FINDING PHYSICAL CONNECTIONS IN A STATISTICAL MODEL: USING GLOBAL UNCERTAINTY AND SENSITIVITY ANALYSIS ON A MULTIPLE PREDICTOR ANALOG METHODOLOGY FOR DOWNSCLING AND BIAS CORRECTING PRECIPITATION FORECASTS

By

ROBERT W. ROONEY

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

To Ash, my family, and my friends who got me through this. It was a long hard journey.

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Dr. Christopher Martinez, for his continued support, encouragement and guidance. This would not have been possible without him. I would also like to thank my graduate committee: Dr. Rafael Munoz-Carpena, Dr. Wendy Graham, Dr. Jim Jones, and Dr. Michael Annable for their interest insight and advice. Specifically I would like to extend my gratitude to Dr. Munoz-Carpena for his support and guidance using the University of Florida HPC for sensitivity analysis.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Figure page

LIST OF ABBREVIATIONS

AM	Analog Method
ANN	Artificial Neural Network
APCP	Precipitation
BS	Brier Skill
BSS	Brier Skill Score
CA	Constructed Analogs
C1	Weighting Coefficient for Precipitation
C2	Weighting Coefficient
C3	Weighting Coefficient for
C4	Weighting Coefficient for
C5	Weighting Coefficient for
C6	Weighting Coefficient for
CRPS	Cumulative Rank Probability Skill
CRPSS	Cumulative Rank Probability Skill Score
ECMWF	European Center for Medium-Range Weather Forecasts
EE	Elementary Effects
EFAST	Extended Fourier Amplitude Sensitivity Test
FAR	False Alarm Rate
GFS	Global Forecast System
GSA	Global Sensitivity Analysis
HR	Hit Rate
KNN	K-nearest neighbor
LI	Linear Interpolation
MBE	Mean Bias Error

MOS	Model Output Statistics
MSE	Mean Square Error
MSESS	Mean Square Error Skill Score
NA	Natural Analogs
NCEP	National Center for Environmental Prediction
NmAn	Number of Analogs
NOAA	National Oceanic and Atmospheric Administration
NWPM	Numerical Weather Prediction Model
OAT	One-At-a-Time (sensitivity test)
RCM	Regional Climate Model
RHUM	Relative Humidity
RMSE	Root Mean Square Error
ROC	Relative Operating Characteristics
PRMSL	Mean Sea Level Pressure
PWAT	Precipitable Water
S_i	First Order Sensitivity Index
SM	Similarity Measure Used for Analog Selection
SM_{tot}	Similarity Measure Used for Multi-predictor Analog Selection
ST_i	Total Sensitivity Index
SW	Search Window
U10m	Meridional Wind Vector
V10m	Longitudinal Wind Vector

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

FINDING PHYSICAL CONNECTIONS IN A STATISTICAL MODEL: USING GLOBAL UNCERTAINTY AND SENSITIVITY ANALYSIS ON A MULTIPLE PREDICTOR ANALOG METHODOLOGY FOR DOWNSCLING AND BIAS CORRECTING PRECIPITATION FORECASTS

By

Robert W. Rooney

August 2013

Chair: Christopher J. Martinez
Major: Agricultural and Biological Engineering

With the recent availability of computationally expensive Numerical Weather Prediction Model archives, the use of analog or pattern matching methodologies has become a viable option for downscaling and bias correction of forecasts. Numerous studies have been performed that show skill using these methods for local scale temperature, reference evapotranspiration and precipitation forecasting. However, as this is a statistical, pattern matching approach, little study has gone into ascertaining a relationship between local climatic physics and the choice of predictors to use in a given condition. This study uses global sensitivity and uncertainty analysis with subsequent Monte-Carlo filtering to determine if there are such linkages with regard to seasonality, lead-day of the forecast, and precipitation event magnitude.

While this study showed that non-atmospheric input factors (search window size and number of analogs used) had the greatest effect in model output variability, it was determined that variations in the weighting of atmospheric predictors does offer the potential for improved forecast skill under different conditions. It was found that for the

case of the Tampa Bay region of Florida, the influence of precipitation, precipitable water, relative humidity, the meridional wind vector and the zonal wind vectors changed with respect to the winter dry season and the summer wet season as well as event threshold. Analysis showed these changes to be statistically significant. Relationships are posited whereby these changes are potentially attributed to climatic physics. An example is the increased influence of the east-west wind vector in the summer season as opposed to winter. Due to the common occurrence of convective storm formation from the sea breeze, it is plausible that this signal is showing up in the analog method. While this study is an initial foray into connecting physics to a statistical model and statistically significant variations in predictor influence was observed, further research is required to validate any causal relationships.

CHAPTER 1
INTRODUCTION

Water managers have the unenviable task of making water allocation decisions on short (one day to 1 month), medium (1 month to 12 months), and long (decades) range temporal scales without a clear idea of the future state of one of the systems primary inputs, precipitation. The majority of managers base their decision making process on the historical record and are quite reticent to incorporate probabilistic forecasts into their operational models. Their hesitation is based upon a number of factors including perceptions of poor reliability, temporal and spatial scale mismatch between the forecasts and the need of the user, and frequent misrepresentation by researchers and users (Callahan et al., 1999; Hartmann et al., 2002; Lemos et al., 2002; Lins et al., 1997; Nicholls, 1999; Pagano et al., 2002; Rayner et al., 2005; Wilks, 2000). Recently, advancements in forecasting capability has instigated water managers to posit the questions: how can they effectively replace historical information used in water allocation decisions with forecasts? And, what level of risk accompanies their incorporation? These questions exemplify that there are two distinct hurdles that must be overcome in order to successfully implement forecasts downscaling in field level decision making practices. The first is that the process has to be transparent and the second is the requirement for improved skill and reliability over what is currently being used. Transparency necessitates the use of models that are easily comprehensible by the field level user without the need for excess training. It also requires that the end users have a good understanding of the limitations of the model, when it should and should not be used as well as the propagation and range of input variability which would give the user a better understanding of the risk involved with using the model output.

With the recent availability of computationally expensive Numerical Weather Prediction Model Archives (NWPM), the use of analog or pattern matching methodologies have become a viable option for downscaling and bias correction of gridded global forecasts. While studies have shown these methods can produce local scale forecasts with skill, little study has gone into ascertaining a relationship between local climatic physics and the choice of predictors to use in a given condition. The first phase of this study is a detailed analysis of Analog Methodologies (AM) to ascertain if it is a viable option for local scale precipitation forecasts in Florida, given the regions unique topography (flat) and close proximity to two large bodies of water, the Gulf of Mexico, and the Atlantic Ocean. This is accomplished by producing multiple AM point-scale (rain-gauge) precipitation forecasts within the study area by varying the model input parameters in a one at a time approach. Subsequently, the AM forecasts are incorporated into an Artificial Neural Network stream-flow model and an evaluation of results is performed. The second phase of the study is an attempt to provide transparency to the AM as well as gain insight into the potential variation of input factor settings with regard to seasonality, lead-day propagation, and variations in precipitation event magnitude. This phase is performed using Global sensitivity and uncertainty analysis with subsequent Monte-Carlo filtering. Results from these tests not only provide information regarding the relative effect individual and combined input factors have on uncertainty propagation through the model, they can also be used to gain insight into linkages between the observed climatic physics and the predictors used for analog selection.

Numerical Weather Prediction Models

Historical Perspective

The modern form of NWPMs, used to predict the state of the atmosphere, began in 1920 when Lewis Richardson developed a cell based procedure using finite difference methods to resolve governing differential equations (Lynch, 2006). However, due to the mathematical complexity of the governing equations, the time it took to produce results proved to be slower than the occurrence of the forecasted weather. In the 1950s a seismic shift in numerical computation capacity occurred due to the introduction of the computer and access to computational simulations. The first super computers in the US had as two of their first applications the determination of the explosive yield of thermonuclear bombs as well as forecasting the state of the atmosphere, both funded by branches of the US military. This fact shows the perceived value of operational forecasts for strategic use as well as the substantial effort brought to this goal. In 1956, Phillips produced a cell based two layer climate forecast model encompassing the United States. The spatial resolution of this model was 270 grid points at 700 km intervals and produced a forecast with a three hour time step. In order to successfully produce timely output from Phillip's model (Phillips, 1956), many limiting assumptions, including most notably barotropic conditions, were used thereby reducing the complexity of Richardson's first model. These assumptions had the effect of separating the modeling algorithm from physical processes.

As computers continued to develop in speed as well as storage capacity, NWPM development was enhanced to a three dimensional model by layering cells for the different levels of the atmosphere and simplifying assumptions were removed as computational ability surpassed the Richardson model's computational complexity. In

effect these newly developed models incorporated physical processes back into the model algorithm.

Currently, both the European Center for Medium-Range Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) are well aware of the "chaotic nature of atmospheric flow" which results in rapid temporal expansion of errors in the initial state (Lynch, 2008). This error limits the forecast predictability. In order to address this concern, ensemble forecasting was seen as a possible solution. Ensemble forecasting is accomplished by running several iterations of an NWPM using slightly varied inputs. The result is an ensemble of forecasts that can be used to produce probabilistic forecasts as well as deterministic forecasts. While NWPMs have undergone substantial improvement in spatial and temporal resolution (ie 200km grid spacing), they are still limited in predicting climate variables on a small spatial scale required for use in local or regional hydrologic models (Wood et al., 2004). This limitation is due to NWPMs inability to resolve local scale climatic physics and surface forcing. Accordingly it is necessary to perform bias correction and downscaling of NWPM forecasts for use on a local or regional scale.

Bias Correction and Downscaling

Modeling bias, in the context of climatic modeling, are systematic errors in model output created by errors in, but not limited to, initial conditions, simulated physics, parameterization of model physics and computational numerical methods used. Bias error can be identified by comparing the model output to an observed reference data set in the same temporal and spatial domain. Numerous bias correction techniques, including quantile mapping and anomaly field correction, have been developed which adjust forecasted climatic variables (Maurer and Hidalgo, 2008; Wood et al., 2004). A

simple example of bias correction would be if a model consistently produces a precipitation forecast from Dec-Mar that is higher than observed conditions, all of the precipitation forecast data within that temporal range would be reduced. There is however concerns when performing bias correction in that spatial and temporal variation can also be effected.

Downscaling is used to generate statistics of local scale climate features that are consistent with the large scale atmospheric states (Vonstorch et al., 1993). Currently there are two means by which bias correction and downscaling are performed - dynamical and empirical/statistical. The dynamic methodology is a process in which downscaling is accomplished using a nested high resolution numerical Regional Climate Model (RCM) with NWPM output providing the initial and boundary conditions. While RCMs have shown improved forecast skill in local spatial and temporal domains (Vasiliades et al., 2009) there are three distinct disadvantages; propagation of bias in boundary conditions provided by NWPMs, bias in the RCM based on deficiencies in modeled cloud formation physics (Vidale et al., 2003), and computational expense.

In contrast, using a statistical methodology accomplishes both downscaling and bias correction by identifying and using a statistical transformation to relate coarse scale NWPM climate variables to local climate variables. Studies have shown forecast skill on a local/ regional spatial scale using regression techniques (Enke and Spekat, 1997), Artificial Neural Networks (ANN) (Schoof et al., 2009), and analogs (Hamill and Whitaker, 2006a). The benefits of using a statistical method over RCM simulations is the transparency in identifying bias by the user, reduced computational expense, and ease of technology transfer and training (Zorita and von Storch, 1999). There are

however limitations to the use of statistical downscaling, such as the necessity of a large and sometimes unobtainable observational data set and there is no way to account for changes in local climate forcing or feedback processes that may occur in the future. In this study the relative skill of analogs, both constructed and natural, to downscale a NWPM to point space locations in Florida is tested.

Analog Method

The AM is based upon the assumption that if it were possible, using the present atmospheric state, to identify an identical atmospheric state in the historical record the temporal propagation of climate from that date could be used to predict future climate conditions. Edward Lorenz (Lorenz, 1969) first used analogs in the meteorological field as a weather predictor. In his study he determined that based upon the length of his historic record, five years, the imperfect matches identified produced large error propagation as early as 2 day lead times. He did however state that with access to a longer historic record, temporal error propagation could be substantially reduced. While analogs were used to forecast weather in 1973, it wasn't until 1995 that the analog method was used as a downscaling technique (Zorita et al., 1995).

Natural Analogs

The use of Natural Analogs (NA) is based on a two-step process; forecast analog selection and statistical analysis of corresponding observed data. To select appropriate analogs it is necessary to choose a predictor for use as a comparator (such as precipitation), determine the temporal domain of potential analogs, and identify the closest matches using a similarity measure. Once the current predictor field has been identified it can then be compared to all of the potential analogs available.

The method used in this study is a static version of the methodology described in Hamill (2006a) and Hamill and Whitaker (2006b). Instead of using a moving spatial window, a single coarse scale window of 9 grid-points surrounding the Tampa Bay area is used (Figure 1-1). The NAs are initially selected by comparing the current day's coarse scale precipitation forecast (Z_{GFS}) and comparing it to an archive of historical forecasts within +/- 45 days for all years in the archive (maintaining seasonality), while omitting the current year. The omission of the current year in identifying potential analogs is used to cross validate the model. The similarity measure (Sm) used is the root mean square error (RMSE) calculated at each of the nine grid-points. The 75 most similar coarse scale patterns ($Z_{analogs}$) are then selected to form the analog ensemble. The observed precipitation, from individual rain-gauges, corresponding to the dates of $Z_{analogs}$ is then used to create the forecast ensemble ($P_{analogs}$). The vector, $P_{analogs,}$ can then be evaluated as a deterministic (by using the mean or median value) or probabilistic forecast.

Improvement of the analog precipitation forecast, in comparison to climatology, will be assessed using variable analog search window sizes, different predictors both singular and combined, variable number of analogs selected, and variable periods of precipitation accumulation. The implementation of this process, however, does have some disadvantages. The method does not work well for extreme or record precipitation events due to the fact that an analog would not be present in the historic record, and implementation of the analog method requires the use of a large forecast record produced from a fixed version of the NWPM.

Constructed Analogs

CAs were developed in order to overcome NA's most significant drawback; difficulty in locating a suitable analog in the historic record (Van den Dool, 1994). Van den Dool proposed that if you constructed a linear combination of the optimum NAs, acceptable "constructed" analogs would emerge. The CA method used in this study is a modification of the Van den Dool method and is described by Hidalgo (2008) as follows. Using the gridded coarse resolution predictor pattern for the current day (Z_{GFS}) and linearly relating it to the pool of coarse scale natural analog patterns ($Z_{analogs}$), an analog of Z_{GFS} can be constructed (Z'_{GFS}) according to Equation 1-1.

$$Z_{GFS} \approx \acute{Z}_{GFS} = A_{analogs} Z_{analogs} \quad (1\text{-}1)$$

where; $Z_{analogs}$ is a matrix with dimensions coinciding with the # of coarse-resolution grid-points (n), and the number of natural analogs (k), while the size of $Z_{analogs}$ is [n:1]. $A_{analogs}$ is a column vector [n:1] of fitted least squares estimates of the regression coefficients that are indicative of the the linear proportions relating the individual contributions of $Z_{analogs}$ to the constructed analog. Z'_{GFS} has dimensions of n:1, but has no further use. The calculation of $A_{analogs}$ is performed by using the definition of the pseudo inverse (Moore-Penrose inverse) according to Equation 1-2.

$$A_{analogs} = \left[\left(\acute{Z}_{analogs} Z_{analogs} \right)^{-1} \acute{Z}_{analogs} \right] Z_{GFS} \quad (1\text{-}2)$$

To downscale the Z_{GFS} pattern the coefficients $A_{analogs}$ are applied to the point scale weather patterns ($P_{analogs}$) corresponding to the same days as the coarse-resolution according to Equation 1-3.

$$P_{CA} = A_{analogs} P_{analogs} \quad (1\text{-}3)$$

Due to the nature of the constructed analogs and the fact that the forecast is being downscaled to a spatial point, the ensemble of NAs is reduced to a single deterministic forecast for each rain-gauge on each date. Therefore the probabilistic verification measures used in assessing the NAs cannot be used in comparison with CAs.

Artificial Neural Network Stream-flow Model Implementation

An Artificial Neural Network (ANN) is a computational model initially developed over 60 years ago in order to understand the functioning of the human brain (McCulloch and Pitts, 1943). In essence it is based on a group of nodes or neurons that receive multiple inputs (from other nodes or from outside inputs) and when activated channel information to other nodes, where eventually the chain leads to model output data. A single node produces output based on the weighted sum of all of its inputs according to its own activation function (Dawson 2001). This type of model, when used in hydrology is considered a lumped, parametric model (Govindaraju and Artific, 2000a). A lumped model, in terms of lumped or distributed model, references that there is no need for a detailed physical understanding of the catchment being modeled. Rather, in order for the model to successfully produce realizations of a catchment, the model must undergo training using observed data in order to determine the weight of the linkages between nodes and to also when necessary remove excessive nodes. When the number of available nodes is too large the model may have too many parameters and the model may over fit the data, too few and there may not be enough information to formulate an effective result. Training (calibration) involves processing an input data set through the model where the output is known. In so doing the nodal structure of the model adapts

to the input/output conditioning and subsequently is capable of receiving new input and producing forecasted outputs.

Although ANN models have been viewed by physical modelers as black boxes due to their structure, several modeling benefits have been identified (Dawson and Wilby, 2001; Govindaraju and Artific, 2000b; Zealand et al., 1999). First they have the ability to empirically identify the relation between input and output data without any prior knowledge, data preprocessing or explicit physical considerations. Second they have the ability to adapt their structure when given a set of inputs and a known output in order to produce consistent results. Third, ANNs have the ability to make generalizations when the data is incomplete, noisy, or distorted by using previous examples. Fourth they are non-linear which gives them better capability to solving problems than linear techniques, and due to the parallel nature of similar nodes, they can computationally run faster.

Forecast Verification

Forecast verification is necessary in climate sciences to assess the forecaster's ability to successfully make predictions. This assessment is known as forecast verification. While there is numerous views on what identifies a good forecast, common techniques include measures that identify the accuracy, reliability, resolution, bias and skill. Where;

- Accuracy: average degree of correspondence between forecast and observation.
- Resolution: the ability of a forecast to distinguish between relative high frequency and low frequency events.
- Reliability: degree of correspondence between the conditional probability forecast and the frequency of observed occurrence.

- Bias: the tendency of a forecast model to consistently over or under predict over a temporal and/or regional domain.
- Skill: relative accuracy of a set of forecasts to that of a standard or reference set of forecasts (i.e. climatology).

The type of forecast output, probabilistic or deterministic also influences the type of verification measure in that probabilistic predictions produced from ensemble forecasts cannot directly produce a scalar verification metric. If the scalar evaluation was necessary it is required that the ensemble forecast be reduced to a discrete value, for example an ensemble mean. Conversely a probabilistic measure cannot be produced from a discrete forecast without spatially or temporally grouping the predictions. In this study both discrete and probabilistic forecasts are evaluated and the verification measures used are described in the following sections.

