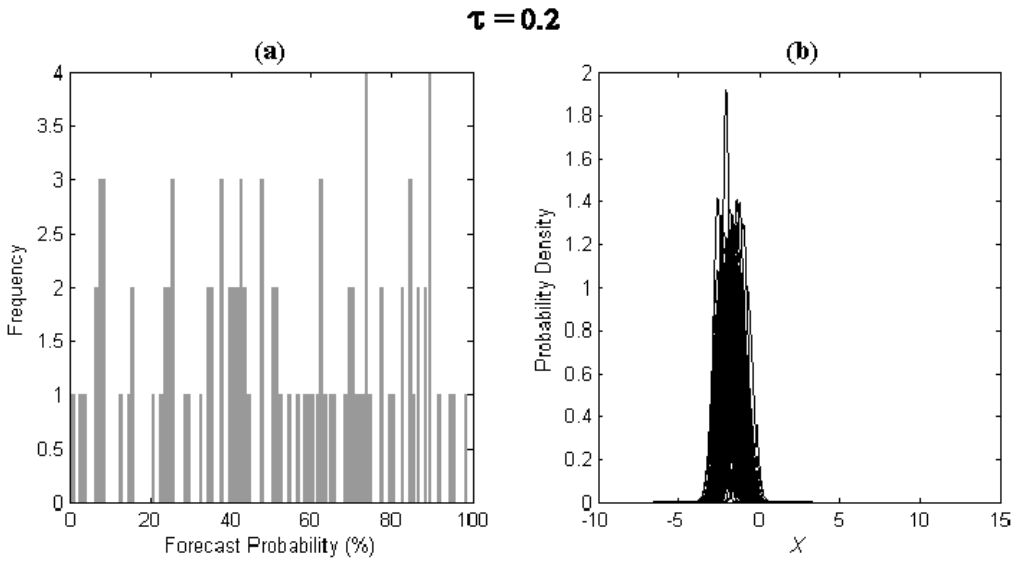
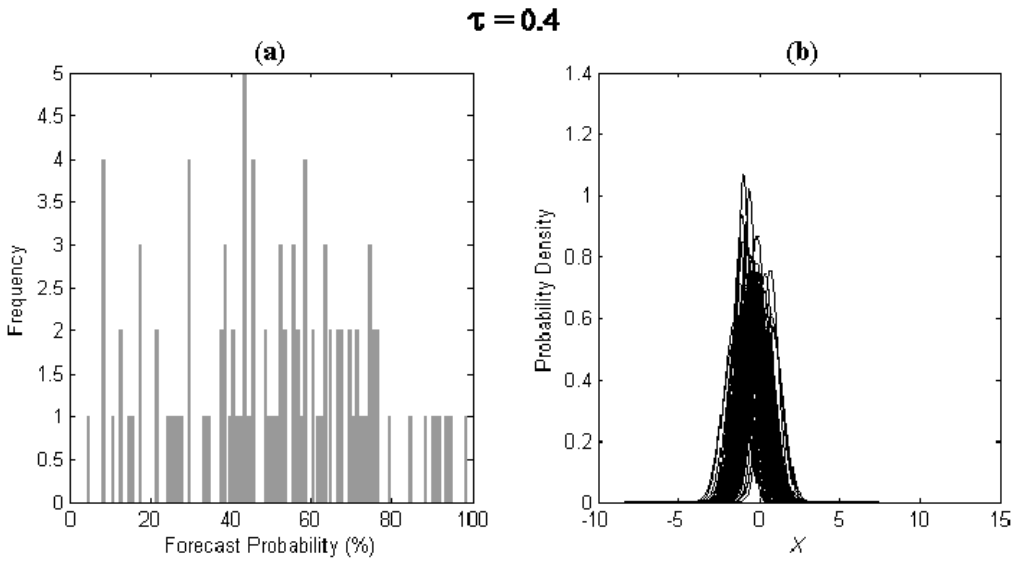
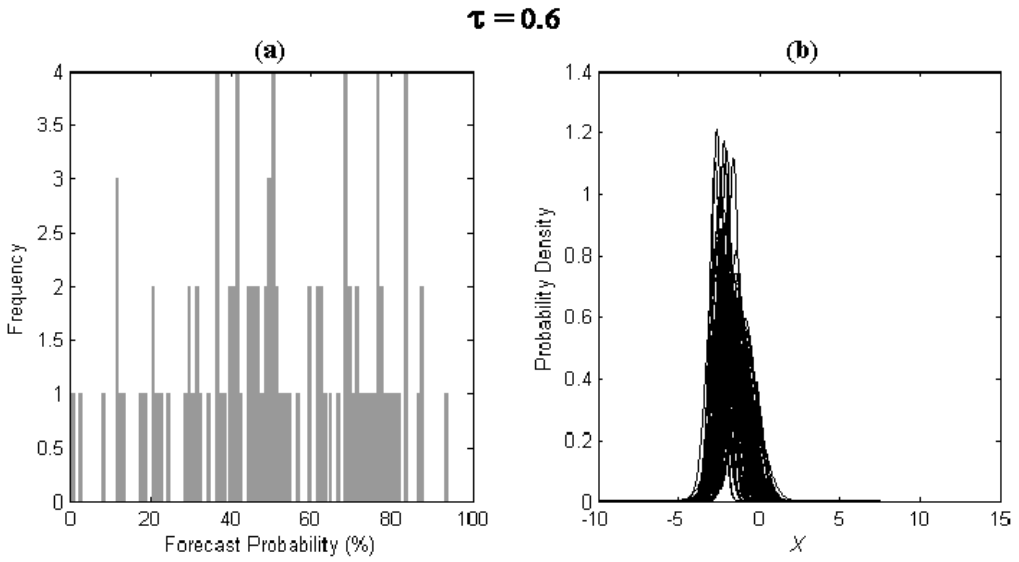


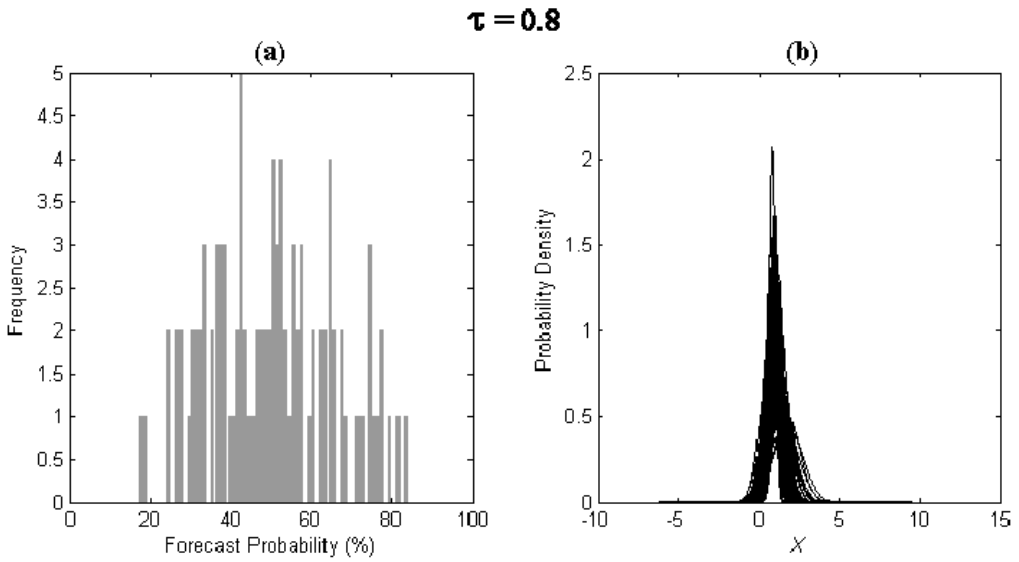

Figure 82. EoE ambiguity evolution showing (a) the histogram of constituent forecast probability values and (b) the constituent PDFs used to find each forecast probability for forecast lead times 0.2 to 5 at 0.2 increment (labeled at the top of each panel). The total ambiguity for this panel equals 85% (7% to 92%).