Continuous Rank Probability Skill Score

The Continuous Rank Probability Score (CRPS) is forecast validation technique for probabilistic forecasts of a continuous variable. The CRPS is based on a comparison of the observed condition to the forecast PDF both of which are cumulative distributions and is calculated according to Equation 1-4 (Hersbach, 2000; Wilks, 2006).

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 dy \quad (1\text{-}4)$$

Where, F(y) and F_0(y) are the cumulative distribution functions of the forecast ensemble and observation respectively. If F_0(y) is a single daily value, which is the case for the AM, it is a Heaviside function or step function. Graphically this can be seen in Figure 1-2, in that the CRPS is indicative of the area between the distributions. The CRPS has a potential range of 0 to 1and is negatively oriented whereby smaller values indicate better overall accuracy.

It can be seen that if the forecast was deterministic and not probabilistic the forecast CDF, F(y) would also be a Heaviside function and the CRPS would reduce to the mean squared error. Once the CRPS is calculated for the AM forecast it is useful to compare it to a reference forecast. A common comparative baseline used in forecast verification is climatology (Hamill et al., 2004). For the case of this study, climatology consists of an average of a 60-day running window of observed precipitation over the 30-year length of the forecast archive. By example, if the lead day was July 1st, climatology would be evaluated using data from June 1st to July 31st for every year in the record. The comparison is a measure of skill or relative accuracy and for the Cumulative Ranked Probability Skill Score (CRPSS) is calculated according to Equation 1-5.

$$CRPSS = 1 - \frac{CRPS_f}{CRPS_c} \quad (1\text{-}5)$$

where $CRPS_f$ and $CRPS_c$ are the cumulative rank probability score of the forecast and climatology, respectively. The CRPSS can range from 1 to $-\infty$ where any value greater than 0 indicates that the forecast is more accurate (useful) than the climatological forecast. If the CRPSS is less than 0, the opposite is true.

Brier Skill Score

While the CRPSS shows the skill of the forecast model on a continuous basis it lacks the ability to discriminate between the amounts of precipitation predicted. The Brier Skill Score (BSS) is a commonly used summary measure that is based on probabilistic forecasts exceeding predefined precipitation event thresholds. CRPSS is in essence an integration of the Brier Score. The BSS, like the CRPSS, is based on comparing an error measure for the forecast to that of a reference forecast. The error

measure in this case is the Brier Score (BS) and is calculated as the average squared error of the probability forecasts, which can be seen in Equation 1-6.

$$BS = \frac{1}{N}\sum_{j=1}^{N}\left(P_j - O_j\right)^2 \quad (1\text{-}6)$$

where j is an ensemble of forecasts, in this case a daily ensemble provided by the AM, P is the probability of the ensemble forecast members exceeding the threshold, and O is an observed precipitation index and N is the number of ensembles represented.

The probability, P, is calculated for each ensemble by determining the fraction of members that exceed the event threshold. The observed precipitation indicator O incorporates observed conditions in the skill assessment, where O = 1 if the observed precipitation exceeded the threshold on that day, and O = 0 if it did not. A perfect Brier Score would be shown with a value of 0. For this to occur every forecast in N ensembles would predict rainfall exceeding that of the threshold and on the corresponding ensemble dates, the observed precipitation would also exceed the threshold thereby validating the forecasts.

Once the Brier Score for the analog method (BS_f) and climatological forecasts (BS_c) are computed, an evaluation of their relative accuracy (Skill) can be determined. This is accomplished by calculating the Brier Skill Score (BSS) (Equation 1-7).

$$BSS = 1 - \frac{BS_f}{BS_c} \quad (1\text{-}7)$$

The BSS has a potential range of values from a perfect score of 1 all the way to $-\infty$. Because this is a relative measure of accuracy it can be said that any score greater than 0 identifies the forecast as better than the reference baseline or in the case of this study, climatology. If a score is achieved below 0 than the forecast based on climatology was more accurate than the forecast.

Reliability and Relative Operating Characteristic

The reliability diagram (Figure 1-3), a comparison of the observed relative frequency to the predicted frequency, is used to assess the forecast model's reliability and give an indication of its resolution. Unlike the CRPSS the reliability diagram does not incorporate all precipitation thresholds, but rather is a discrete verification method that is applied for each threshold level. A perfectly reliable forecast model occurs when the forecasted event probability coincides with the rate of occurrence and is indicated by a plot that falls upon a 45 degree line (from [0,0] to [1,1]). Deviations from the 45 degree line indicate a reduction in reliability with the forecast overestimating if it falls below and underestimating when above. Plots that do not propagate across all probability bins indicate that the forecast has shown no event probability at that respective level. This can be due to the event being a rare/extreme event which by nature does not have a high probability of occurrence, or it can be due to limitations in the forecast model's ability to predict an event (i.e. predict a large precipitation event with a 10 day lead). A horizontal line is indicative of a model with poor resolution or inability to discriminate between events. However deviations from the perfectly reliable line can be a result of sampling limitations and not forecast reliability (Hamill and Colucci, 1998; Jolliffe and Stephenson, 2003; Wilks, 2006) therefore a histogram based on sample size can be included in order to improve the user's confidence as well as indicate the sharpness of the model output. Sharpness is used in this context as the spread or distribution of the forecasts.

A Relative Operating Characteristics (ROC) curve (Figure 1-4) is used to determine the resolution of the forecast model by showing the models ability to discriminate between two outcomes. The ROC curve is created by comparing the

number of hits (predicted occurrence is observed to occur) and the false alarm rate (predicted occurrence does not occur) in each probability bin. The ROC curve does not have the ability to give information about reliability in that bias is not incorporated in its construction. In effect the ROC curve shows the potential skill of a forecast model and not necessarily the accuracy of the model. Frequently the ROC curves and Reliability diagrams are presented together due to the complimentary nature of the information presented.

Discrete Forecast Evaluation Techniques

Three scalar verification measures are used in this study to assess forecast quality. RMSE identifies the accuracy of the forecast, Mean Square Error Skill Score (MSESS) identifies the relative accuracy compared to a control, and Mean Bias identifies the forecast tendency to over or under predict.

The MSESS was calculated (Equation 1-8) by determining the daily MSE of the discrete forecasts.

$$MSESS = \left[1 - \frac{MSE_f}{MSE_c}\right] \quad (1\text{-}8)$$

where MSE_f is the mean square error of the forecast and MSE_c is the mean square error of climatology. Climatology was calculated in the same manner as stated for probabilistic forecasts. It can be seen from Equation 1-8 that the range of MSESS is from 1 (best case) to$-\infty$. A negative MSESS implies that it would be better to use the climatological forecast over that of the modeled forecast.

The MBE determines if there is a bias in the modeled forecast when compared to observed precipitation records by calculating the mean difference between forecast and observed data (Equation 1-9).

$$MBE = \frac{1}{N}\sum_{i=1}^{N}\left(P_{obs} - P_{forecast}\right) \quad (1\text{-}9)$$

Where N is the number of forecasts, and P is the observed and forecasted precipitation amounts. It should be noted that it is possible to obtain low values of MBE while the forecast has high RMSE. This would occur if the forecast deviated from the observed values by both higher and lower readings. Therefore it is a common practice to present this evaluation metric along with the RMSE. The RMSE is not affected by low and high differences cancelling out due to the difference being squared in the calculation.

Uncertainty and Sensitivity Analysis

Sensitivity analysis in essence is an assessment of the change in model output in relation to variations in model input mathematically represented by a derivative. The simplest form would be varying a single input parameter for each consecutive model realization. This is known as One-At-a-Time (OAT) sampling. However there is a limitation to this technique in that the derivative is localized to a single position (base point) thereby giving a local sensitivity analysis only. The primary limitation to local analysis is that it cannot give an assessment of model sensitivity throughout the input parametric space (Saltelli, 2008; Saltelli et al., 2006). This local assessment of sensitivity can be effective for linear models, where linear extrapolation can be used to produce results away from the base point; it is not effective for non-linear models. To overcome this limitation a Global Sensitivity Analysis (GSA) which is based on input sampling throughout the space is required (Saltelli, 2004).

It is important to define the goal of sensitivity analysis before selecting a method or combination of methods for use. Applicable goals consist of Factor Fixing, Factor Prioritization, Variance Cutting and Factor Mapping (Cacuci et al., 2003; Saltelli et al.,

2006). Factor Fixing consists of identifying input factors that can be fixed to a single value due to their limited impact with regard to output uncertainty and variance. Factor Prioritization assesses the level of impact an input factor has on output variability in relation to the other input factors. Variance Cutting identifies the minimal set of input factors that can be fixed in order to achieve a predefined reduction in uncertainty. Factor Mapping identifies the input factors that have the greatest impact in producing model output in a given region. While all can be used to gain insight into output model uncertainty, factor mapping is based on Monte Carlo filtering while the other three base their analysis on model variance (Saltelli et al., 2006). From these applications there are three distinct methodological categories that can be used for GSA (screening, regression and variance based) and each has its own inherent strengths and weaknesses

Screening methods are used in models to provide a short list of important input factors when the number of input parameters or the length of computational time for one model realization is excessive to the point that it is untenable to perform the number of realizations required for more extensive regression and variance methodologies. Typically screening methods, such as the design of Cotter, Iterated fractional factorial design, sequential Bifurcation, the Method of Morris and Modified Method of Morris are based upon OAT realizations (Morris, 1991; Saltelli et al., 2000). The benefit of these methods is based on their low computational expense and ability to prioritize input parameter effects. Of the methods listed, the Method of Morris is the only one that covers the entire range of input parametric space (Campolongo et al., 2005; Morris, 1991). Once performed, a subset of input factors that dominate output variability is

identified for use in more extensive methodologies. The downside of using screening measures is that results are only indicative of qualitative relative input factor sensitivities and not quantitative measures.

Regression methods for sensitivity analysis incorporate the use of regression techniques to estimate the effect of input parameters on output variability (Saltelli et al., 2000). A major limitation to these techniques is they are not effective for nonlinear or non-monotonic models (Saltelli et al., 2006). Saltelli also showed concern in that GSA is designed for use in uncertainty propagation while general regression techniques remove variability from the input parameter thereby giving an indicator value that does not incorporate the full range of uncertainty (Saltelli, 2008). Variance based methods can effectively overcome these limitations.

Variance based methods, such as the method of Sobol, (Sobol, 1993) Fourier Amplitude Sensitivity Test (FAST) (Reusser et al., 2011) and extended FAST (eFAST) (Saltelli et al., 1999) are based on Monte Carlo simulations using all of the input parameters and a predefined sampling strategy that covers the entire input parametric space. Analysis is performed by reviewing the resulting sensitivity indices. For a given input factor, the sensitivity index is a value that corresponds to the relative effect or contribution that factor has on the model output variance. The sensitivity indices are calculated by decomposing the total variance of the output as a sum of the individual input effects as well as the interactive effects. In this study the modified Morris Method is used as a qualitative screening of NA input parameters prior to the computationally expensive Sobol method for GSA.

Morris Method

The Morris (1991) method originally developed in 1991 is based on OAT simulations where each input factor's range of variability is split into scalar or discrete intervals or levels resulting in a k-dimensional p-level gridded input domain. The variable k coincides with the number of input factors tested and p is commonly set to 4, 6 or 8. Using this gridded domain and a One at a Time sampling (OAT) scheme, whereby each input factor is moved one by one using a step of Δ, an input trajectory of (k+1) points is produced. The coordinates of each point (i.e. X_1) in a given trajectory is used to produce a model realization where $Y=f(X_1, X_2, \ldots, X_k)$ is a vector of corresponding model output. Once the screening realizations are performed the input factor Elementary Effects (EEs) are calculated (Equation 1-10).

$$EE_i(X) = \frac{[y(x_i, \cdots\cdots x_{i-1}, x_i+\Delta, x_{i+1}, \cdots\cdots, x_k) - y(X)]}{\Delta} \quad (1\text{-}10)$$

Where i represents input factor I, $EE_i(X)$ is the elementary effect of input factor i from a single trajectory, and y(X) is the model output produced using the input coordinate prior to the jump of Δ. It can be seen that each input will have a single jump where all other coordinates remain the same. The Equation 1-10 simply takes the difference of the model output realizations before and after the jump, and divides it by the size of the jump. As the input space is originally based on uniform distributions from 0 to 1 for all input factors, relative comparative analysis can be applied.

Once the elementary effects are calculated for each input the average, μ, can be used in relation to the other inputs as a relative importance factor, the higher the value the more important the parameter with regard to output variability. A second statistic that is used is standard deviation, σ, of the Elementary effects. A large standard

deviation indicates there is a significant interactive effect with other parameters while a small standard deviation indicates little to no interactive effects on the output. A modification was made to the Morris method by Campolongo (2005) in that the mean of the EEs is calculated using the absolute values. The thought was that if you had positive and negative EEs they could shift the mean towards zero while using the absolute values would take into account output variation in either the positive or negative direction. This measure is denoted as μ^*.

One of the reasons for performing a screening method is in the ability to obtain the relative importance of an input parameter in regards to its effect on output variability using a limited amount of model realizations (Saltelli et al., 2006). In the case of the Morris method the number of simulations required (N) is calculated by Equation 1-11.

$$N = r(k+1) \tag{1-11}$$

Where k is the number of inputs used and it has been shown that for good results r = 10 (Campolongo et al., 1999; Saltelli et al., 2000). Therefore for this study using 9 inputs, 100 simulations would be required for analysis.

Sobol Method

The Sobol (1993) method is based on an ANOVA-like decomposition of model output variance and attributing fractions to individual input factors as well as their interactive effects. The generic decomposition of the unconditional output variance V(Y) can be seen in Equation 1-12.

$$V(Y) = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<m} V_{ijm} + \ldots + V_{12\ldots k} \tag{1-12}$$

where; V_i is the output variance attributed to input factor i, V_{ij} is the variance attributed to the interactive effects of inputs i and j, and k is the number of inputs factors.

The number of terms in ANOVA decomposition is based on the number of factors and increases at a rate of 2^k, therefore two sensitivity indices are commonly used to assess the results (Saltelli et al., 2000).

Sobol sensitivity indices are ST_i and S_i for total and first order sensitivities, respectively. Si is the ratio of the fraction of the total output variance due to the input factor I (Equation 1-13), to the total unconditional variance of the output;

$$S_i = \frac{V_i}{V(Y)} \tag{1-13}$$

V_i can be written in terms of conditional variance transforming Equation 1-13 into Equation 1-14.

$$S_i = \frac{V(E[Y|X_i])}{V(Y)} \tag{1-14}$$

where; X_i represents input factor i. In simple terms the sensitivity index S_i can be viewed as the fraction of output variance directly attributed to input factor i. Values of S_i can fall anywhere within the range [0, 1]. When there are no interactive effects present in the model, then all output variance can be described by the conditional variance term of the RHS of Equation 1-14. Correspondingly the total unconditional variance would be equal to the sum of input factors' first order conditional variance. Such a model is called a purely additive model.

The second sensitivity index, ST_i, accounts for first order effects as well as variations in model output due to higher order interactive effects. In order to calculate ST_i a second conditional variance is used, $V(E(Y|X_{-i})$ (Homma and Saltelli, 1996). This is the closed contribution to the output variance due to the remaining k-1 factors (Saltelli, 2004). This conditional variance can then be subtracted from the total variance

leaving the sum of the variance contributions from all higher order (interactive) terms in Equation 1-12 that include X_i. Therefore ST_i can be according to Equation 1-15.

$$ST_i = \frac{V(Y) - V(E(Y|X_{-i}))}{V(Y)} = 1 - \frac{V(E(Y|X_{-i}))}{V(Y)} \quad (1\text{-}15)$$

For i=1 and k=3, Equation 1-16 shows all direct and interactive sensitivity effects for input factor 1.

$$ST_1 = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3} \quad (1\text{-}16)$$

where; S_1 is the sensitivity index for the first order effects and the remaining terms ($S_{1,2}$, $S_{1,3,}$ and $S_{1,2,3}$) correspond to the interactive effects for input factor 1. A quantitative assessment of a factor's interactive effects, in relation to first order effects, can be made by subtracting S_i from ST_i. Input factors with a S_i approaching the magnitude of ST_i indicate very little interactive effects, while small values of S_i with corresponding large values of ST_i are indicative of a factor whose influence on output variability is based on interactions with other factors. Due to the computational expense required to calculate interactive effects for all input factor combinations, the individual interactions cannot be parsed out.

Objectives

The focus of this study is to gain a better understanding of the relationship between the selection of predictors used to find analogs and the corresponding observed climatic conditions. The study is presented in two phases; 1) comparative assessment of Analog Methodologies used to produce local scale precipitation forecasts and subsequent implementation into a stream-flow model, and 2) Global sensitivity and uncertainty analysis with subsequent Monte-Carlo filtering..

The specific objectives are:

- Evaluate improvement of the Natural Analog Method for use in downscaling and Bias correcting Global Forecast System (GFS) short-term Forecasts in the Tampa Bay region

- Compare optimized NAs with two other downscaling methods, CAs and Linear interpolation (LI).

- Assess the use of Short Term Precipitation Forecasts produced from the NA method into Tampa Bay Water's operational ANN Stream-flow Forecast Model.

- Using the Method of Morris to evaluate the importance of the NA input parameters using multiple predictors.

- Using the method of Sobol, perform a quantitative Global Sensitivity Analysis on the NA downscaling method to assess input parameter variation on short term precipitation forecasts by analyzing the temporal dynamics of output sensitivity.

- Using the GSA model simulations, perform MC filtering to identify variations in input factor settings with regard to seasonality, lead-day propagation and event magnitude.

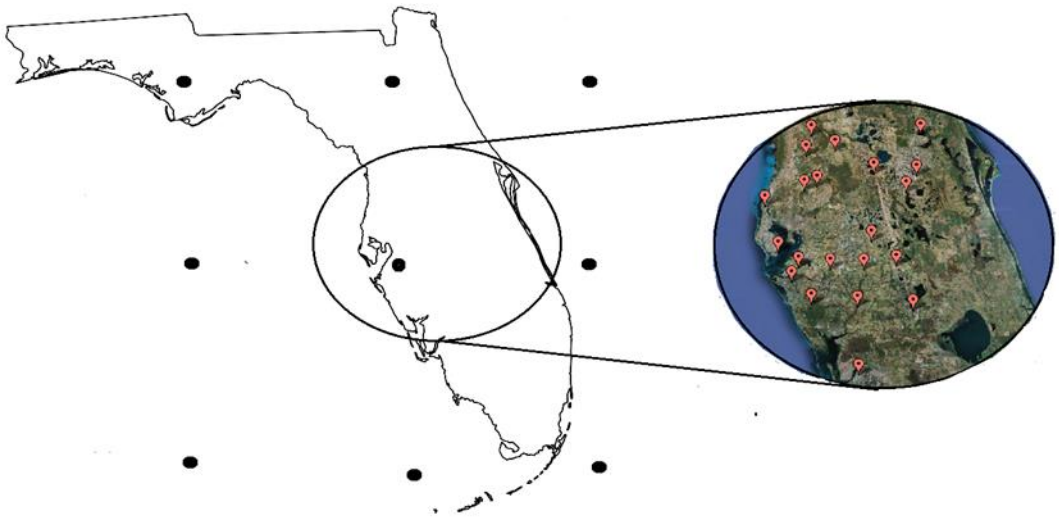


Figure 1-1. Coarse scale NWPM grid and rain-gauge locations overlaid on an outline map of Florida.

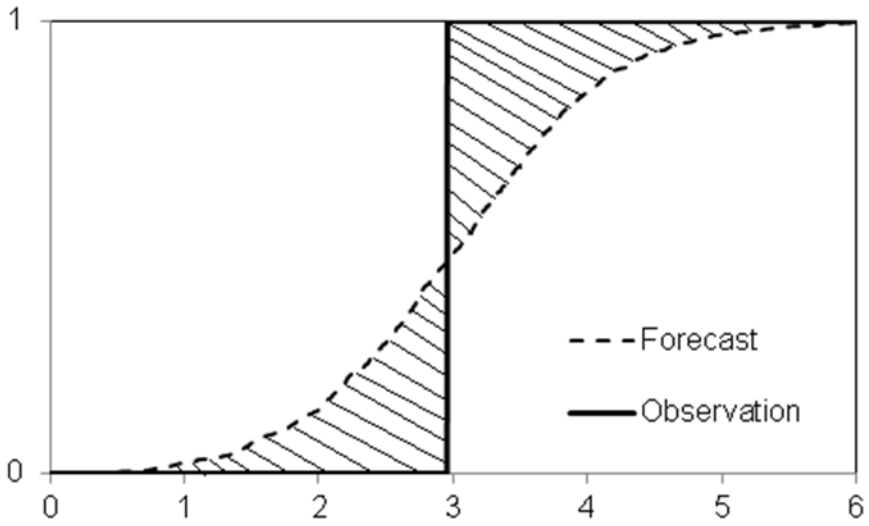


Figure 1-2. Example of CRPS calculations if the observed data is a step function. The value of CRPS is the area that is shaded.

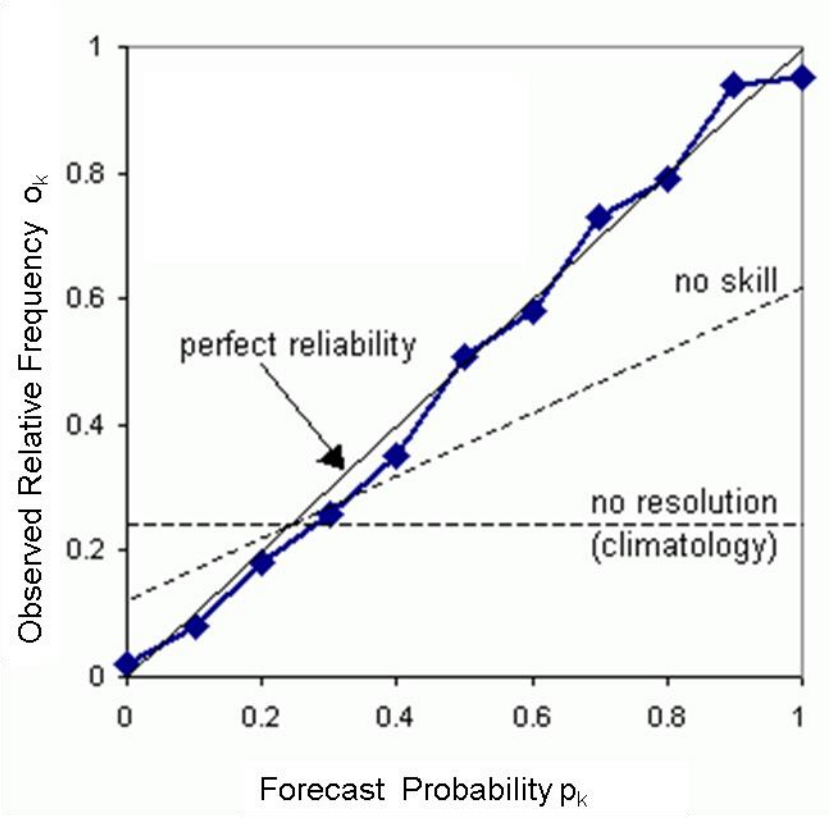


Figure 1-3. Example of a reliability diagram comparing forecast probability to the observed relative frequency. Perfect reliability is indicated by the 45 degree line, while the horizontal line is the climatological frequency.

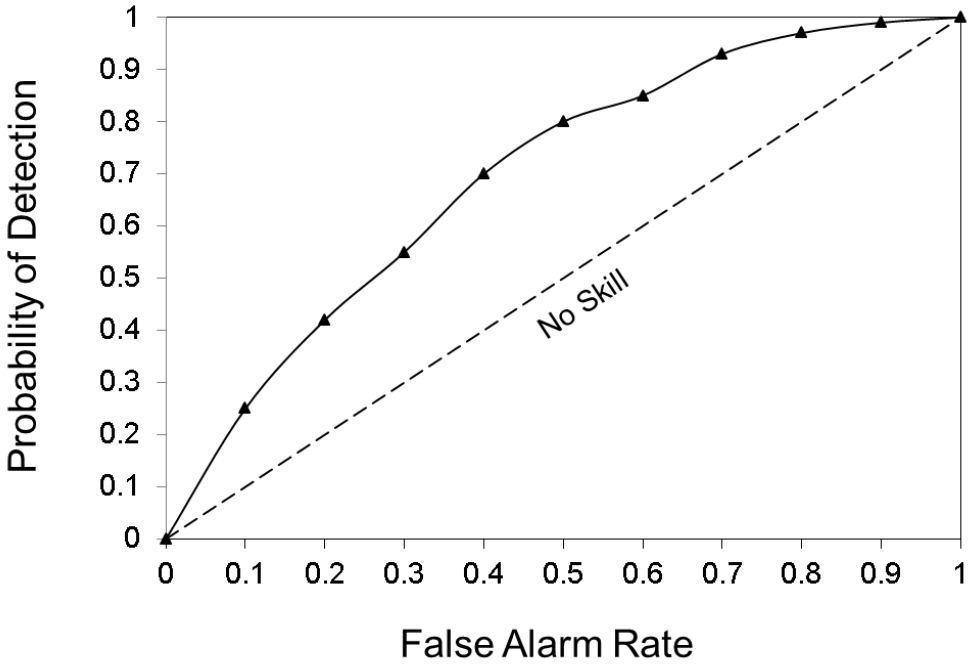


Figure 1-4. Example of a relative operating characteristic (ROC) plot comparing the forecast probability of detection to the false alarm rate. Any data that falls below the 45 degree, or 1:1 line, denotes a forecast with no skill.

CHAPTER 2
EVALUATION OF ANALOG METHODOLOGIES FOR USE AS PRECIPITATION FORECAST DOWNSCALING AND BIAS CORRECTION TECHNIQUES IN FLORIDA'S UNIQUE CLIMATE

Background

While Global NWPM have shown marked improvement in grid resolution (150-200 km) and forecast skill, agricultural hydrologists as well as water management agencies involved in planning require improved precipitation prediction on a local spatial scale (10-35 km) that cannot be directly obtained from the coarser scale NWPM resolution (Clark et al., 2004). In order to provide usable local scale forecasts, NWPM output must undergo bias correction and downscaling to account for sub-resolution (<150km) regional forcing such as coastlines, lakes, and land surface characteristics that are known to influence local climatic conditions (Murphy, 1999).

Downscaling is used to generate statistics of local scale climate features that are consistent with the large scale atmospheric states (Vonstorch et al., 1993). Currently there are two means by which bias correction and downscaling are performed - dynamical and empirical/statistical. The dynamic methodology is a process in which downscaling is accomplished using a high resolution numerical Regional Climate Model (RCM) using NWPM output to provide the initial and boundary conditions. Though dynamic downscaling has shown skill in producing regional scale forecasts there are several noted issues; 1) computational cost, 2) bias in coarse scale NWPM forecasts propagate through the RCM to the downscaled output, 3) RCMs can introduce their own bias into the fine scale forecast. In contrast, using a statistical methodology accomplishes both downscaling and bias correction by identifying and using a statistical transformation to relate NWPM climatic variables to local observed weather conditions.

Zorita (1999) showed that statistical techniques resulted in a similar skill to current dynamic downscaling methodologies. Other benefits of using a statistical method over RCM simulations is the transparency in identifying bias by the user, reduced computational expense, and ease of technology transfer and training (Zorita et al., 1995). There are, however, limitations to statistical downscaling including: the necessity of a large sometimes unobtainable fixed physics forecast and observational data sets and there is no way to account for changes in local climate forcing or feedback processes that may occur in the future (Wilby and Wigley, 1997). The difficulty in obtaining a large enough archive of NWPM data is that most NWPMs periodically undergo revision in an effort to improve forecast skill, and once changed lose comparative value with the past forecast record.

In an attempt to provide water managers with easily accessible long term fixed physics forecast information, both the Climate Diagnostic Center (CDC) of the National Oceanic and Atmospheric Administration (NOAA) and the European Centre for Medium-Range Weather Forecasts (ECMWF) produced extensive archives of retrospective forecasts (reforecasts) using a fixed version of a NWPM. These archives are usually not available primarily due to their large computational expense. The reforecast data set is a valuable tool for bias correction and downscaling due to its temporal length thus enabling the identification of model error or bias much more readily (Hamill et al., 2006b).

Recently studies have been performed to assess the usable application of reforecast data sets including week 2 forecasts, short range precipitation forecasts (Hamill and Whitaker, 2006a; Jung and Vitart, 2006), forecasts of normally distributed

atmospheric fields (i.e. geo-potential height and temperature) (Hagedorn et al., 2008; Hamill and Whitaker, 2007), and stream-flow predictions (Voisin et al., 2011; Wang et al., 2011). The use of both the GFS and the ECMWF reforecast data sets as a means of probabilistic forecast calibration showed that with an archive as limited as 20 years the resulting calibrated output exceeded the predictive capacity of either models raw output of both temperature (Hamill et al., 2008) and precipitation (Hagedorn et al., 2008). Whitaker et al (2006) showed that by using Model Output Statistics (MOS) and a reforecast archive from a single model can significantly improve the skill of week 2 forecasts. Although extensive reforecast data sets are seldom produced due to their excessive computational expense, studies have validated Hamill's assertion that they can be used to diagnose model bias resulting in improved model development, to statistically correct weather forecasts to aide in developing user specific products, or they can be used to study atmospheric predictability.

One of the benefits of using a reforecast data set with a large temporal scale produced using a fixed model is the variety of methods available for bias correction and downscaling. Wilkes and Hamill (2007) preformed a comparison of the relative abilities of linear regression (LR), non-homogeneous Gaussian regression (NGR), and Gaussian ensemble dressing using the GFS reforecast archive, as well as an analysis of the potential of a NA and a Bias Corrected Constructed Analog Methodology. Other downscaling and bias correction techniques that have been used in accompaniment with reforecast data sets are Bayesian Model averaging (Schmeits and Kok, 2010), MOS (Robertson et al., 2012), and multi-model reforecast ensembles (Whitaker et al., 2006). As this study was undertaken to improve the local precipitation forecasts for the

Tampa Bay region to potentially aide in water management decision making, a statistical downscaling methodology, AMs were selected for analysis due to its lack of complexity and production of ensemble forecasts.

The AM is based upon the assumption that if it were possible, using the present atmospheric state, to identify an identical state in the historical record, the temporal propagation of climate from that date could be used to predict future climate conditions (Hamill and Whitaker, 2006a; Vandendool, 1989). Edward Lorenz (1969) first used analogs in the meteorological field as a weather predictor. It wasn't until 1995 that the analog method was used as a downscaling technique (Zorita et al., 1995). The analog method has shown that given the availability of a large forecast archive to act as a training set the skill of raw model output can be improved for precipitation (Hamill and Whitaker, 2006a), temperature (Hagedorn et al., 2008), soil moisture (van den Dool et al., 2003), and evapotranspiration (Tian and Martinez, 2012). It also has the benefit of using observed weather patterns and hence the spatial covariance structure of local scale weather is maintained in the simulated field and it does not assume the form of the probability distribution of downscaled variables, which is the case with other statistical technuiques (Matulla et al., 2008). Hamill et al. (2006b) (Hamill and Whitaker, 2006a) demonstrated that skill improvement using the reforecast analog method was equivalent to 10 years of model improvement/development. They evaluated the improvement of skill over the entire continental United States, using a gridded observed precipitation field, and made a recommendation for uniform input parameter values for use throughout the nation. The first part of this study is to ascertain if Florida's unique conditions (i.e. extended coastline, numerous water bodies including the largest

everglades in the country, and lack of topography) require an adjustment in these values in order to obtain optimum predictive skill and secondarily comparing the NA results with results produced by Constructed Analogs, and direct Spatial Interpolation.

Objectives

The objectives of this work is to assess the viability of using the single-predictor AM as a precipitation forecast downscaling and bias correction methodology in Florida's unique climatic conditions. The hypothesis is that the single-predictor AM can produce skillful local scale 1-15 day precipitation forecasts for all seasonal conditions, however the optimum input factor settings differ from the reference literature continental settings. This work involves simulation results produced from one at a time input variation which are subsequently analyzed for skill and compared to two other downscaling methodologies.

Study Area and Data Collection

Tampa Bay Waters area of responsibility includes a 6,583 km^2 watershed which intersects 3 counties; Hillsborough, Pasco, and Pinellas. This study was performed using a 9 point GFS reforecast grid ranging from a latitude of 25-30^{0}N and a longitude of 80-85^{0}W (Figure 1-1) with a spatial resolution 0f 2.5^0 or approximately 200 km. While there are 53 rain-gauges in the watershed of interest, the AM requires that observational precipitation records span the entire temporal range of the reforecast archive and that missing data is limited. These requirements severely limited the pool of viable rain gauge records. Therefore the observed data spatial region was expanded to include rain-gauges from the surrounding area (Table 2-1).

The downscaling methodologies applied in this study require the use of two different sources of data; simulated gridded NWPM forecast data and observed rain-

gauge data, both spanning the same temporal range. The NWPM data is used in the analog selection process and the second for subsequent statistical analysis of corresponding observational data. The forecast data archive used during the selection process is the reforecast data set produced by the National Center for Environmental Protection using a fixed 1998 operational version of the GFS model. The archive has a temporal range of over 30 years, 1/1/1979 – present, and a spatial resolution of 2.5 degrees or approximately 200 km A 15-member ensemble forecast was produced every day from January 1979 to present resulting in a 30+ year data set. The ensemble initial conditions consist of a control initialized with NCEP-NCAR reanalysis and a set of seven bred pairs. The breeding method was consistent with the 1998 operational model. The forecast extends to 15 days and was archived every 12 hours (Hamill, 2006b). The reforecast archive can be downloaded at: http://www.esrl.noaa.gov/psd/forecasts/reforecast/data.html.

For rain-gauges that had data gaps or where the temporal domain did not cover the range of the reforecast data set no attempt was made to fill in missing data, instead all missing data was annotated NAN and if encountered by the MATLAB programming, the NA simply went to the next available analog with the closest match.

Methodology

This study has two phases of investigation; the assessment and optimization of using direct NAs using variations in input parameters, and subsequently the comparison of the optimized NAs with two other downscaling methodologies; Linear Interpolation and Constructed Analogs (LI and CA).

Natural Analogs

The use of NAs is based on a two-step process; analog selection and statistical analysis of corresponding observed precipitation data. To select appropriate analogs it is necessary to choose a predictor for use as a comparator (precipitation), determine the temporal domain of potential analogs, and identify the closest matches using a similarity measure. Once the current predictor field has been identified it can then be compared to all of the potential analogs available using the similarity measure.

The method used in this study is a static version of the methodology described in (Hamill and Whitaker, 2006a). Instead of using a moving spatial window, a single coarse scale window of 9 grid-points surrounding the Tampa Bay area is used (Figure 1-1). The NAs are initially selected by comparing the current day's coarse scale precipitation forecast (Z_{GFS}) and comparing it to an archive of historical forecasts within +/- 45 days for all years in the archive (maintaining seasonality), while omitting the current year. The omission of the current year in identifying potential analogs is used to cross validate the model. The similarity measure used is the RMSE calculated at each of the nine grid-points. The 75 most similar coarse scale patterns ($Z_{analogs}$) are then selected to populate an analog ensemble. The observed precipitation, from individual rain-gauges, corresponding to the dates $Z_{analogs}$ is then used to create the forecast ensemble ($P_{analogs}$). The vector, $P_{analogs,}$ can then be evaluated as a deterministic or probabilistic forecast.

Improvement of the local precipitation forecast, in comparison to climatology, will be assessed using variable search window sizes, different predictors both singular and combined, variable number of analogs selected, and variable periods of precipitation accumulation. The implementation of this process, however, does have some

disadvantages. The method does not work well for extreme or record precipitation events due to the fact that a suitable analog would not be present in the historic record, and implementation of the analog method requires the use of a large forecast record produced from a fixed version of the GCM.

Constructed Analogs

CAs were developed in order to overcome NA's most significant drawback; difficulty in locating a suitable analog in the historic record (Van den Dool, 1994). Van den Dool proposed that if you constructed a linear combination of the optimum NAs, acceptable analogs would emerge. The CA method used in this study is a modification of the Van den Dool method and is described by Hidalgo (2008) as follows. Using the gridded coarse resolution predictor pattern for the current day (Z_{GFS}) and linearly relating it to the pool of coarse scale natural analog patterns ($Z_{analogs}$), an analog of Z_{GFS} can be constructed (Z'_{GFS}) according to Equation 2-1.

$$Z_{GFS} \approx \acute{Z}_{GFS} = A_{analogs} Z_{analogs} \quad (2\text{-}1)$$

Where; $Z_{analogs}$ is a matrix with dimensions coinciding with the # of coarse-resolution grid-points (n) and the number of natural analogs (k), while the size of $Z_{analogs}$ is [n:1]. $A_{analogs}$ is a column vector [n:1] of fitted least squares estimates of the regression coefficients that are indicative of the linear proportions relating the contributions of $Z_{analogs}$ to the constructed analogs. Z'_{GFS} has dimensions of [n:1], but has no further use. The calculation of $A_{analogs}$ is performed by using the definition of the pseudo inverse (Moore-Penrose inverse) according to Equation 2-2.

$$A_{analogs} = \left[\left(\acute{Z}_{analogs} Z_{analogs} \right)^{-1} \acute{Z}_{analogs} \right] Z_{GFS} \quad (2\text{-}2)$$

To downscale the Z_{GFS} pattern the coefficients $A_{analogs}$ are applied to the point scale weather patterns corresponding to the same days as the coarse-resolution according to Equation 2-3.

$$P_{CA} = A_{analogs} P_{analogs} \quad (2\text{-}3)$$

where $P_{analogs}$ has dimensions of [n:1].