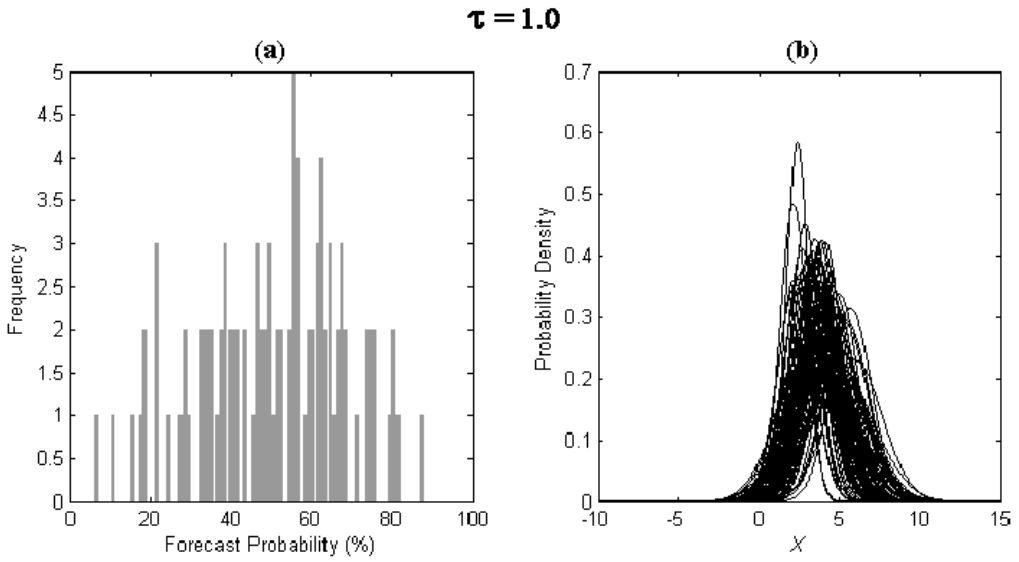

(Figure 82 continued.) The total ambiguity for this panel equals 75% (12% to 87%).



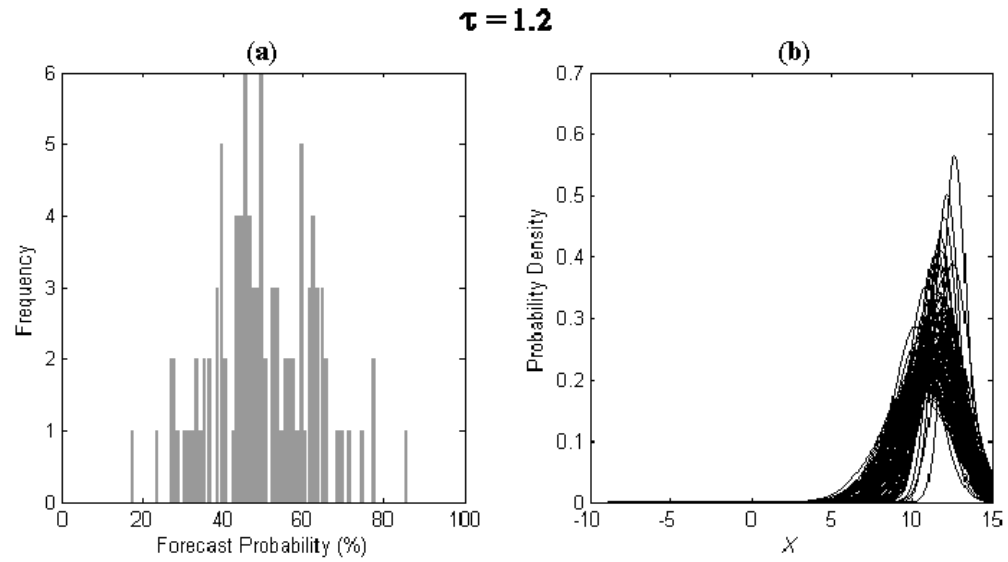

(Figure 82 continued.) The total ambiguity for this panel equals 74% (13% to 87%).



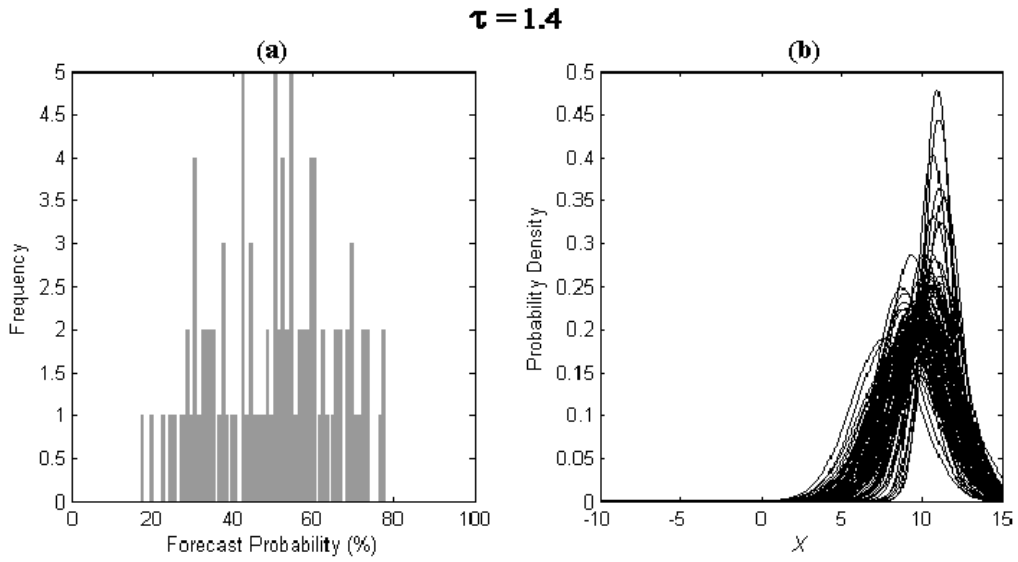

(Figure 82 continued.) The total ambiguity for this panel equals 49% (25% to 74%).