Due to the nature of the constructed analogs and the fact that the forecast is being downscaled to a spatial point, the ensemble of NAs is reduced to a single deterministic forecast PCA for each rain-gauge on each date. Therefore the probalistic verification measures used in assessing the NAs cannot be used in comparison with CAs.

Linear Interpolation

Although linear interpolation as a downscaling methodology is considered simplistic, it is used in this study as a reference. In this case the coarse scale GFS reforecast data is linearly interpolated to the latitude and longitudinal coordinates of each rain-gauge. The resulting precipitation forecasts can then be used as an indirect comparator to GFS model skill.

Probabilistic Verification Measures

The CRPS is forecast validation technique for probabilistic forecasts of continuous variable. The CRPS is based on a comparison of the observed condition to the forecast PDF both of which are cumulative distributions and is calculated according to Equation 2-4 (Hersbach 2000, Wilkes 2006).

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 dy \quad (2\text{-}4)$$

Where, F(y) and $F_0(y)$ are the cumulative distribution functions of the forecast and observation respectively. If the observed function ($F_0(y)$) is a single daily value, which is the case for the AM, it is a Heaviside function or step function. The CRPS is indicative of the area between the distributions. The CRPS has a potential range of 0 to 1 and is negatively oriented whereby smaller values indicate better overall accuracy.

It can be seen that if the forecast was deterministic and not probabilistic the forecast CDF would also be a Heaviside function and the CRPS would reduce to the mean squared error. Once the CRPS is calculated for the AM forecast it is useful to compare it to a reference forecast. A common comparative baseline used in forecast verification is climatology (Hamill et al., 2006a). In this study, climatology consists of an average of a 60-day running window of observed precipitation over the 30-year length of the forecast archive. By example, if the lead day was July 1st, climatology would be evaluated using data from June 1st to July 31st for every year in the record. The comparison is a measure of skill or relative accuracy and for the CRPSS is calculated according to Equation 2-5.

$$CRPSS = 1 - \frac{CRPS_f}{CRPS_c} \quad (2\text{-}5)$$

The CRPSS can range from 1 to -∞ where any value greater than 0 indicates that the forecast is more accurate (useful) than the climatological forecast. If the CRPSS is less than 0, the opposite is true.

While the CRPSS shows the skill of the forecast model on a continuous basis it lacks the ability to discriminate between the amounts of precipitation predicted. To ascertain the quality of forecasts produced with regard to precipitation event magnitude, the Brier Skill Score (BSS) a commonly used measure of forecast quality, was selected

as a verification method. CRPSS is in essence an integration of the BSS. The BSS, similar to CRPSS, is based on a comparison of the error of a forecast to that of a reference forecast. In this case the Brier Score (BS) is used as the comparator. The BS is calculated as the average squared error of an ensemble of probability forecasts and can be calculated according to Equation 2-6.

$$BS = \frac{1}{N}\sum_{j=1}^{N}\left(P_j - O_j\right)^2 \quad (2\text{-}6)$$

where; N is the number of forecasts, P_j is the probability of forecast j's occurrence, and O is the observed precipitation index.

Given an ensemble (j) of forecasts, P_j is determined by the number of forecasts that exceed a predefined threshold divided by the size of the ensemble. In simple terms the probability can be thought of as the fraction of forecasts in the ensemble exceed the threshold. Daily forecast ensembles provided by the AM are used in this study. The observed precipitation indicator O incorporates observed conditions in the skill assessment, where O = 1 if the observed precipitation exceeded the threshold on that day, and O = 0 if it did not. A perfect Brier Score would be indicated by a value of 0. This would indicate that for all ensembles (N) every forecast predicted rain higher than the threshold, and on the dates corresponding to each ensemble, rainfall occurred equal to or greater than the threshold. That would be a perfect prediction record.

Once the Brier Score for the analog method (BS_f) and climatological forecasts (BS_c) are computed, an evaluation of their relative accuracy (Skill) can be determined. This is accomplished by calculating the BSS (Equation 2-7).

$$BSS = 1 - \frac{BS_f}{BS_c} \quad (2\text{-}7)$$

The BSS has a potential range of values from a perfect score of 1 all the way to $-\infty$. Because this is a relative measure of accuracy it can be said that any score greater than 0 identifies the forecast as better than the reference baseline or in the case of this study, climatology. If a score is achieved below 0 than the forecast based on climatology was more accurate than the analog method forecast.

The reliability diagram (Figure 1-3), a comparison of the observed relative frequency to the predicted frequency, is used to assess the forecast models reliability and give an indication of its resolution. Unlike the CRPSS the reliability diagram does not incorporate all precipitation thresholds, but rather is a discrete verification method that is applied for each threshold level. A perfectly reliable forecast model occurs when the forecasted event probability coincides with the rate of occurrence and is indicated by a plot that falls upon a 45 degree line (from [0,0] to [1,1]). Deviations from the 45 degree line indicate a reduction in reliability with the forecast overestimating if it falls below and underestimating when above. Plots that do not propagate across all probability bins indicate that the forecast has shown no event probability at that respective level. This can be due to the event being a rare/extreme event which by nature does not have a high probability of occurrence, or it can be due to limitations in the forecast models ability to predict an event (i.e. predict precipitation with a 10 day lead). A horizontal line is indicative of a model with poor resolution or inability to discriminate between events (Jolliffe and Stephenson, 2003).

A Relative Operating Characteristic (ROC) curve is used to determine the resolution of the forecast model by showing the models ability to discriminate between two outcomes. The ROC curve is created by comparing the number of hits (predicted

occurrence is observed to occur) and the false alarm rate (predicted occurrence does not occur) in each probability bin. The ROC curve does not have the ability to give information about reliability in that bias is not incorporated in its construction. In effect the ROC curve shows the potential skill of a forecast model and not necessarily the accuracy of the model. Frequently the ROC curves and Reliability diagrams are presented together due to the complimentary nature of the information presented. The 45 degree line in a ROC plot is called the “No Skill” line with ROC curves falling above the line showing forecast skill and curves falling on or below the 45 degree line showing no skill.

Deterministic Verification Measures

While NAs produce an ensemble of forecast predictions for each day, CAs reduce the daily ensemble into an individual deterministic forecast. Therefore it is necessary to use deterministic verification measures to establish relative forecast skill. In the case of this study daily RMSE, MBE, and MSESS are used. In order to reduce the daily NA ensemble forecast to a discrete prediction the mean of the ensemble was used. Subsequently, monthly averages of the daily RMSE, MBE and MSESS were used for comparative purposes.

The MSESS was calculated (Equation 2-8) by determining the daily MSE of the discrete forecasts,

$$MSESS = \left[1 - \frac{MSE_f}{MSE_c}\right] \quad (2\text{-}8)$$

where MSE_f is the mean square error of the forecast and MSE_c is the the mean square error of climatology. Climatology was calculated in the same manner as stated for probabilistic forecasts. It can be seen from Equation 2-8 that the range of MSESS is

from 1 (best case) to $-\infty$. A negative MSESS implies that it would be better to use the climatological forecast over that of the modeled forecast.

The MBE determines if there is a bias in the modeled forecast when compared to observed precipitation records by calculating the mean difference between forecast and observed data (Equation 2-9).

$$MBE = \frac{1}{N}\sum_{i=1}^{N}\left(P_{obs} - P_{forecast}\right) \quad (2\text{-}9)$$

where; N is the number of forecasts, and P is the observed and forecasted precipitation amounts. It should be noted that it is possible to obtain low values of MBE while the forecast has high RMSE. This would occur if the forecast deviated from the observed values by both higher and lower readings. Therefore it is a common practice to present this evaluation metric along with the RMSE. The RMSE is not affected by low and high differences cancelling out due to the difference being squared in the calculation.

Results and Discussion

NA Method Results

Initially the overall skill of the AM, using Hamill's (2006a) recommended values of a 91 day search window and 75 analogs with 24 hour accumulated precipitation was assessed using the CRPSS. It can be seen in Figure 2-1 that there is improved skill of the AM forecast in comparison to climatology particularly in the cooler dry period of Dec through Mar and with short lead times. By lead day 5 there are three months (May, July, and September) that show a negative CRPSS value which is indicative of the AM being less skill-full than using a forecast based on climatology. However the AM showed that it has comparatively better forecasting ability from Nov through Apr up to lead day 7.

In order to determine if there are input parameter variations that would improve the AM usefulness in the Tampa Bay region, the search window size, number of analogs used, and the accumulation period were changed individually. Figure 2-2 shows the annual average CRPSS for lead days 1 through 7 using a variable search window size, from +/- 10 days to +/- 75 days while maintaining all other input parameters. Lead-day 1 and 2 show peak skill scores using a 61 day window while 3-7 peak at 15. Once a peak is reached, lead-days 2 and 4-7 show decreasing CRPSS scores. Lead-day 1 and 3 maintain their peak score to the next higher search window size and then they subsequently show declining skill.

While maintaining the search window size at +/- 45 and a 24 hour accumulation period, variations in the # of analogs used show a trend in CRPSS values that is similar to that of search window size variation (Figure 2-3). For lead day 1, the lowest CRPSS is produced when using the smallest number, 10 analogs, with increasing skill until it plateaus at 45 analogs. From this point the skill remains consistent with a marginal drop all the way to 200 analogs. For the subsequent lead days, the skill shows marginal improvement after 75 analogs all the way to 200 analogs, with no drop. The CRPSS for lead day 7 shows negative values when less than 45 analogs were used. The number of analogs and search window size were varied in all combinations and analyzed in the same manner using search windows of 31, 91, and 120 days and a number of analogs ranging from 45 to 150 at intervals of 15. It was found that using a search window size of 61 days and 125 analogs showed the most overall (annually averaged) improvement in skill.

Figure 2-4 shows the effect of the time of accumulation on the forecast skill using +/- 30 day search window and 125 analogs. The results show that 48 and 72 hour accumulation at times in the year had more skill than the 24 hour accumulation forecast. The skill then falls as the accumulation period gets longer. The week 2 forecast, that is the forecast based on accumulation from lead day 8-14, had little to no skill compared to the forecast based on climatology. This can be seen from the corresponding negative CRPSS values for the majority of the year.

Although the CRPSS shows the AM has forecast value compared to that of climatology overall, it is a skill score based on a continuous model; therefore it can't show trends in the skill for different precipitation thresholds. The BSS, a discrete verification measure, does give us this insight. Using the optimum input parameter values, the monthly averaged BSS for a 2.5 and a 25mm threshold can be seen in Figure 2-5. The BSS shows the same trend as the CRPSS in that the highest skill occurs in the cold, dry months and shows a significant decline in the wet summer months, as well as a decrease in skill as the AM forecasts progress temporally (lead-day 1-7). The AM also shows an inverse relationship in comparative forecast effectiveness in relation to threshold values in that the BSS decreases significantly as the threshold increases from 2.5mm to 25mm. This is indicative of better skill in forecasting the occurrence of precipitation compared to larger events.

The BSS has shown the AM has a reduced effectiveness as a precipitation forecasting tool with respect to increasing threshold values yet it doesn't specifically address the reliability or resolution of the output specifically. The relative reliability and resolution of the AM using a 24 hour accumulation, 61 day search window, and 125

analogs as input parameters is presented in Figure 2-6. Both the 2.5 and 25mm thresholds are represented. It can be seen in Figure 2-6 by their proximity to the perfect reliability line, that for lead day 1 both the 2.5 and 25 mm thresholds show similarly good reliability while slightly over forecasting the events. It should be noted that in some instances the reliability curve will not propagate through all probability bins (i.e. 25mm lead-day 7), this is due to the fact that the forecasts did not produce higher forecast probabilities. Therefore these bins do not show data points. This plot incorporates bias into the calculation therefore we cannot necessarily see potential skill that can be obtained if the forecasts were calibrated. Figure 2-6, the ROC diagram, by its nature has no bias incorporated and it can be seen that for lead day 1 there is good resolution (area under the curve) for both the 2.5 and 25mm thresholds. The resolution is significantly reduced for the 2.5 and 25mm thresholds of lead day 7. This lowered potential skill coincides with the previous diagrams that showed an inverse relationship between skill and threshold value. In comparison to the 24 hour results discussed, the weekly accumulation runs presented in Figure 2-7 shows good reliability. The resolution presented in the ROC diagram shows similar trends are found for 24 hour accumulation forecasts.

A determination of the role reliability and resolution played in the significant reduction of skill between winter and summer, as identified by BSS and the CRPSS is presented in Figure 2-8 and Figure 2-9. Figure 2-8 shows the Reliability and ROC diagrams for forecasts run using identical input parameters as Figure 2-6, however only data from the seasons, Jan-Mar and Jun-Aug respectively were used in the calculation of reliability and ROC plots. Figure 2-8 (winter) indicates similar reliability trends to

those produced using all available forecasts. Specifically, lead day 1 shows good reliability for both low and high thresholds. Lead-day 7 again shows good reliability, but indicate lack of forecasts in the higher probability range. Figure 2-8 B shows that the AM has very good resolution forecasts for lead-day 1, both low and high threshold, lead-day 7 shows a substantial drop. For summer, Figure 2-9 indicates similar reliability to winter, however the resolution is substantially decreased for both lead-day 1 and 7. In fact while lead-day 1 shows that the AM can resolve some low and high threshold events, lead-day 7 shows virtually no resolution at all. It is also noted that in both summer and winter, day 7 forecasts don't propagate across the entire plot. This is indicative of the limited ability of the forecast model to give high probabilities at long lead days. These results indicate that a drop in skill from the winter to the summer months is influenced by lack of resolution.

Discrete Forecast Comparison Results

Comparative assessment is shown for lead day one and lead day 5 due to the marginal skill found for NAs for later lead-days in the first section of this study. Figure 2-10 and Figure 2-11 represent a lead-day (1 and 5) comparison of the RMSE and MBE of the NA, CA, and LI downscaling techniques for lead-day 1 and lead-day 5 respectively. It can be seen that linear interpolated reforecast data showed the greatest bias with predictions in the temporal period changing from the dry season (winter) to the wet season (summer) consistently being higher than observation. Conversely seasonal transition from wet to dry showed consistent under prediction of precipitation amounts. From the boxplots it can be seen that lead day 1 NA and CA showed very little bias in forecast prediction with NAs having a smaller spread in daily MBEs. This is to be expected in that the Analog method accomplishes both downscaling and bias correction

of the reforecast data set. An increase in bias error is seen in lead day 5 forecasts for both Analog forecasts with CAs showing a higher bias with no monthly temporal trend present.

Comparatively the RMSE of the CA and NA methods for lead-day 1 and 5 are similar in trend and magnitude with significantly increase in error in the summer wet season and, as expected, performed better than LI forecasts. Figure 2-10 also shows that the NA method performs slightly better from June to September for both lead days. Interestingly, the LI method did not follow the same trend lead day increase in RMSE. The LI RMSE showed a slight decrease in error, although aside from Lead day 5, July through September it still showed greater error and therefore poorer performance than the analog methodologies.

The third discrete evaluation metric used to compare forecasts is the MSESS. In this case each method is compared to a climatological forecast based on observed precipitation. The negative values of MSESS for LI shown in Figure 2-13 shows that using climatology would be a preferred method over LI for all months in the year. The NA and CA methods, however, do show improved performance for lead day 1 and NA shows improved performance in lead day 5. In both instances the NA method outperformed CA, with CA lead day 5 forecasts showed no skill in the summer and very little in the winter. A question arises as to how the mean of a NA forecast ensemble could outperform constructed analogs. The answer may lie in the number of analogs used. When the analog dates are selected by the AM, they are ordered from the best predictor field match to the worst. In this case using an optimized number of analogs for

NAs may have skewed the regression used in the CA method and weighted some of the "worst" analogs higher.

Chapter Summary

This study has evaluated the use of NA and CAs as a means of improving the skill of short term probabilistic and discrete precipitation forecasts locally in the Tampa Bay region. It was found that having access to a large reforecast data archive and with simple selection criteria a regional scale forecast can be developed that has higher skill when compared to climatology. The trend showed a higher skill in the winter months, the dry season for Florida, and a lower skill in the wet summer season. However, the forecast produced was for a 1-14 day lead time yet lead times of 5 or more days tended to have declining or no skill. It was also determined that the BSS in all instances dropped as the precipitation threshold was increased. This may be impacted by the validation method used, BSS, in that the calculation has an indicator value (O) where it is a 1 if the precipitation on that day is greater than the threshold value and a 0 otherwise (Equation 2-6.). This in effect gives the BSS more weight at wet grid points than at dry grid points (Hamill and Whitaker, 2006a; Juras, 2000). Therefore raising the precipitation threshold would in effect create "drier" days and less weight.

There are virtually an unlimited number of variations that can be implemented using the AM, considering you can vary the search window size, number of analogs used to build the observational forecast ensemble, and accumulation period. Therefore this study varied the input parameters individually rather than concurrently. In so doing, it was found that the optimum setup was 125 analogs, a search window of +/- 30 days and a 48 or 72 hour accumulation period. The last item actually is more dependent upon the needs of the end user. These finding points out the limitations of the analog

method if one were to use a static set of input parameters across variable spatial and temporal domains.

Comparison of NA to CA and LI showed both Analog methods outperformed LI as well as forecasts based on climatology. The analog methods were able to directly remove bias from the GFS reforecast predictions which is seen by their low monthly MBE. NAs outperformed CAs in reducing RMSE and increasing MSESS. Case in point, for lead day 5 forecasts the MSESS comparison showed NAs were the only downscaling methodology that showed positive scores. In this study the number of analogs used was optimized for NAs and a comparison was made using the same number of analogs for CAs. This number may not have been optimum for CAs. CAs are created by regressing the reforecast predictor fields of the 125 analog dates to that of the date of forecast. In this way each analog date is given weight. It is possible to have different regression coefficients depending on the method of calculation. Therefore it is possible that the calculated CA forecast applied more weight to analog dates that were not the best match, as all analogs used will have a value. If the number of analogs used were reduced from the optimum for NAs, then the calculation of CA regression coefficients would be limited to better analog matches. It is recommended that an optimization of the number of analogs used for CA forecast calculation be performed.

Table 2-1. Temporal range and spatial location of rain-gauge sites used in this study.