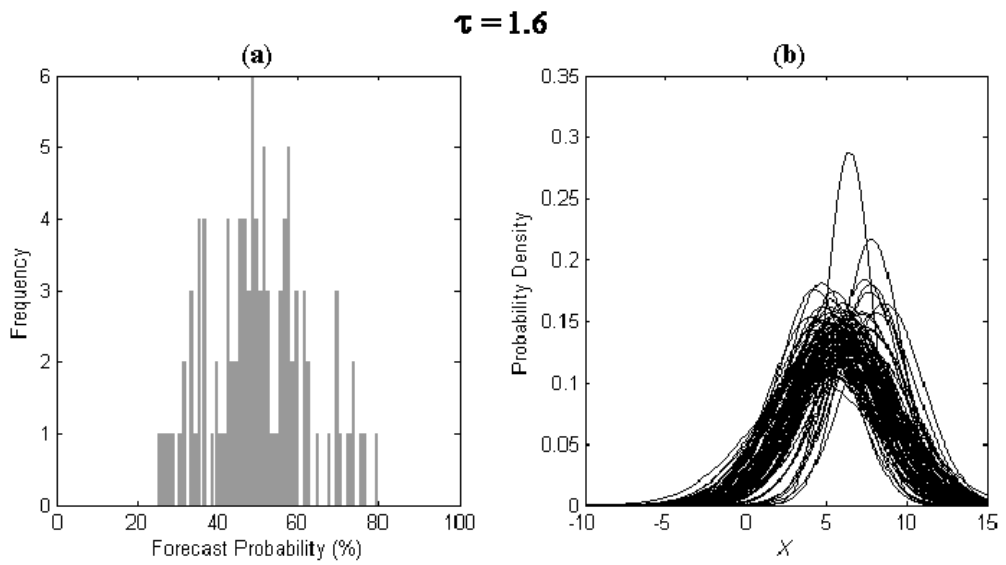

(Figure 82 continued.) The total ambiguity for this panel equals 58% (21% to 79%).



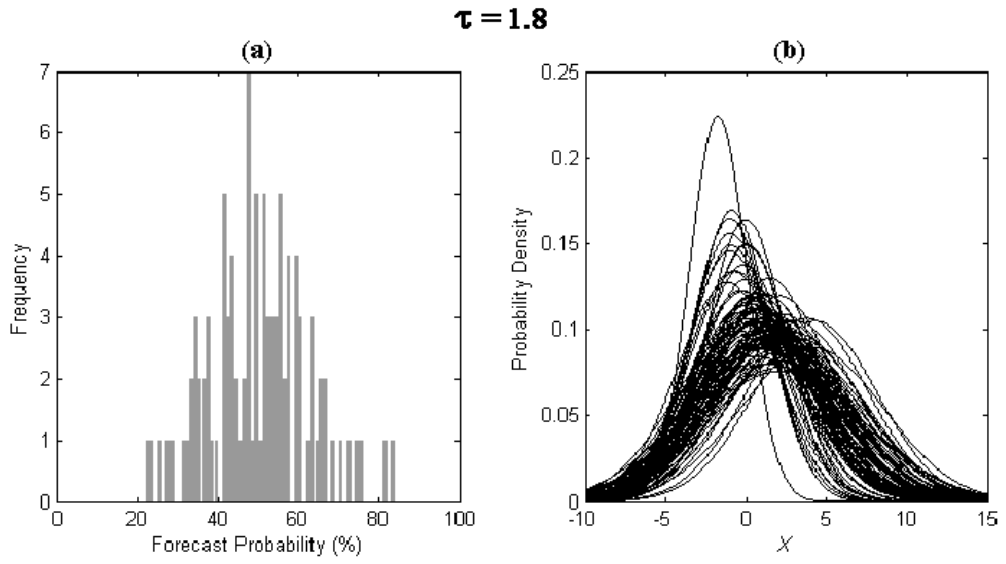

(Figure 82 continued.) The total ambiguity for this panel equals 41% (29% to 70%).



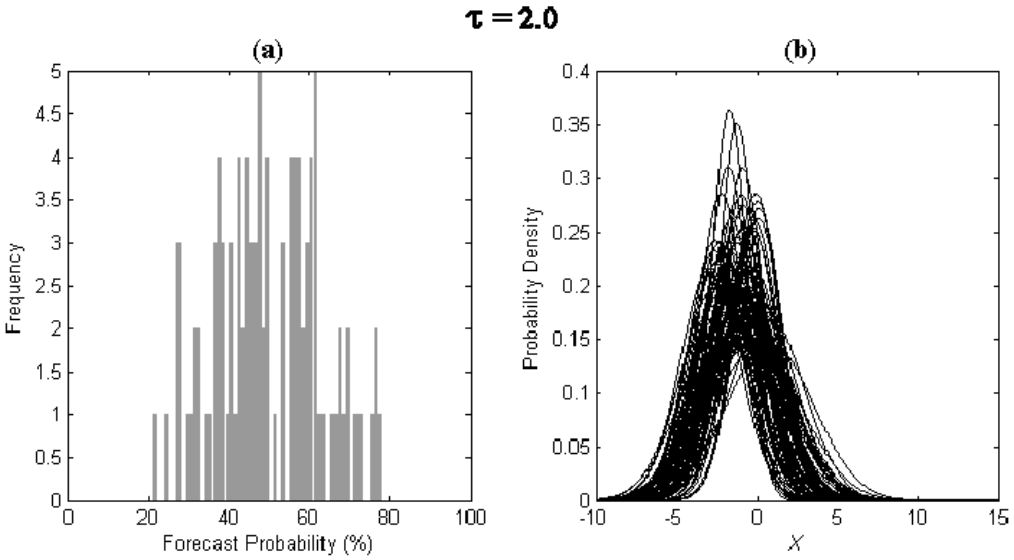

(Figure 82 continued.) The total ambiguity for this panel equals 48% (26% to 74%).



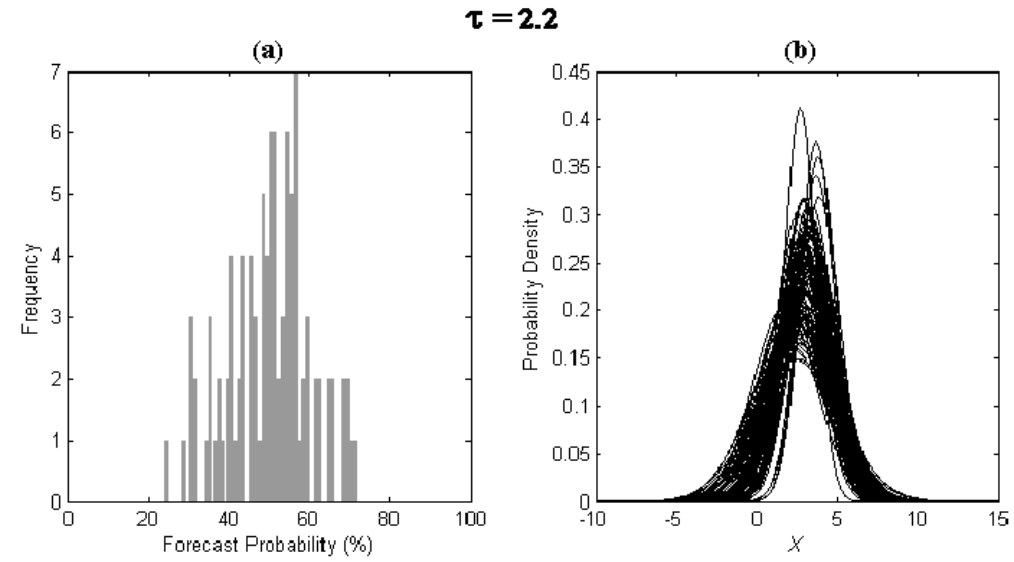

(Figure 82 continued.) The total ambiguity for this panel equals 40% (30% to 70%).