ID Num	Starting Date	Ending Date	Site Name	Lat	Lon
1	10/31/1899	Present	Arcadia	27.218	81.874
2	12/31/1968	Present	Archbold Bio. Stn.	27.183	81.35
3	3/31/1892	Present	Avon Park	27.601	81.509
4	12/31/1891	Present	Bartow	27.829	81.743
5	4/30/1965	Present	Bradenton	27.447	82.501
6	6/30/1918	7/23/2012	Brooksville	28.616	82.366
7	3/31/1918	Present	Bushnell	28.664	82.09
8	6/30/1948	Present	Clermont	28.455	81.723
9	9/1/1973	Present	CYCC3	28.301	82.383
10	7/31/1948	Present	Fort Green	27.567	82.133
11	12/31/1891	Present	Fort Myers AP	26.583	81.867
12	1/31/1899	Present	Inverness	28.803	82.312
13	9/1/1948	Present	Kissimmee	28.283	81.411
14	8/9/1943	Present	Myakka Rvr. St. Park.	27.242	82.316
15	5/15/1952	Present	Orlando Intl. AP	28.433	81.316
16	7/31/1948	Present	Parrish	27.583	82.422
17	6/1/1958	Present	Sanford	28.815	81.278
18	2/21/1895	Present	St. Leo	28.336	82.261
19	12/31/1891	Present	St. Petersburg AP	27.727	82.626
20	12/31/1892	Present	Tarpon	27.567	81.817
21	12/31/1932	Present	Wauchula	28.154	82.753

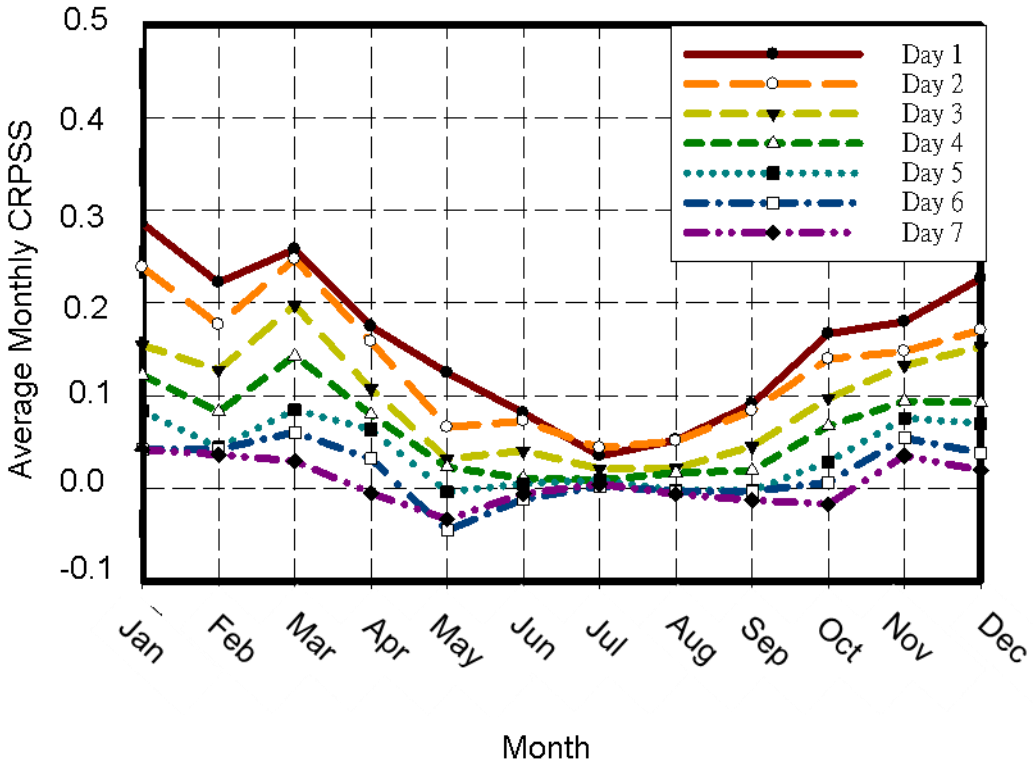


Figure 2-1. Natural analog average monthly CRPSS for lead day 1 through 7. Results are based on simulations using 75 analogs, a 91 day search window and a 29 year (Oct 1 1979- Sept 31 2008) reforecast Data Set. Twenty-one rain-gauges were used.

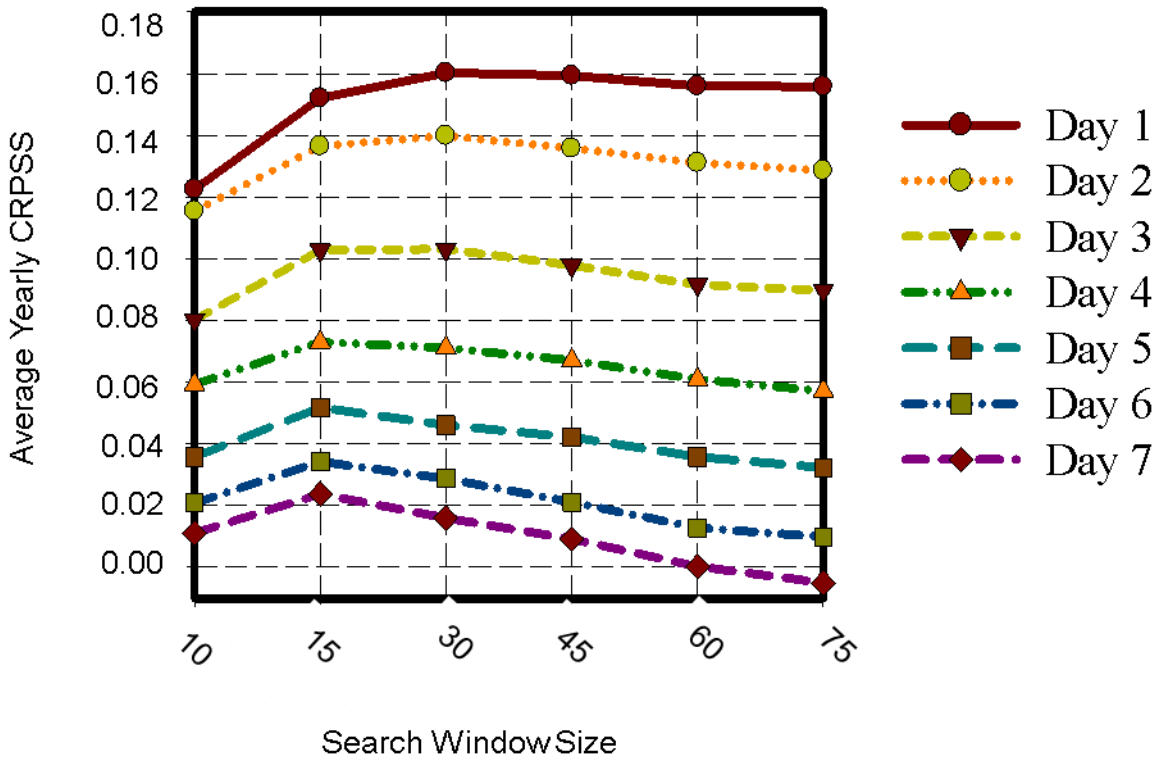


Figure 2-2. Natural analog average monthly CRPSS for lead day 1 through 7 for variable search window size. The results are based on simulations using 75 analogs, a 91 day search window and a 29 year (Oct 1 1979- Sept 31 2008) reforecast Data Set. Twenty-one rain-gauges were used.

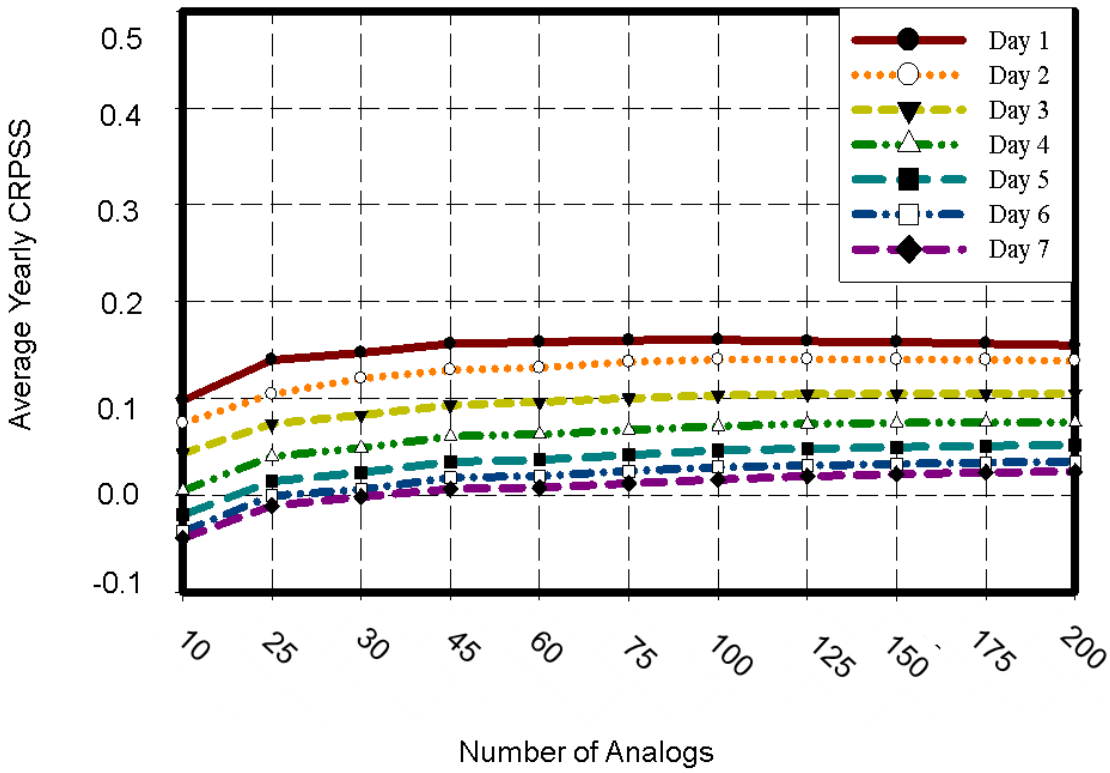


Figure 2-3. Average daily CRPSS for lead day 1 through 7 for variable number of analogs used. The results are based on simulations using a 91 day search window and 45 day search window on a 29 year (Oct 1 1979- Sept 31 2008) reforecast Data Set. Twenty-one rain-gauges were used.

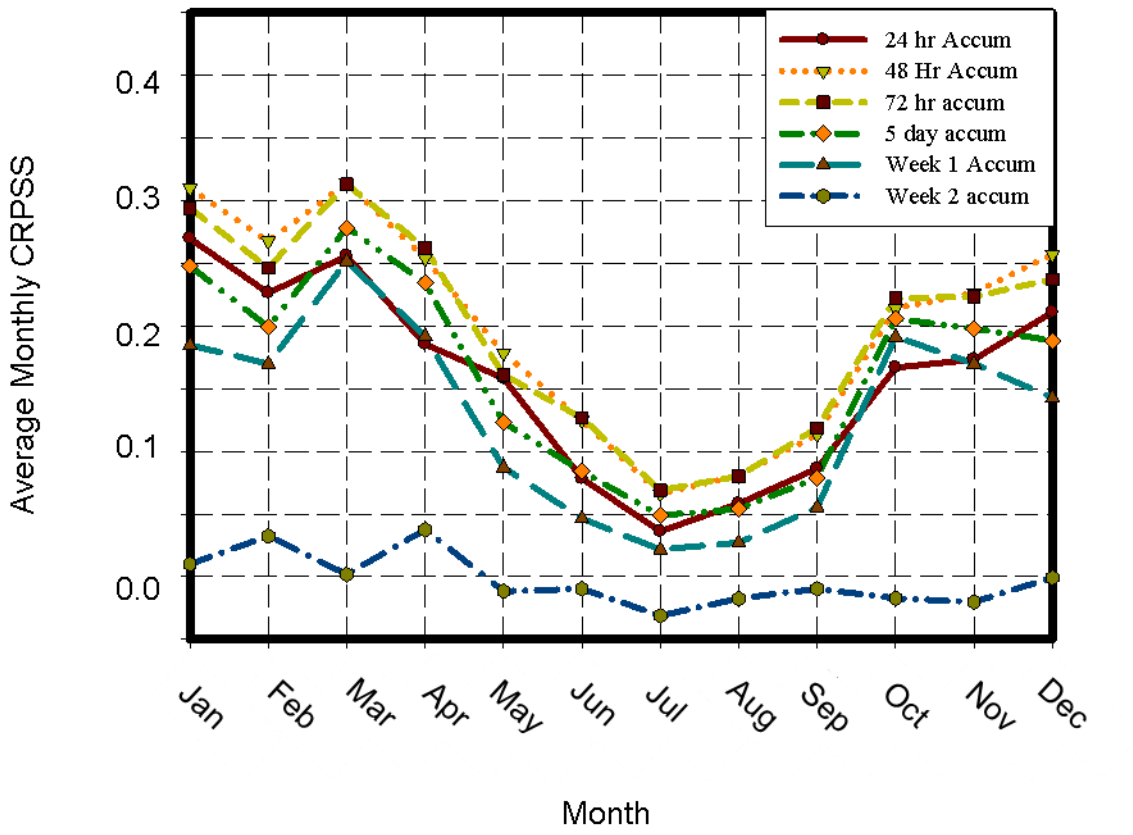


Figure 2-4. Average monthly CRPSS for variable accumulation periods. The results are based on 125 analogs, and a 61 day search window on a 29 year (Oct 1 1979- Sept 31 2008) reforecast Data Set. Twenty-one rain-gauges were used.

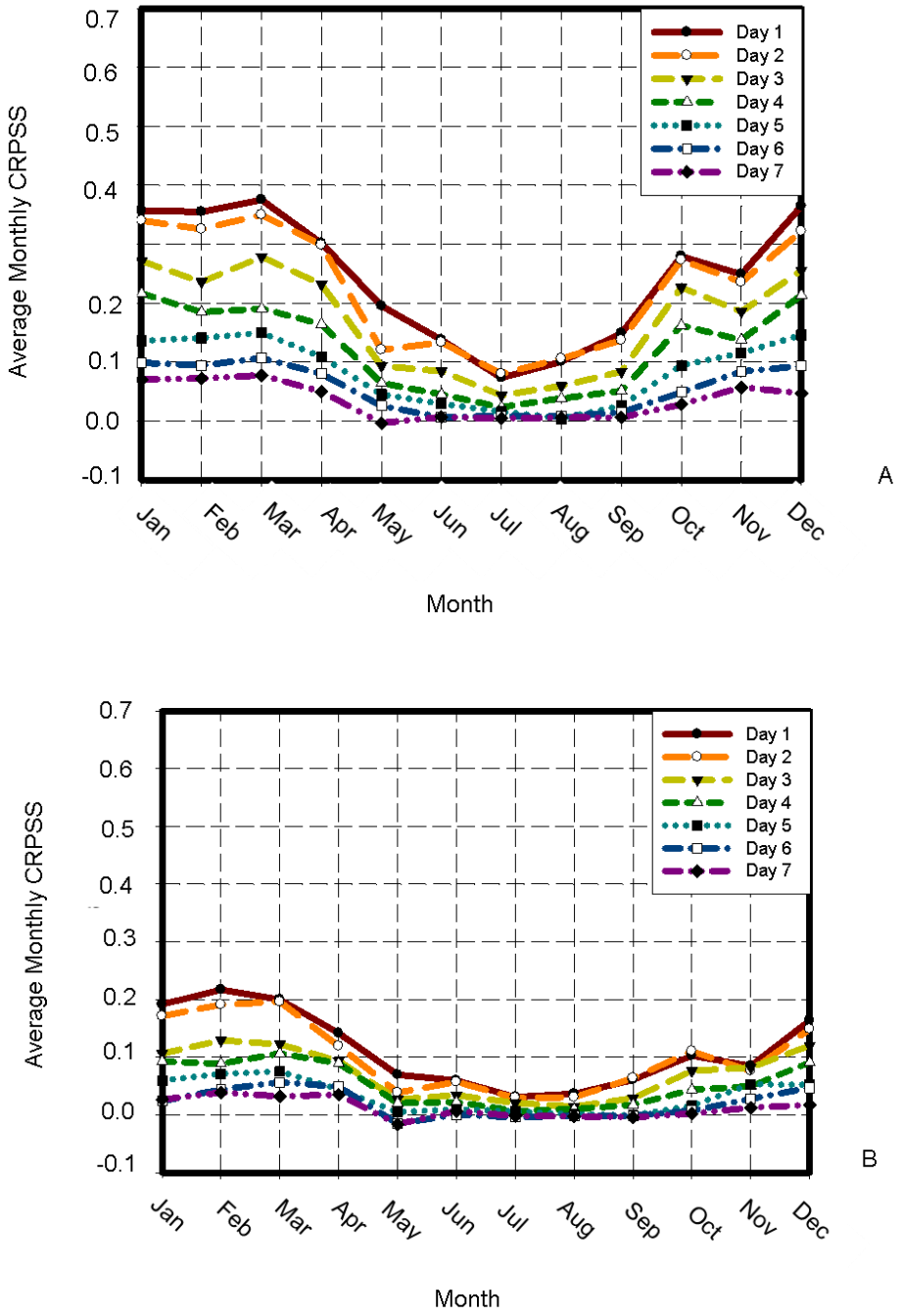


Figure 2-5. Natural analog average monthly BSS for lead day 1 through 7. The results are based on simulations using 125 analogs, a 61 day search window and a 29 year (Oct 1 1979- Sept 31 2008) reforecast Data Set. 21 rain-gauges were used. A) 2.5mm BSS threshold. B) 25mm BSS threshold.