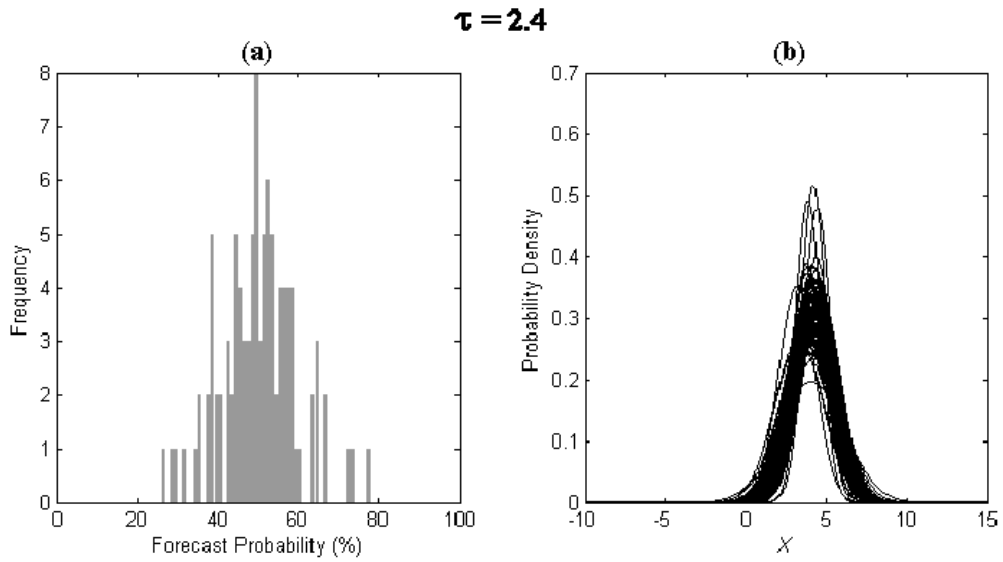

(Figure 82 continued.) The total ambiguity for this panel equals 40% (30% to 70%).



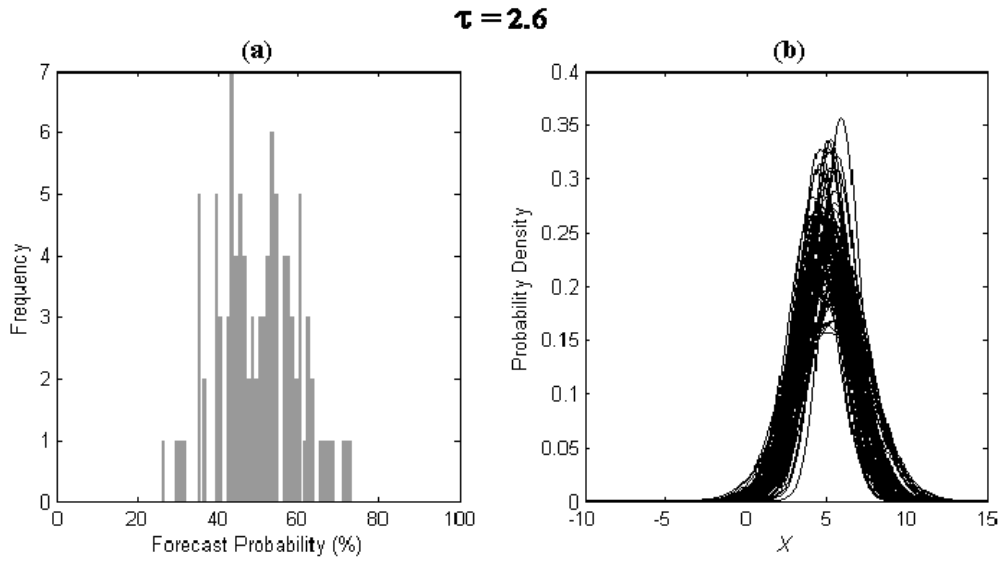

(Figure 82 continued.) The total ambiguity for this panel equals 42% (29% to 71%).



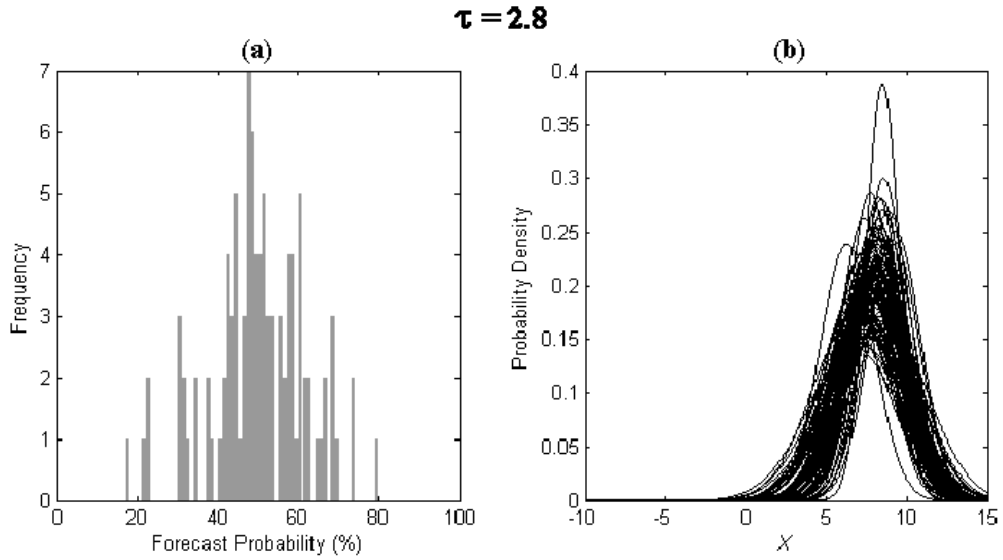

(Figure 82 continued.) The total ambiguity for this panel equals 34% (33% to 67%).