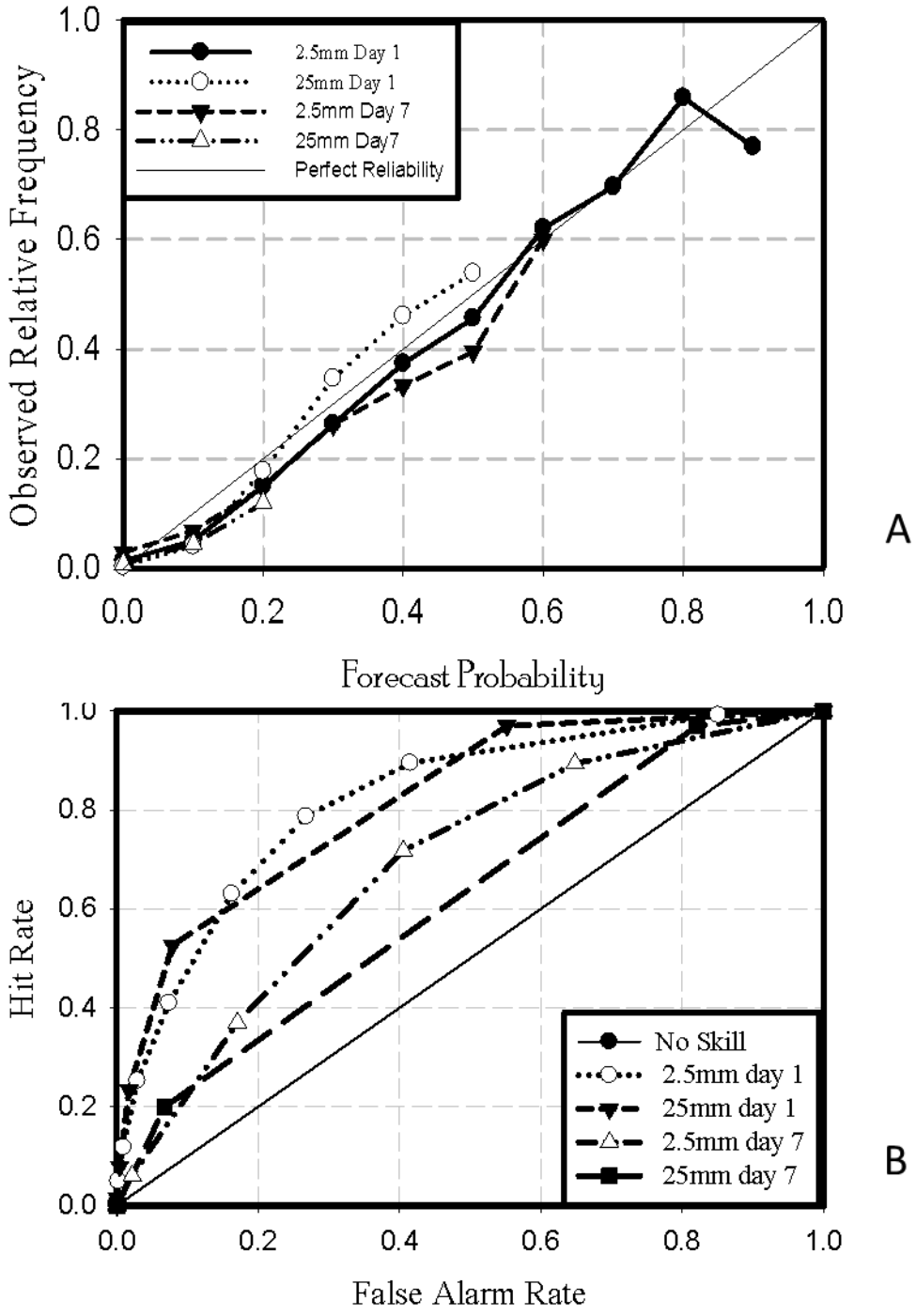


Figure 2-6. Resolution and reliability results for AM forecasts using 24hr accumulation, 125 analogs, and 61 day search window. A) Reliability diagram. B) Relative operating characteristic plot.

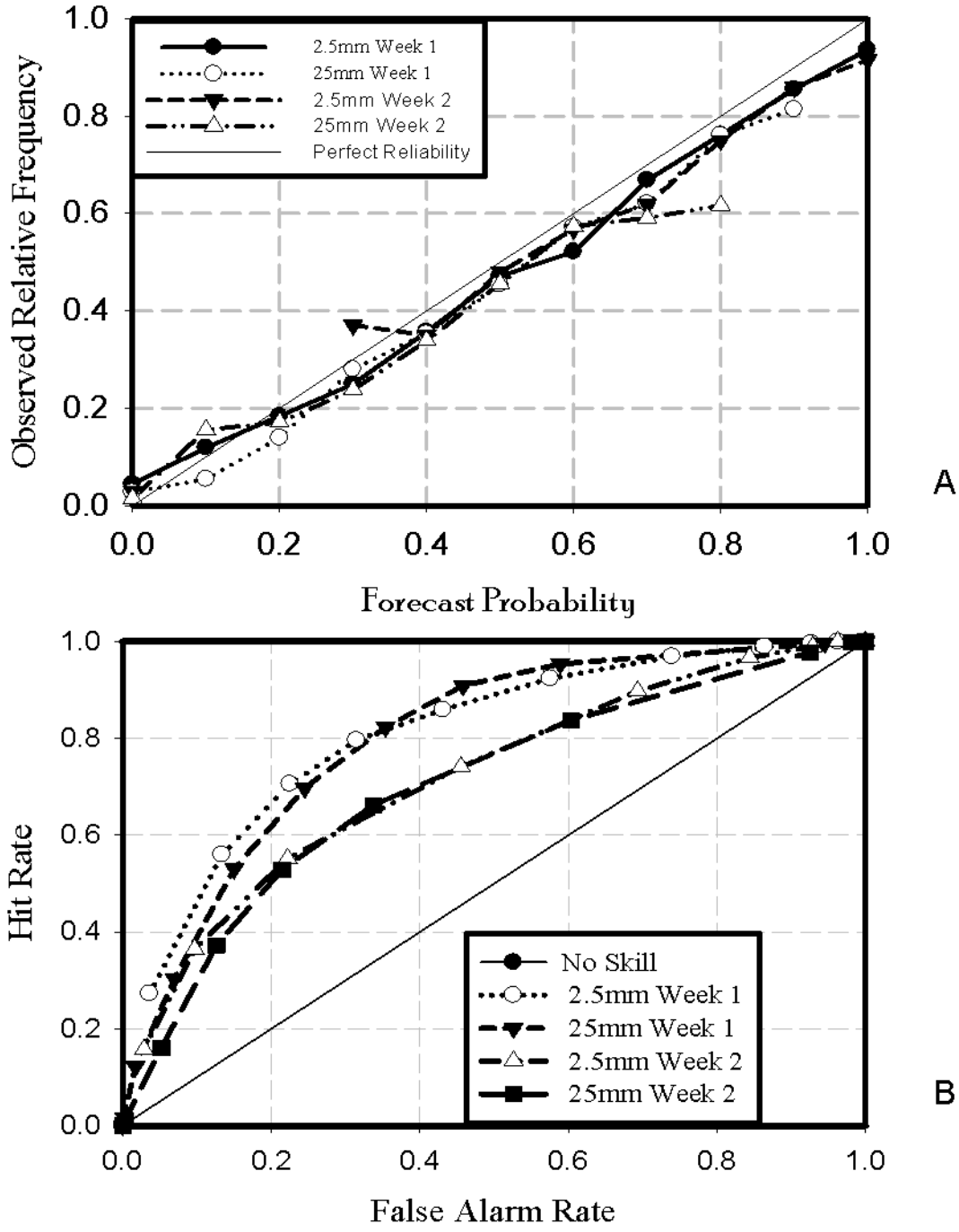


Figure 2-7. Resolution and reliability results for AM forecasts using weekly accumulation, 125 analogs, and 61 day search window. A) Reliability diagram. B) Relative operating characteristic plot.

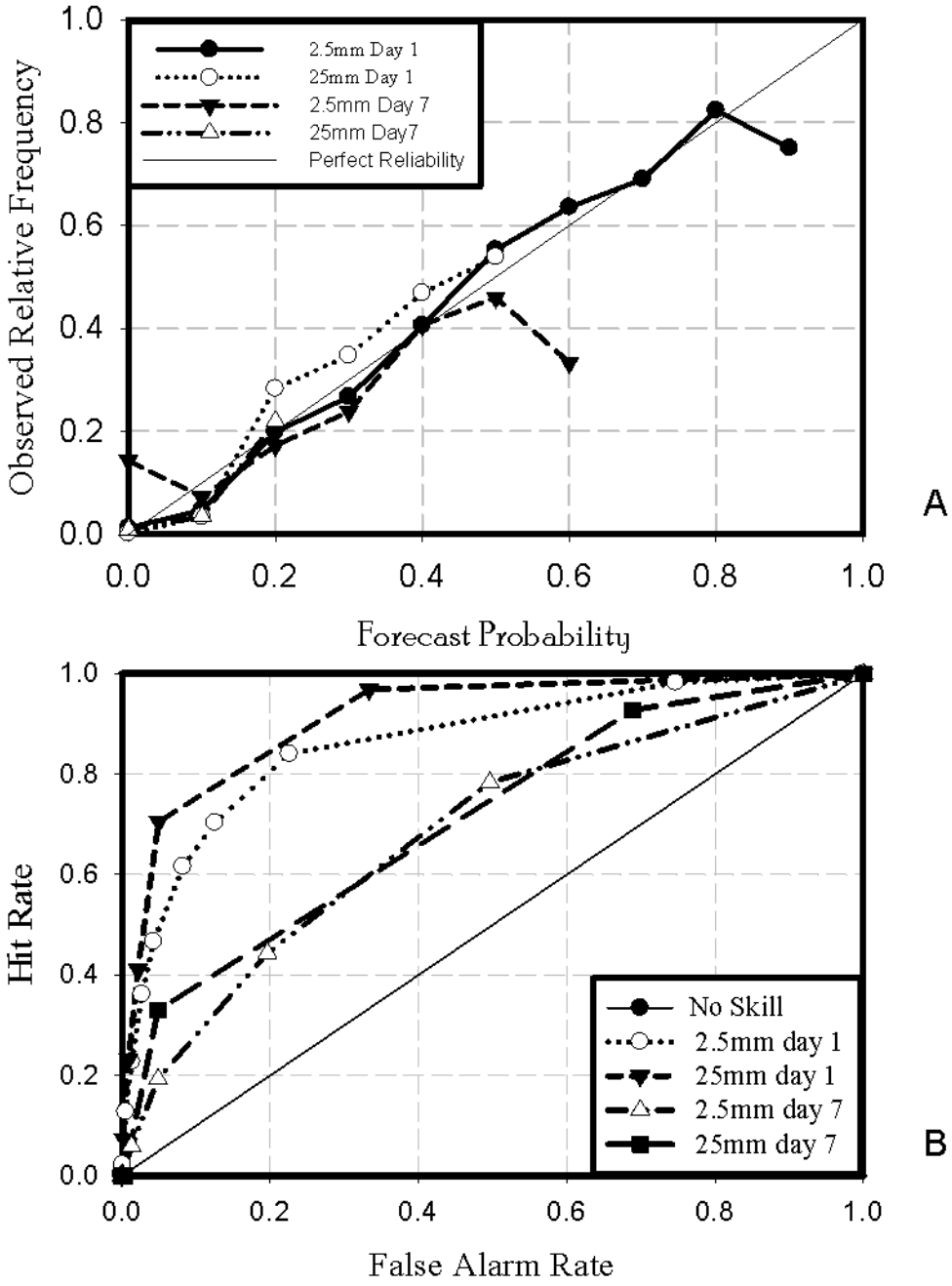


Figure 2-8. Winter (Jan., Feb., and Mar.) resolution and reliability results for AM forecasts using 24hr accumulation, 125 analogs, and 61 day search window. A) Reliability diagram. B) Relative operating characteristic plot.

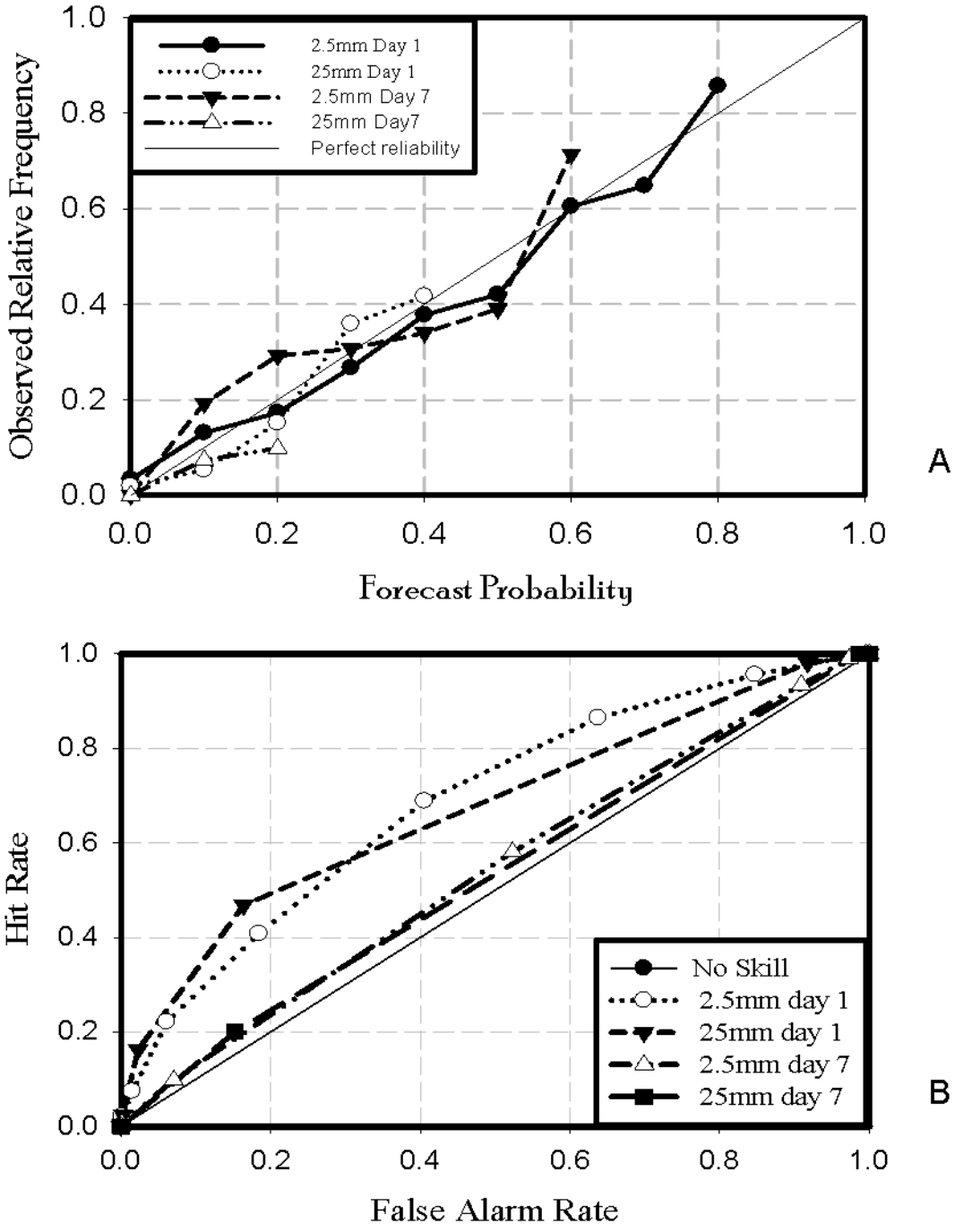


Figure 2-9. Summer (Jun., Jul., and Aug.) resolution and reliability results for AM forecasts using 24hr accumulation, 125 analogs, and 61 day search window. A) Reliability diagram. B) Relative operating characteristic plot.

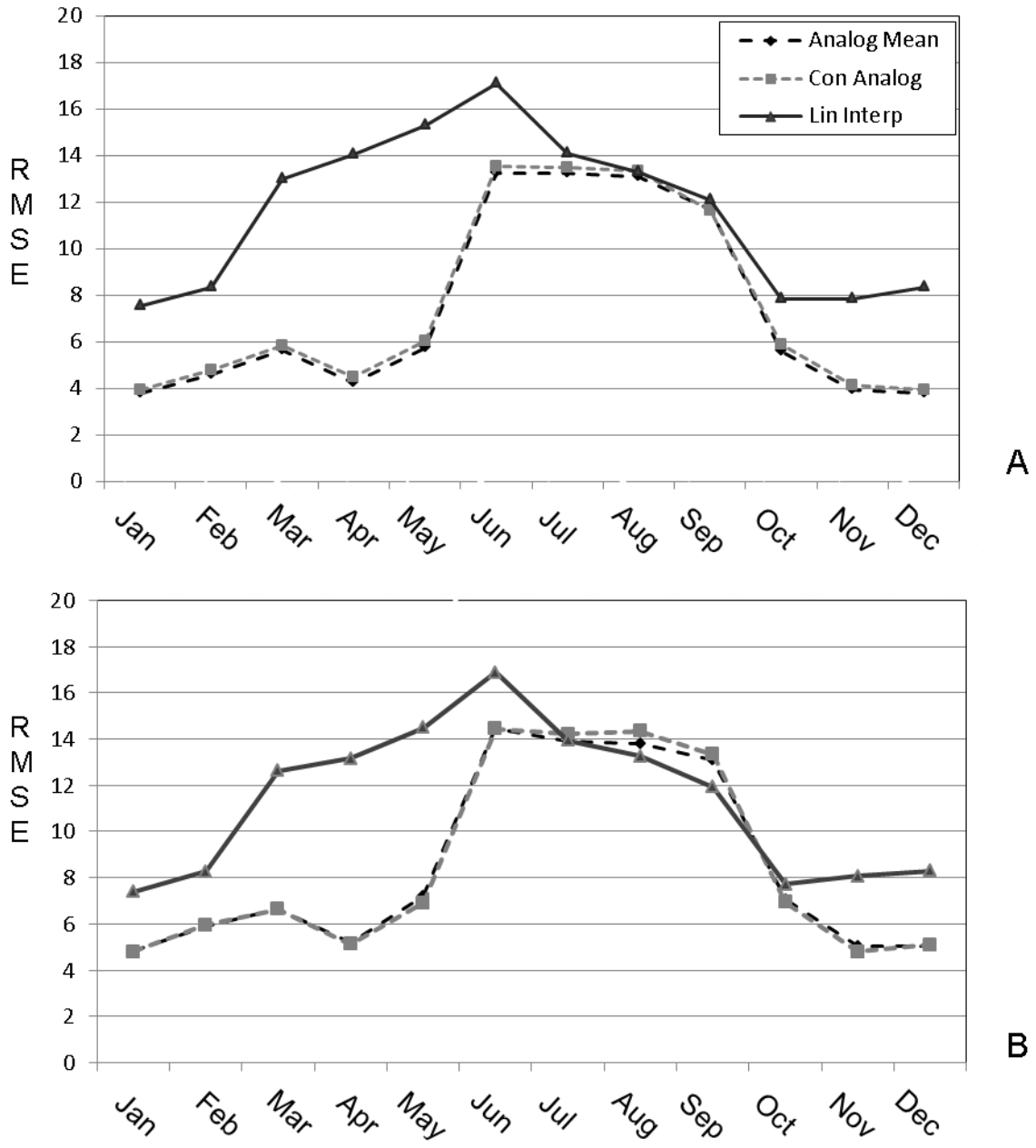


Figure 2-10. RMSE results for the LI, CA, and NA downscaling methodologies. A) Lead-day 1. B) Lead-day 5.

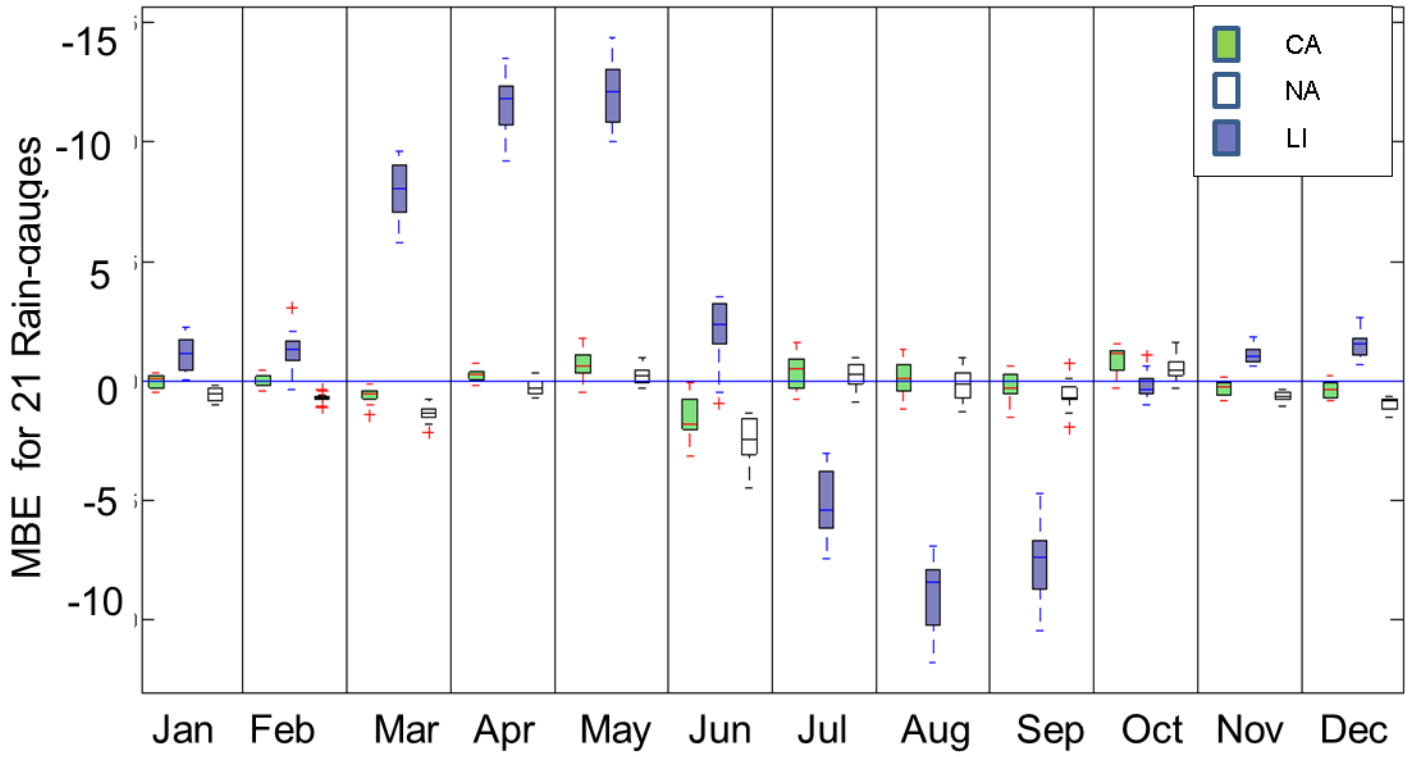


Figure 2-11. Lead-day 1 monthly MBE results for LI, NA, and CA downscaling methodologies.

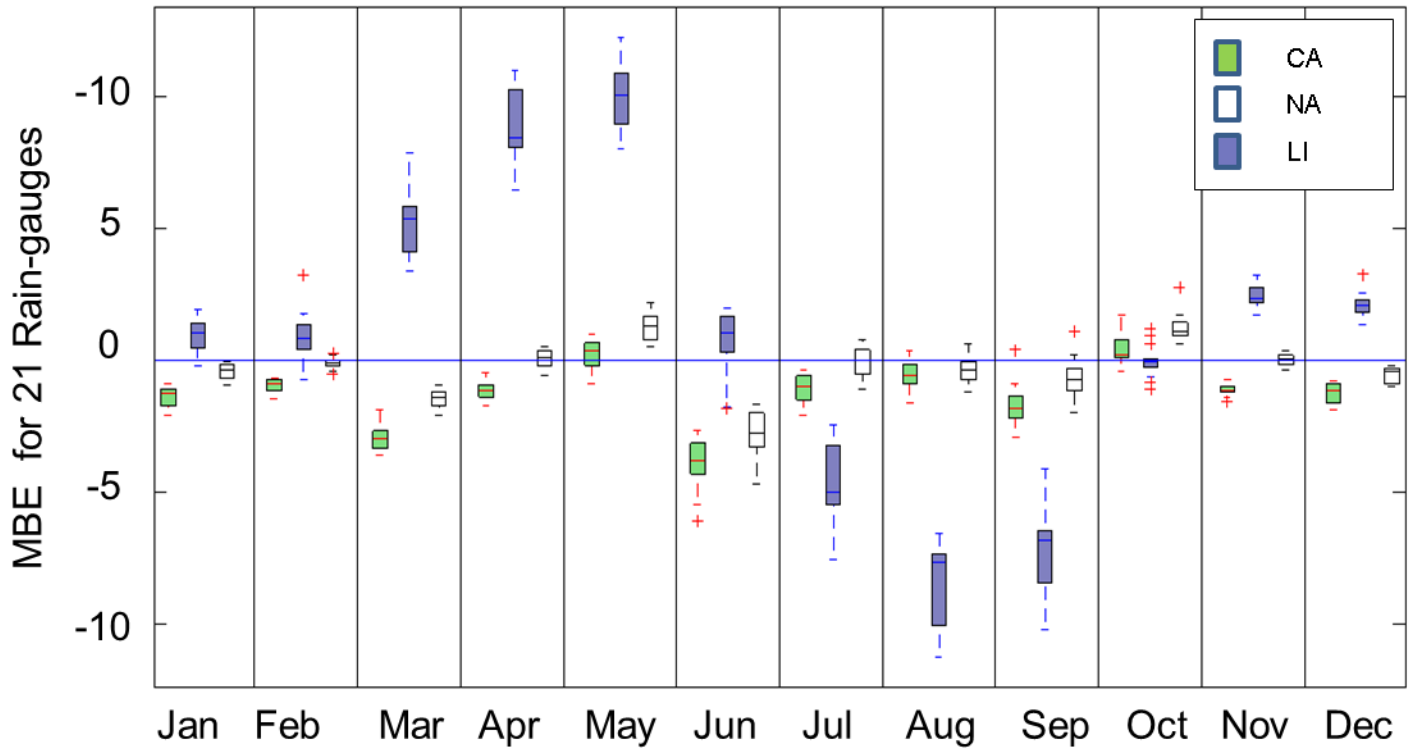


Figure 2-12. Lead-day 5 monthly MBE results for LI, NA, and CA downscaling methodologies.

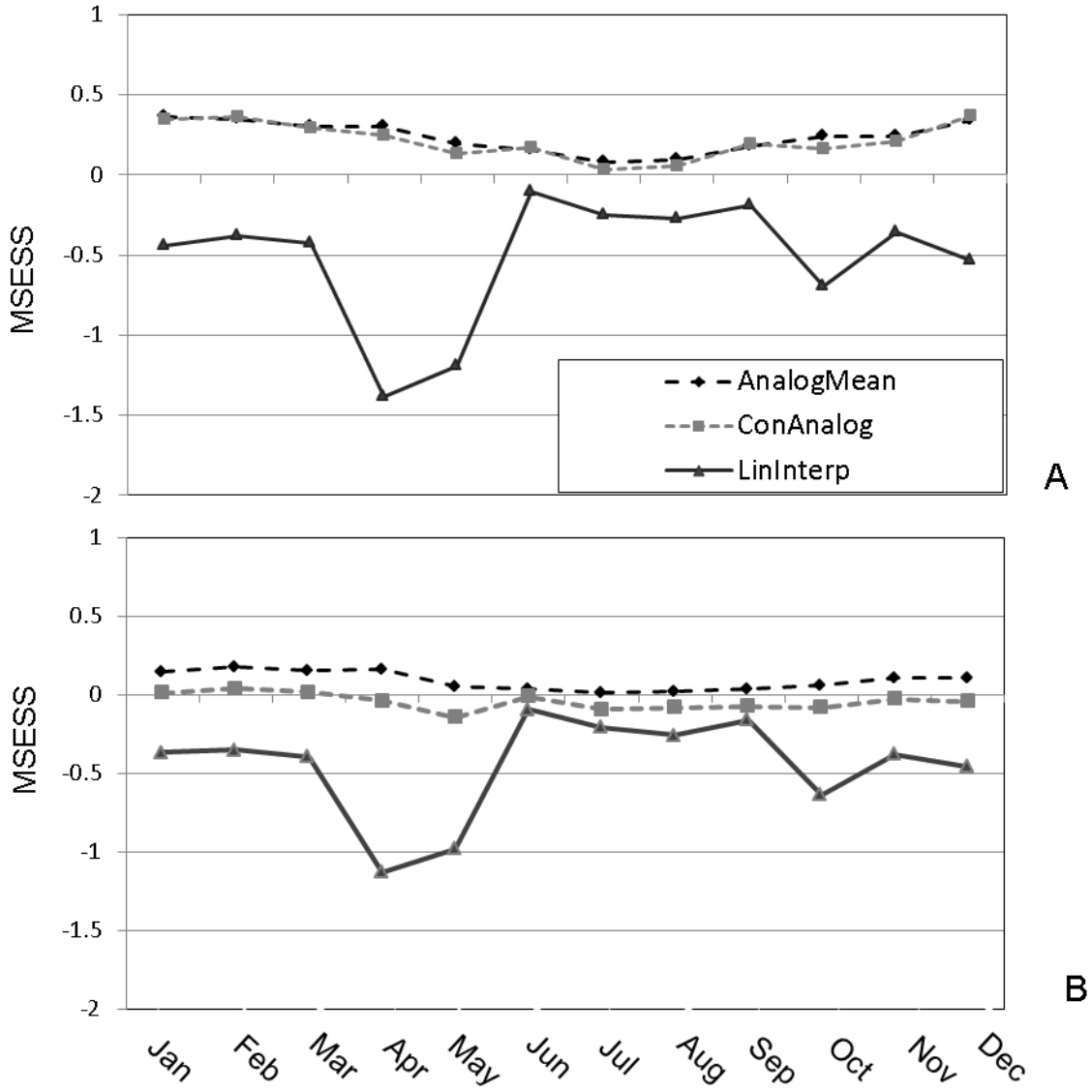


Figure 2-13. MSESS results for the LI, CA, and NA downscaling methodologies. A) Lead-day 1. B) Lead-day 5

CHAPTER 3
IMPLEMENTATION OF ANALOG METHOD FORECASTS USING TAMPA BAY WATERS PROPRIETARY ARTIFICIAL NEURAL NETWORK STREAM FLOW MODEL

Background

In this study weekly precipitation forecasts produced from an AM are incorporated into a proprietary based Artificial Neural Network hydrologic stream-flow model (SWANN) in order to improve its one week lead time stream-flow forecasting potential. Comparisons are made to the original model precipitation forecast settings and inputs based on Natural Analog (NA), Constructed Analog (CA), and forecasts based on Linear Interpolation (LI)

There are two basic needs that drive the improvement and use of hydrologic surface water models: a) to predict the temporal nature of local floods as well as their magnitude, and to determine the long and short term availability of water. This study is an example of the latter. In order to meet the water consumption demands of over two million customers, Tampa Bay Water has at its disposal surface water, groundwater, and desalinated water for this purpose. However, even with what some would consider an excess of available water, decisions have to be made regarding how much water should come from each source. These decisions are based on regulatory, physical and financial constraints. Regulations stipulate a maximum allotted volume that can be taken from ground water after which the water management company is heavily fined. Desalination on the other hand is not regulated in this way however the desalination process is prohibitively expensive. Surface water, fed to a treatment plant, is the most economically friendly, and the most desirable choice of the three. Decisions on the proportional contribution of these sources are made four weeks in advance with a weekly revision based on current and predicted conditions. One of the primary drivers

of these plans are the volumetric short term prediction of water available for extraction from the Hillsborough river (Asefa, 2009). Additionally if an excess of water is available, the reservoir can be recharged for use in times of surface water scarcity such as winter. Hydrologic modeling is used to aide in these decision making processes.

There are two main categories of surface water models; stochastic and numerical. Numerical models, also called process driven models, are designed by identification and mathematical incorporation of physical processes in the domain that are representative of the hydrologic cycle. Such processes include, but are not limited to, surface water flow, evapotranspiration, surface to groundwater interaction and soil storage. This type of model by its nature is complex and parameterization of variables in a watershed is difficult.

Statistical models are based on a black-box approach. The watershed is considered a control volume with inputs and outputs. Potential inputs are precipitation, groundwater and surface water flows. In general, the output of the control volume identifies the volumetric flow rate of a river or stream selected depending on the information pursued by the modeler. Physical relationships that affect the propagation of water from the input the output to the output are not specifically taken into account; instead behavioral links between the input(s) and output (s) are identified and these relationships are modeled. Three common techniques have been used to model these linkages: transfer functions (Astatkie and Watt, 1998), regression (Zhu and Day, 2009), and Artificial Neural Networks (Govindaraju and Artific, 2000b). In this study the stream-flow model is an ANN that is conceptually based on the General Likelihood Uncertainty estimation approach (discussed in SWANN Model Theory section).

Artificial Neural Network Theory

An Artificial Neural Network (ANN) is a computational model initially developed over 60 years ago in order to understand the functioning of the human brain (McCulloch and Pitts, 1943). Although these models are considered stochastic, the modeler does not apply statistical or mathematical representation to the model, this type of model is trained. Training is accomplished by processing an input(s)/output paired data series, in a teaching mode, and the model identifies input/ output relationships and stores this information. Subsequently when new data is processed the model is capable of identifying the appropriate linkages and uses them to produce representative output. It can be seen that this is a true black-box; even the statistical linkages are unknown. This process is similar to the steps the human brain uses to process and predict information, which is in fact where this field of computational modeling had its roots.

The structure of an ANN is based on a group of nodes or neurons that receive multiple inputs (from other nodes or from outside inputs) and when activated channel their stored piece of information down-stream to a receiving node, where eventually the chain leads to model output data. A single node produces output based on the weighted sum of all of its inputs according to its own activation function (Dawson and Wilby, 2001). A visual representation of a simple 3 layer neural network can be seen in Figure 3-1. Nodes that have access to external information are considered the input layer, nodes that produce output are the output layer and the remaining processing nodes are the hidden layer.

The training of an ANN has three main purposes, storage of input output relationships, refinement of network structure, and in terms of time series modeling, to determine a lag-space relationship of input data. The storage of data is based on

internal manipulation of the nodal weights and activation functions. When the changes achieve a threshold where a comparison (mean square error) of the simulated output to the observed output reaches a predefined value, the ANN is considered trained (Cannas et al., 2006). The training methodology should be performed numerous times with a different number of nodes available. In this way the network can achieve optimum results using the least number of nodes. It has been pointed out that having too many nodes in a network provides space for the storage of unnecessary information which consequently can deteriorate optimum performance, while too few nodes and there may not be enough information stored to formulate an effective result (Govindaraju and Artific, 2000a).

The third aspect of training, for time-series ANN models is the identification of the lag space. Lag space in an ANN can be conceptually thought of as the residence time in a numerical hydrologic model. The input data from the previous time step may not have a relationship with the current time steps output. The linkage may be, due to residence time, from three steps into the past. Therefore training of time series is accomplished with a window of input data from previous time-steps. In this way the ANN can identify how far back the network has to reach in order to correctly represent the input output links. It follows that each input will have its own representative lag term, together called lag space.

Although ANN models have been viewed by physical modelers as black boxes due to their structure, several modeling benefits have been identified (Dawson and Wilby, 2001; Govindaraju and Artific, 2000a; Zealand et al., 1999). First they have the ability to empirically identify the relation between input and output data without any prior

knowledge, data preprocessing or explicit physical considerations. Second they have the ability to adapt their structure when given a set of inputs and a known output in order to produce consistent results. Third, ANNs have the ability to make generalizations when the data is incomplete, noisy, or distorted by using previous examples. Fourth they are non-linear which gives them better capability to solving problems than linear techniques, and due to the parallel nature of similar nodes, they can computationally run faster.

SWANN Model Theory

SWANN is the proprietary ANN stream-flow forecast model used by Tampa Bay Water to predict weekly surface water volumetric flow rates for four streams within the contributing watershed. This information is used to aide in the source allocation planning and decision making process. Conceptually this model is a change from earlier versions of stream-flow ANN models in that the input parameter and lag space uncertainties are handled in a manner similar to the GLUE method.

The GLUE method of evaluating uncertainty is based on the proposition that there is no one correct model realization that is optimum. Instead, there is an ensemble of model conditions that can produce optimal output. This is accomplished using Monte-Carlo simulations propagated by samples generated from input sample PDFs. The model results are then analyzed by comparing the output data to observed data using a likelihood measure. Based on the measure, multiple acceptable results can be identified. SWANN follows the same core principles (Asefa, 2009).

SWANN output is produced using an ensemble of ANN results, based on the number of inputs, variations in lag space and number of nodes. The development of SWANN (Asefa, 2009) required four distinct phases: selection of relevant inputs,

determination of optimum lag space, Monte-Carlo simulation and assessing the acceptability of results using a likelihood measure. For the first step all of the available rain-gage, groundwater and stream flow observation stations were used. The reason none were omitted is that this GLUE based methodology can discriminate between influential and non-influential inputs based on likelihood.

The optimum range of lag space was determined by using a technique (He and Asada, 1993) that provide a measure, Lipchitz quotient, of the influence of independent to dependent time series data. By repeating the test with different lags a relationship of lag space to influence can be determined. Once the lag space, input relationship is known Monte-Carlo sampling is implemented. The Monte-Carlo samples are based on the input factor PDF with their corresponding lag and the size of the ANN in terms of number of input nodes as well as the number of hidden nodes. The number of input nodes limits the number of inputs that can be used, which as stated before, removes the burden of eliminating inconsequential inputs during the selection process. Monte-Carlo samples then undergo ANN training. The output from the trained networks, are then used in the final step, the likelihood comparison of ANN outputs to observed data. The likelihood measure provides an indication of the optimal ensemble of ANNs to use for a specific output. The likelihood measure used is similar to the Nash Sutcliff efficiency measure and was calculated according to Equation 3-1.

$$\mathrm{L}(\Theta_i|Y) = \left(1 - \frac{\sigma_i^2}{\sigma_{obs}^2}\right)^n \quad (3\text{-}1)$$

where; $\mathrm{L}(\Theta_i|Y)$ is the likelihood, and σ_i^2 is the variance of error for output from ANNi and σ_{obs}^2 is the observed variance. Based on the likelihood measure the members of the

ANN ensemble for a given output are then weighted as to their value. These weights along with the optimum ANN ensemble are imbedded in the SWANN model.

In ANN hydrologic forecasting models there are times when input factor data is not available because it has not occurred yet. In this case forecasted data either climatology or from a different model is used. SWANN introduced a simplified K nearest neighbor routine to forecast these values. The data used to identify nearest neighbors has a limited temporal range of 3 years and available data has intervals of 7 days. The theory of this limitation being that when the model is applied it will use the current days forecast for weekly accumulated rain. The aim of this study is to replace improve the forecasted precipitation by replacing the Knn forecast with those produced by the AM.