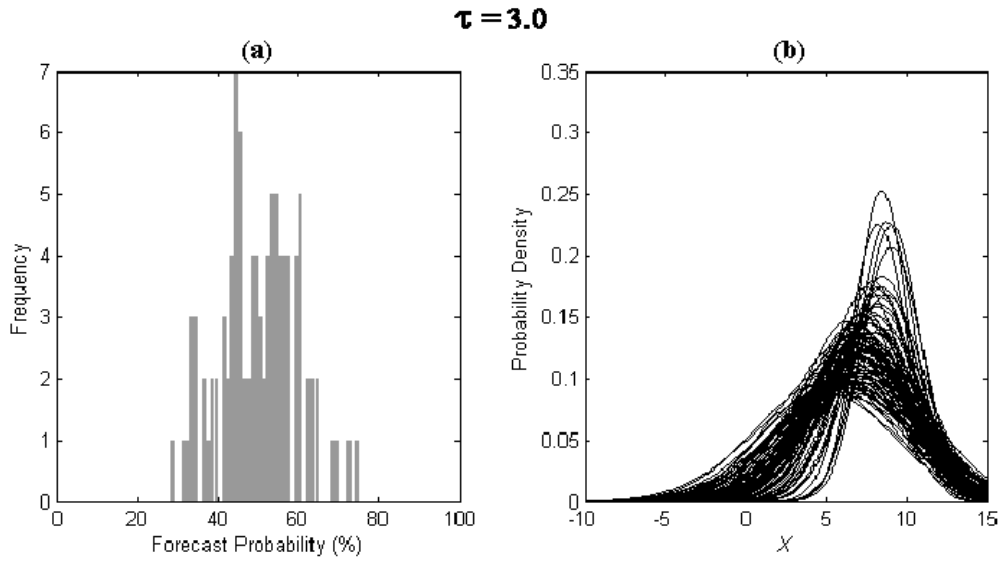

(Figure 82 continued.) The total ambiguity for this panel equals 31% (35% to 66%).



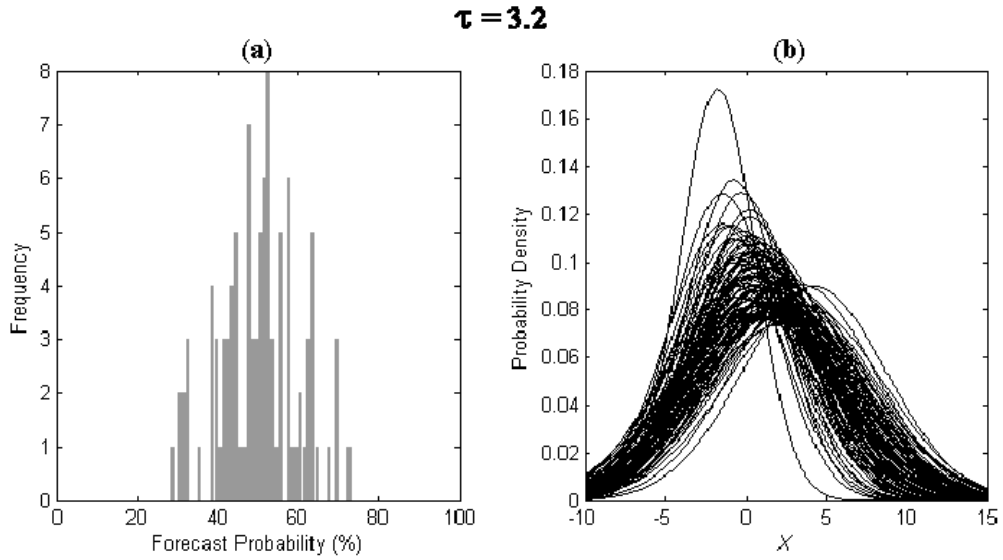

(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).



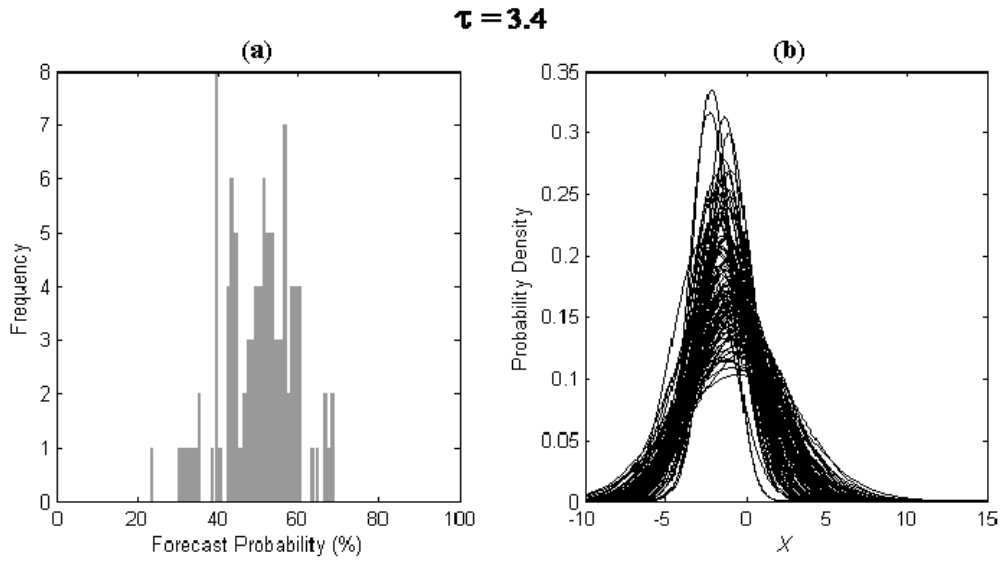

(Figure 82 continued.) The total ambiguity for this panel equals 38% (31% to 69%).



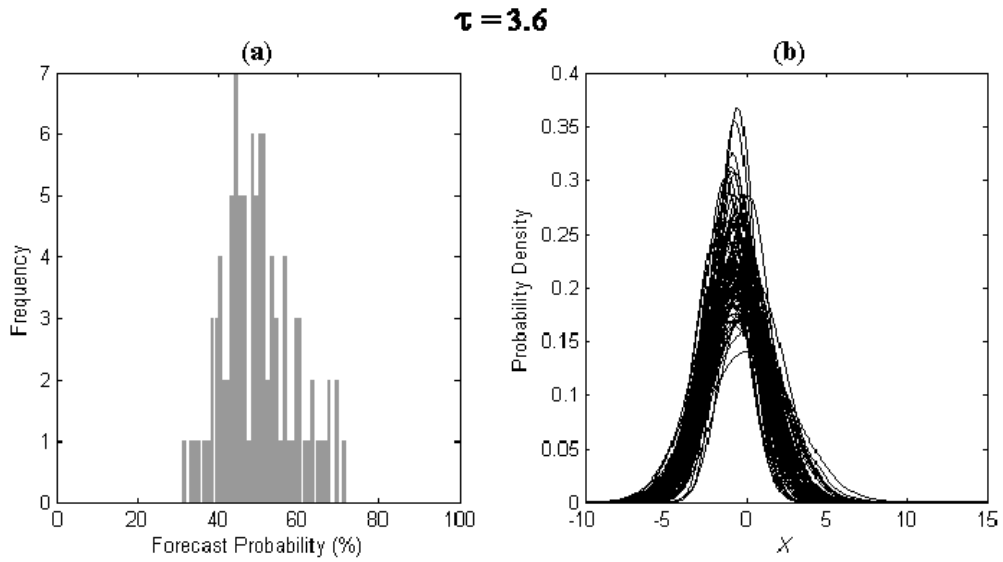

(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).



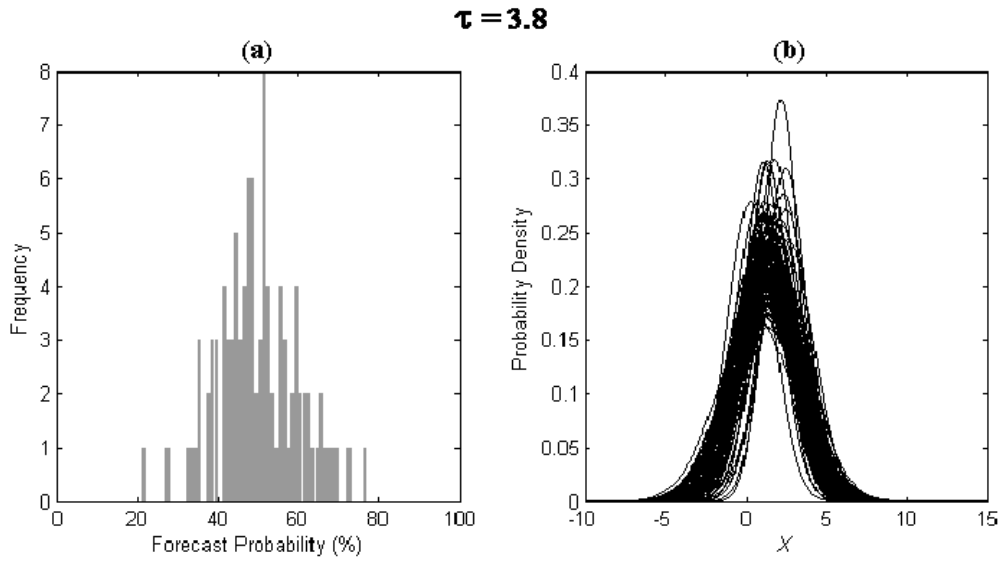

(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).



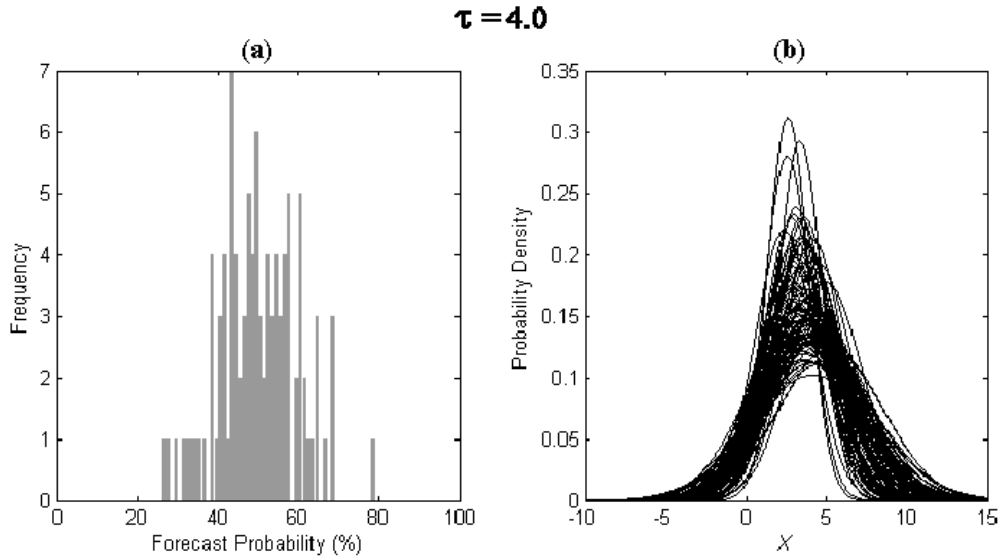

(Figure 82 continued.) The total ambiguity for this panel equals 30% (35% to 65%).



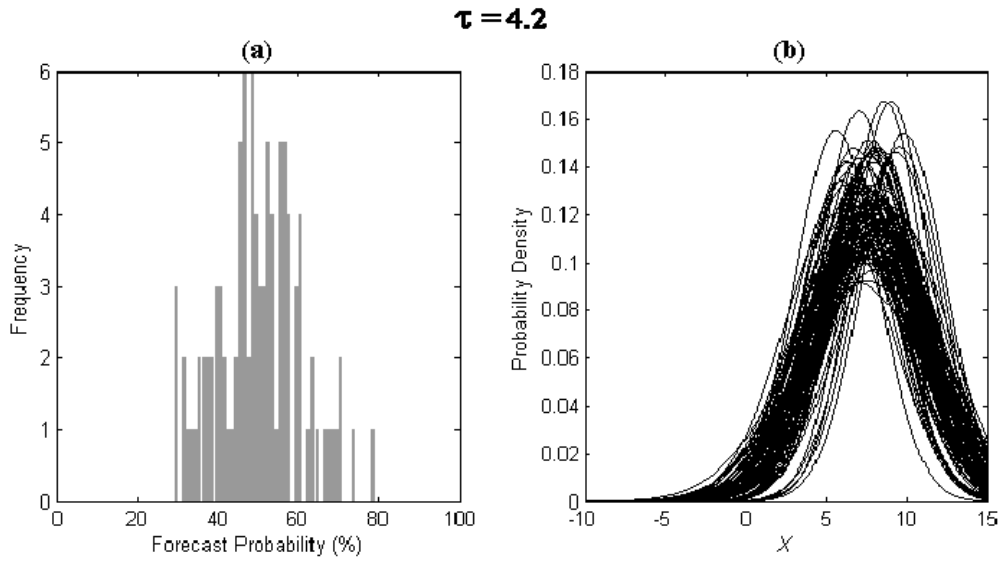

(Figure 82 continued.) The total ambiguity for this panel equals 28% (36% to 64%).



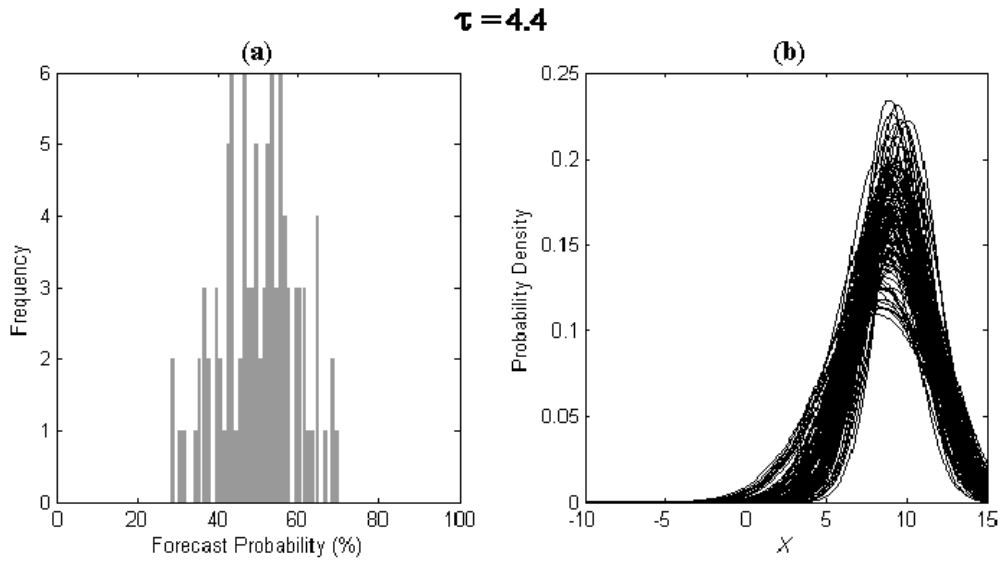

(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).



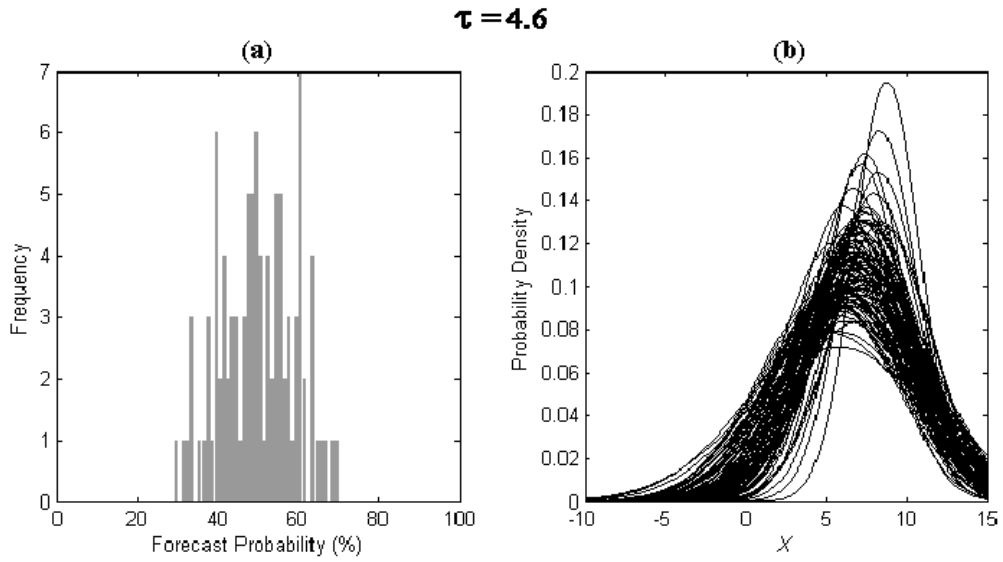

(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).



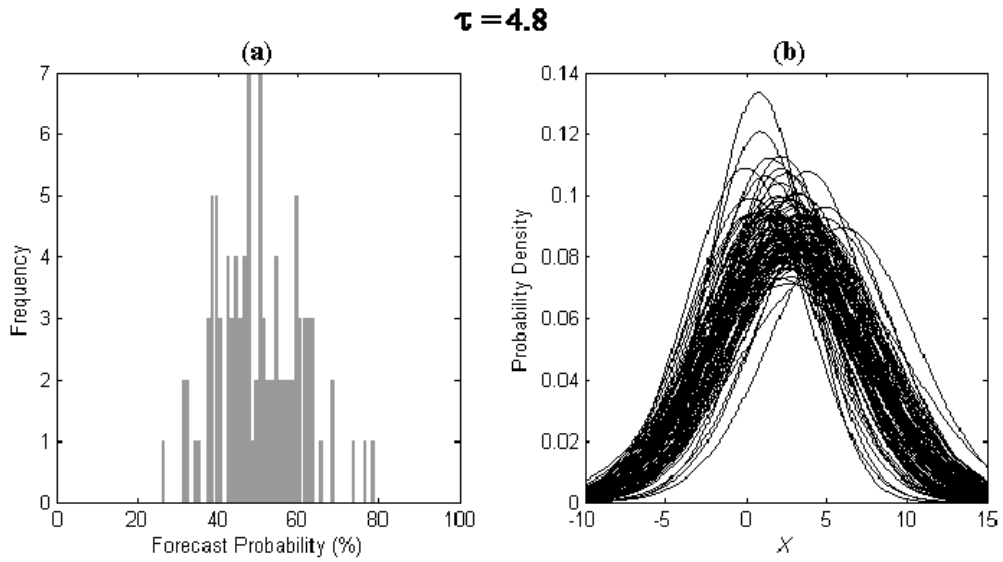

(Figure 82 continued.) The total ambiguity for this panel equals 34% (33% to 67%).



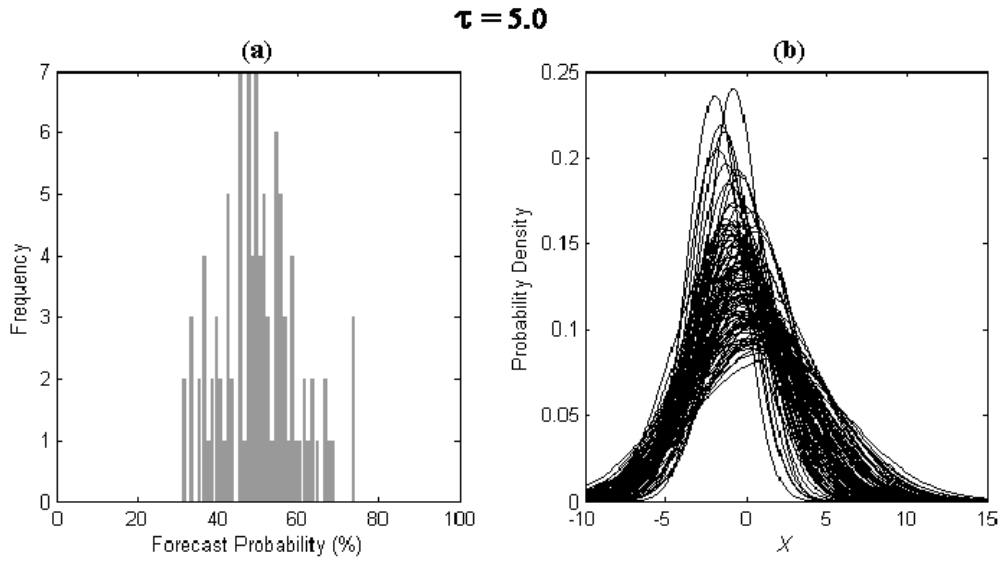

(Figure 82 continued.) The total ambiguity for this panel equals 30% (35% to 65%).




(Figure 82 continued.) The total ambiguity for this panel equals 30% (35% to 65%).




(Figure 82 continued.) The total ambiguity for this panel equals 34% (33% to 67%).




(Figure 82 continued.) The total ambiguity for this panel equals 32% (34% to 66%).

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF REFERENCES


Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Anderson, J. L., 1997: The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Wea. Rev.*, **125**, 2969–2983.

Anderson, J. L., 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634–642.

Anderson, J. L., S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.

Bowler, N. E., 2006: Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus A,* **58**, 538–548.

Brooks, H. E., C. A. Doswell, 1993: New technology and numerical weather prediction—a wasted opportunity? *Weather*, **48**, 173–177.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2907.

Burgers, G., P. Jan van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.,* **126**, 1719–1724.

Camerer, C., M. Weber, 1992: Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty,* **5**, 325–370.

Cohn, S. E., 1997: An introduction to estimation theory. *J. Meteor. Soc. Japan*, **75**, 257–288.

Descamps, L., O. Talagrand, 2007: On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Wea. Rev.,* **135**, 3260–3272.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Eckel, F.A., 2003: Effective Mesoscale, Short-Range Ensemble Forecasting. Ph.D. Dissertation, University of Washington Department of Atmospheric Sciences, Seattle, WA., 224 pp.

Eckel, F. A., C. Mass, 2005: Aspects of effective mesoscale, short-range ensemble. *Wea. Forecasting*, **20**, 328–350.

Eckel, F.A., M.S. Allen, Draft 2009: Estimating Ambiguity in Ensemble Forecasts. Submitted to Weather and Forecasting.

ECMWF, cited 2009: Singular Vectors: Linear Perturbation Growth [Available online http://www.ecmwf.int/research/predictability/projects/IC_pert/SV_method/index.html] (Accessed July 30, 2009).

Ellsberg, D., 1961: Risk, ambiguity, and the Savage axioms. *Quart. J. Econ.*, **75**, 643–669.

Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from The Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104–3127.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10143–10162.

Evensen, G., 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Wea. Rev.*, **125**, 1342–1354.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hamill, T. M., 2006: Ensemble-based atmospheric data assimilation. *Predictability of weather and climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 124–156.

Hamill, T. M., S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835–1851.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.,* **129**, 2776–2790.

Hansen, J.A., 2009: Personal correspondence.

Hansen, J. A., C. Penland, 2006: Efficient approximate techniques for integrating stochastic differential equations. *Mon. Wea. Rev.*, **134**, 3006–3014.

Houtekamer, P. L., J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123**, 2181–2196.

Houtekamer, P. L., H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126**, 796–811.

Jolliffe, I. T., D. B. Stephenson, 2003: *Forecast verification: a practitioner's guide in atmospheric science.* John Wiley & Sons, Ltd., 240 pp.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.,* **82**, 35–45.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press, 341 pp.

Katz, R. W., A. H. Murphy, 1997: *Economic value of weather and climate forecasts.* Cambridge University Press, 222 pp.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Leutbecher, M., T. N. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics,* **227**, 3515–3539.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.*, **133**, 1865–1885.

Lorenc, A. C., 2003: The potential of the ensemble Kalman filter for NWP-a comparison with 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **129**, 3183–3203.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21**, 289–307.

Lorenz, E. N., 1993: *The essence of chaos.* University of Washington Press, 227 pp.

Lorenz, E. N., 1996: Predictability-a problem partly solved. *Proc. Proc. Seminar on Predictability,* 1–18.

Magnusson, L., E. Kallen, and J. Nycander, 2008: Initial state perturbations in ensemble forecasting. *Nonlin. Processes Geophy.,* **15**, 751–759.

Mullen, S. L., R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173–191.

Murphy, A. H., 1985: Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Mon. Wea. Rev.*, **113**, 362–369.

Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.,* **128**, 361–384.

National Research Council Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts.* National Academies Press, 112 pp.

Nutter, P., D. Stensrud, and M. Xue, 2004a: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.,* **132**, 2358–2377.

Nutter, P., M. Xue, and D. Stensrud, 2004b: Application of lateral boundary condition perturbations to help restore dispersion in limited-area ensemble forecasts. *Mon. Wea. Rev.,* **132**, 2378–2390.

Orrell, D., 2003: Model error and predictability over different timescales in the Lorenz'96 systems. *J. Atmos. Sci.*, **60**, 2219–2228.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128**, 747–774.

Reichle, R. H., J. P. Walker, R. D. Koster, and P. R. Houser, 2002: Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometeor,* **3**, 728–740.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126**, 649–668.

Richardson, D. S., 2001: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.*, **127**, 1847–1864.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.,* **131**, 3079–3101.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Wea. Forecasting,* **12**, 809–818.

Szczes, J.R., 2008: *Communicating Optimized Decision Input from Stochastic Turbulence Forecasts.* M.S. Thesis, Graduate School of Engineering and Applied Sciences, Naval Postgraduate School. 159 pp. [Available from the Defense Technical Information Center].

Szunyogh, I., Z. Toth, 2002: The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts. *Mon. Wea. Rev.*, **130**, 1125–1143.

Tallagrad, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. from the Workshop on Predictability,* Anonymous European Center for Medium-Range Weather Forecasts, 1–25.

TIGGE, cited 2009: Thorpex Interactive Grand Global Ensemble [Available online: http://tigge.ecmwf.int/] (Accessed July 30, 2009).

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.,* **131**, 1485–1490.

Toth, Z., E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74**, 2317–2330.

Toth, Z., Y. Zhu, I. Szunyogh, M. Iredell, and R. Wobus, 2002: Does increased model resolution enhance predictability? Preprints. *Proc. Symp. on Observations, Data Assimilation, and Probabilistic Prediction,* Orlando, FL.

Tracton, M. S., E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting,* **8**, 379–398.

Tribbia, J. J., D. P. Baumhefner, 2004: Scale interactions and atmospheric predictability: An updated perspective. *Mon. Wea. Rev.,* **132**, 703–713.

Wallsten, T. S., 1990: Measuring vague uncertainties and understanding their use in decision making. G.F. Furstenberg, Ed., Kluwer Academic Publishers, 377–399.

Wang, X., C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60**, 1140–1158.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, C. Bishop, and X. Wang, 2006: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus,* **58**, 28–44.

Weisstein, Eric W. "Runge-Kutta Method." From *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/Runge-KuttaMethod.html (Accessed July 30, 2009).

Whitaker, J. S., T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.,* **130**, 1913–1924.

Wilks, D. S., 2005: Effects of stochastic parametrizations in the Lorenz'96 system. *Quart. J. Roy. Meteor. Soc.,* **131**, 389–407.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences.* 2nd ed. Academic Press, 467 pp.

WMO, cited 2009: The Observing System Research and Predictability Experiment [Available online: http://www.wmo.int/pages/prog/arep/wwrp/new/thorpex_new.html] (Accessed July 30, 2009).

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83**, 73–83.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus A,* **52**, 280–299.

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Air Force Weather Technical Library
   14th Weather Squadron
   Asheville, North Carolina

4. Major Tony Eckel
   Naval Postgraduate School
   Monterey, California

5. Dr. Wendell Nuss
   Naval Postgraduate School
   Monterey, California

6. Dr. Patrick Harr
   Naval Postgraduate School
   Monterey, California

7. Dr. Eva Regnier
   Naval Postgraduate School
   Monterey, California

8. Dr. James Hansen
   Naval Research Lab
   Monterey, California

9. Dr. Philip Durkee
   Naval Postgraduate School
   Monterey, California