Objectives

The objective of this work is to increase the skill of a stream-flow model (SWANN) forecasts by introducing weekly accumulation local scale precipitation forecasts using the AMs (CA and NA). The hypothesis is that the artificial neural network stream-flow model although trained using persistence forecasts, can be improved by using single predictor AM forecasts. The work is accomplished by modifying SWANN code to accept natural analog, constructed analog and linearly interpreted precipitation forecasts. Comparisons are then made between the original and the modified model results for high and low flow throughput streams.

Study Area and Data Collection

The SWANN model provides stream-flow forecasting results for Tampa Bay Water based on the contributions to the Hillsborough River watershed. Model inputs consist of; nine rainfall stations, four groundwater level monitoring points, and two

surface water flow gauges within the watershed. The locations of these sited can be seen in Figure 3-2 (Asefa, 2009). The rain-gauge data used by the model are weekly accumulated totals with an interval of 7 days. The locations and temporal range of the data can be seen in Table 3-1. The Analog method(s) forecasts incorporated into the SWANN model use the same data sets as the study performed in Chapter 2. A summary of these rain-gauge sites and the available corresponding data can be viewed in Figure 1-1 and Table 2-1.

SWANN produces four quantitative rate outputs for the streams; Morris Road Bridge (MRB), Cypress Creek, Trout Creek, and the Alafia River. Observed stream-flow data for these four locations was obtained from Tampa Bay Water and had a range of 9/15/07-3/21/2009 (obtained from Tampa Bay Water). A summary of descriptive statistics for the stream-flow flow rates, for the 80 week historical range used by SWANN, is recorded in Table 3-2.

Methodology

Incorporation of the NA, CA and LI forecasts into SWANN consisted of two parts, data filling, and input forecast replacement. It can be seen in Table 3-2 that the temporal range of the rain-gauge data currently used in the SWANN model does not encompass the 30 year range used by the AMs. In order for these new forecasts to be used the current rain-gauges would require some form of temporal expansion. Two approaches were pursued in an attempt to fill the rain-gauge records. The first was when the code encountered a precipitation input with no available record; data was extracted from the mean of the corresponding data from all available rain-gauge sites. This method comes with a potential issue, the spatial variability of precipitation events are not maintained. In an attempt to overcome this issue, the second method extracted

data from an individual site with available data based on a hierarchy of rain gauge data correlation coefficients. The current rain-gauge record has a data interval of 7 days perhaps based on the assumption that when using weekly accumulation it is not appropriate to record a precipitation event twice in the record, which would be the case for a daily record of weekly accumulated precipitation. Using the AM, this does not appear to be the case. When identifying analogs based on weekly accumulations you are looking for a match to the precipitation event, or combination of events to produce the same volume as the current weekly accumulated forecast. There is no reason to believe there is a climatic signal based on the days of the week; therefore the climatic physics in the atmosphere for a week of rainfall starting on Monday has the same chance of matching as a week of rainfall starting on Friday. With this in mind the current rain-gauges were adjusted to a daily data interval using daily precipitation records for the current rain-gauges. When SWANN encounters a missing value from the new data sets it implements the same data extraction procedure cited previously.

The second aspect is the implementation of the new forecasts for use as inputs to SWANN. The SWANN algorithm for determining a precipitation input for each run in the ANN ensemble consists of finding the two best Knn forecasts using a euclidian distance measure between all of the input factors current values and those of the historical archive. This in effect is a type of analog methodology. Once the dates of the two best forecasts are identified, inputs are collected from the record corresponding to the Knn dates plus one week. The two Knn forecast values are then averaged the result of which is an input to an all members of the ANN ensemble. Individual stream-flow forecasts are then produced from each network. The ensemble stream-flow

forecast outputs are then reduced to a single value using the ANN weights (Chapter 1). It is worthy to note that if either date supplied as the two best KNN, then the model defaults the input variables to the previous week's values.

The NA method provides 125 analog dates to SWANN in order to determine ANN inputs. Two approaches were used to replace the Knn forecasts. The first consisted of running the same routine using Knn to produce input parameters, and then replacing the precipitation inputs with data from the current rain-gauges corresponding to a single analog date. This process is repeated 125 times, once for every date in the analog ensemble. As an example, suppose there are 10 ANNs in the network ensemble, then the original approach by SWANN would produce 10 stream-flow forecasts from 1 input set, and 1 weighted average forecast. Using the NA method, 10*125 (1025) stream-flow forecasts would be produced and 125 weighted average forecasts are produced.

The second approach used all 125 analog dates at once. This was accomplished by locating the observed precipitation for all dates in the analog ensemble and using the mean for replacement. This gives a 1:1 relationship between the number of forecasts provided by the original methodology and the NA methodology. For the CA analogs, there is only one analog and this is used in place of the Knn. For LI, no dates were used. Replacement of the KNN was done with the current LI forecast.

Evaluations of the SWANN results were conducted using RMSE and the Nash Sutcliff Coefficient of Efficiency (Ceff). The comparison consists of the mean and weighted average of the forecast ensemble for Knn, Ca, and LI methodologies; and the mean of the forecast for NA as well as the mean of the weighted average forecasts. A

third measure was used to ascertain if there was bias in the model results, the Mean Bias Error (MBE).

Results and Discussion

The SWANN model was executed in its original form using both the original rain-gauge data sets and the dataset with daily intervals. This was done to evaluate if any improvement in the results based on precipitation input replacement was not due to data manipulation. The output was then analyzed according to the RMSE and the CEFF. These results indicated a marginal improvement in the output stream-flow forecast with the original method, for Morris Road Bridge, having an RMSE of 45.32 and a Ceff of 0.458 and the modified data results were 45.29 and 0.462 respectively. As these results are only marginally different, the rest of the study uses the modified (expanded) rain-gauge records for all methods including the original Knn method.

In order to determine whether replacement of missing rain-gauge data should be performed using the mean of available rain-gauge data or by a correlation hierarchy, both methods were implemented while using the mean NA as replacement forecasts. The results indicated that the correlation hierarchy achieved better results. The RMSE and Ceff for MRB using the mean fill method was 4.97 and 0.46 while the correlation fill approach resulted in 44.85, and 0.47. While there are only marginal differences in output error and efficiency, the correlation hierarchy approach is used to maintain precipitation event spatial variability.

Figure 3-3 and Figure 3-4 are plots representing a comparison of the original SWANN Knn approach with that of CAs. This is presented as a comparison of the effectiveness of new forecast implementation in forecasts for stream with a high flow-rate (MRB, Figure 3-3) and low flow-rate (Trout Creek, Figure 3-4). Each plot

represents an ensemble of forecasts, therefore the mean, 25th percentile and the 75th percentile is displayed, along with observed data points. In this way the range of forecasts is shown which, as noted by Asefa (2009), indicates the models ability to overcome some of the input and structural uncertainty. Figure 3-4 indicates that both methodologies are capable of following the behavior of observed rainfall at low flow conditions and are equally capable of forecasting flow-rate changes. It should be noted that while the Knn method was able to identify a peak on April 22, 2008 it was not able to capture the magnitude (360 mgd). While the CA methodology reached the observed flow-rate with the mean of the forecast ensemble. For the low flow-rate Trout Creek results (Figure 3-4) indicate that both methodologies did not capture peak events. The spread of the CA method does indicate that it could potentially be a better technique to reach the peak flows, but the increased spread pulls the forecast distribution away from the low flow conditions.

A summary of the two error measures, RMSE and MBE as well as the Ceff can be seen in Table 3-3 for all four streams. The bold values indicate forecasts based on the weighted average of the ANN ensemble, while the corresponding (normal text) is representative of the mean. The data indicated that the model (using all techniques) does a substantially better job of forecasting for high-flow streams (MRB and Alafia). The Ceff of these two streams a range of 0.46-0.52 and 0.35-0.38, respectively. The low flow streams (Cypress and Trout Creeks) have ranges from -0.13 to -0.06 and from -0.05 to 0.24. Due to the underwhelming performance of the low flow forecasts they are not evaluated further.

The results for MRB indicate that, while all of the new methods perform better than the Knn approach, the LI performs the best in terms of error (RMSE and MBE) and CA has the highest efficiency (marginally better than LI and NA). Trout Creek forecast results follow a similar pattern in that all of the new methods outperformed the nn approach with the CA showing the highest Ceff. However in this instance LI is outperformed by NAs except for the MBE.

Interestingly in both cases (MRB and Trout Creek) using the mean of the 125 NAs for precipitation input caused the MBE to almost double all other methods. On the other hand, running the analogs separately put the MBE in the same range as the other techniques. On first impression it appeared that maybe using the median of the analog ensemble would improve results, the theory being that the tail end of the analog ensemble (poorest matches) show a significant drop in forecasting ability. This would then skew the average value used as an input. When using all 125 analogs individually, the smoothing that occurs with the sheer size of the ensemble counters the analogs with poor ability. To test this theory, the median of the analog ensemble was used. The results showed that it performed worse than the mean of the analogs.

Figure 3-4 and Figure 3-5 illustrate the resolution and reliability of the techniques by plotting the results using a ROC plot and reliability diagram. The representation of the results for MRB are split into three terciles (low, middle, and high). The ROC plot for the lower tercile indicated that for the lower tercile, the NA technique using all of the analogs individually showed the greatest resolution with Knn the least. At the other extreme LI had the best ability to resolve peak events followed by the CA method. The

reliability diagrams show that all of the techniques performed similarly with declining reliability from the lowest tercile to the highest.

Chapter Summary

In this study four precipitation forecast techniques replaced the precipitation inputs for SWANN, Tampa Bay Waters weekly stream-flow forecasting tool. SWANN is an ensemble ANN based model that uses a Knn approach to identify precipitation forecast inputs. The implemented techniques consisted of CA, LI, and NAs processed as an ensemble mean as well as 125 individual members. Model output consisted of four weekly stream-flow forecasts; MRB, Cypress Creek, Trout Creek and Alafia. Performance was based on two error measures, RMSE and MBE, and the efficiency measure Ceff. Subsequently, the resolution and reliability of each technique was assessed using a tercile based reliability diagram and ROC plot.

All of the results produced from using these techniques, original and new, indicated that this model does not produce results with any quality for the two low flow-rate streams, Trout Creek (0.42mgd mean) and Cypress Creek (0.67mgd mean). In only one instance do forecasts for these creeks obtain a Ceff above 0 (0.24). The model does, however, show skill in forecasting the flow rates for the larger tributaries MRB (60.26mgd mean) and the Alafia (51.74mgd mean). MRB results indicated that while all of the new techniques outperformed the original KNN method, using precipitation forecasts produced from LI and CA improved stream-flow forecasts the most. LI showed the least amount of bias error of any of the methods. This is interesting because it is solely based on a spatial interpolation of the Numerical Weather Prediction Model (NWPM) and is the only technique that does not use

observed data in any aspect of its production. Therefore it is the least bias corrected, yet it shows the least bias.

Conversely the NA method using the mean of the 125 member analog ensemble had bias almost twice that of the other techniques. Initially it was considered that of the 125 analogs in the ensemble the poorest matches, with the least similarity to current forecasted patterns, and skews the mean away from quality analogs. Therefore, the median was tested but to no avail. The median of the NA ensemble performed worse than the mean. It would appear, based on the superior results of using the NA ensemble members individually, that the benefit of using an ensemble of NA forecasts propagates through a receiving model, only when the ensemble remains intact. However, reducing the ensemble using more complicated techniques, i.e. CAs, is effective and in fact improved stream-flow forecast results.

The resolution of the model when using the new techniques outperformed the original Knn method for all three terciles. For low flow (0-33%), the NA method using individual members had the best resolution followed by CA and LI. In high flow conditions, although all of the output resolution decreased, LI outperformed the other techniques by a good margin. Again this is an unexpected result. As for reliability, all of the techniques performed similarly with no one technique distancing itself from the others.

Although all of the aforementioned results were identified, Table 3-3 shows that based on the RMSE, MBE, and Ceff, the differences were marginal. Therefore using the Morris Method (Chapter 2) a simple input screening was performed to see if precipitation is even a factor in the model. The idea was to vary all of the stream-flow

inputs, all of the groundwater flow inputs, and all of the precipitation inputs as a group. The range of variation was based on 25% of the historical distributions. When the results came back, stream flow ranked highest for all 80 weeks of output, then groundwater and consistently lowest was precipitation. Now due to the crude nature of the test no information could be validated, however these results instigated a reanalysis of the code.

During the reanalysis there is a part of the code that is determining the Knn inputs that says, if either date for the two best matches is at the end of the data set, then use today's data as an input. The reason for this is that next week's forecast is determined from finding the Knn dates and taking the information from the following week. This is based on the theory that if today's pattern matches a historical pattern then the climatic conditions will propagate forward in the same manner. It turns out that the code written to perform this step always answers yes and therefore always uses the current values as a forecast.

The implementation of the new techniques is coded after this so they are applied, however when the ANN was trained and validated, the network was looking for relationships based on today's values predicting tomorrows. This could explain two things; LI which is a spatial interpolation of today's forecast performed so well and potentially why a more significant improvement was not seen using NA and CA.

Table 3-1. Descriptive statistics for observed stream-flow rates used in the study.

Statistic	Morris Bridge	Cypress Crk.	Trout Crk.	Alafia
mean	60.26	0.67	0.42	51.74
25%	36.91	0.08	0.08	32.25
75%	120.71	12.00	4.50	132.26
5%	23.95	0.02	0.00	19.69
95%	354.38	32.21	29.72	300.08
95% CI	330.43	32.19	29.72	280.39

Note: All values are given as weekly averaged daily flow rate (mgd)

Table 3-2. Temporal and spatial details of rain-gauge sites used by the SWANN model

Starting Date	Ending Date	Site Name	Latitude	Longitude
10/1/1988	Present	CNR-T1	28.120	82.131
10/1/1988	Present	CNR-T2	28.086	82.077
10/1/1988	Present	CNR--T3	28.148	82.093
12/7/1990	Present	CYB-CY-7	28.176	82.354
2/21/1986	Present	CYB-TOT	28.224	82.365
9/1/1976	Present	CYC-C-3	28.301	82.380
11/1/1985	Present	CYC-PLANT	28.286	82.425
2/25/1999	Present	SCH-SC-1	27.863	82.202
1/13/1989	Present	SCH-SC-4	27.863	82.144

Table 3-3. Summary of error measures for MRB and CYC stream-flow rates based on five different precipitation forecast methods.

Input Forecast	Morris Road Bridge			Cypress Creek		
	RMSE	CEFF	MBE	RMSE	CEFF	MBE
Knn Original	45.29 (44.79)*	0.46 (0.43)	4.94 (-2.73)	11.75 (11.72)	-0.06 (-0.19)	3.16 (3.02)
Linear Interpolation	42.38 (42.81)	0.51 (0.53)	3.34 (2.09)	11.75 (11.61)	-0.06 (-0,16)	2.91 (2.77)
Natural Analogs (mean)	44.65 (44.09)	0.47 (0.49)	8.16 (6.81)	12.11 (11.97)	-0.13 (-0.24)	3.34 (3.19)
Natural Analogs (125)	43.87 (43.2)	0.5 (0.51)	5.33 (4.05)	12.78 (13.66)	-0.06 (-0.04)	3.26 (3.02)
Constructed Analogs	42.63 (42.04)	0.52(0.54)	4.72 (3.43)	11.66 (11.53)	-0.04 (-0.15)	2.86 (2.71)

* Mean (Weighted Average) of ANN network ensemble stream-flow forecasts

Table 3-4. Summary of error measures for TRC and ALA stream-flow rates based on five different precipitation forecast methods.

Input Forecast	Trout Creek			Alafia		
	RMSE	CEFF	MBE	RMSE	CEFF	MBE
Knn Original	8.88 (8.75)*	0.24 (0.26)	1.13 (0.88)	88.55 (88.14)	0.35 (0.36)	-2.81 (-6.95)
Linear Interpolation	10.29 (10.14)	-0.05 (-0.02)	3.43 (3.11)	88.08 (88.64)	0.36 (0.36)	-2.17 (-6.41)
Natural Analogs (mean)	10.02 (9.98)	0 (-0.03)	3.47 (3.14)	87.94 (87.84)	0.36 (0.36)	-6.22 (-10.28)
Natural Analogs (125)	10.24 (11.42)	-0.01 (-0.01)	2.76 (2.35)	87.55 (87.36)	0.37 (0.37)	-3.75 (-8.06)
Constructed Analogs	10.09 (9.94)	-0.01 (0.02)	3.47 (3.14)	87.31 (87.02)	0.38 (0.39)	-3.25 (-7.43)

* Mean (Weighted Average) of ANN network ensemble stream-flow forecasts

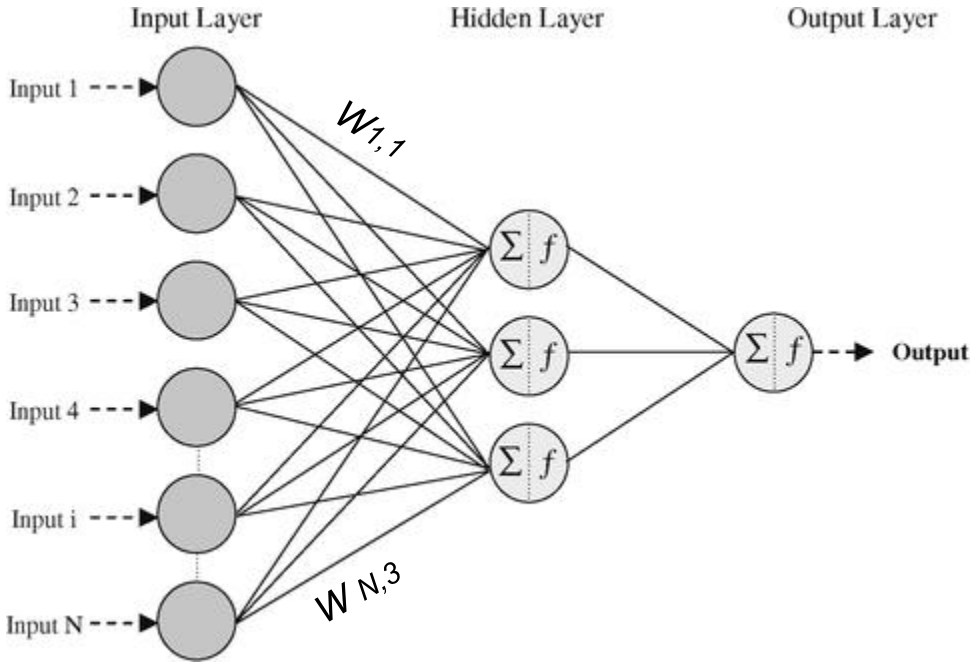


Figure 3-1. Example of an artificial neural network framework with output, input and hidden nodes.

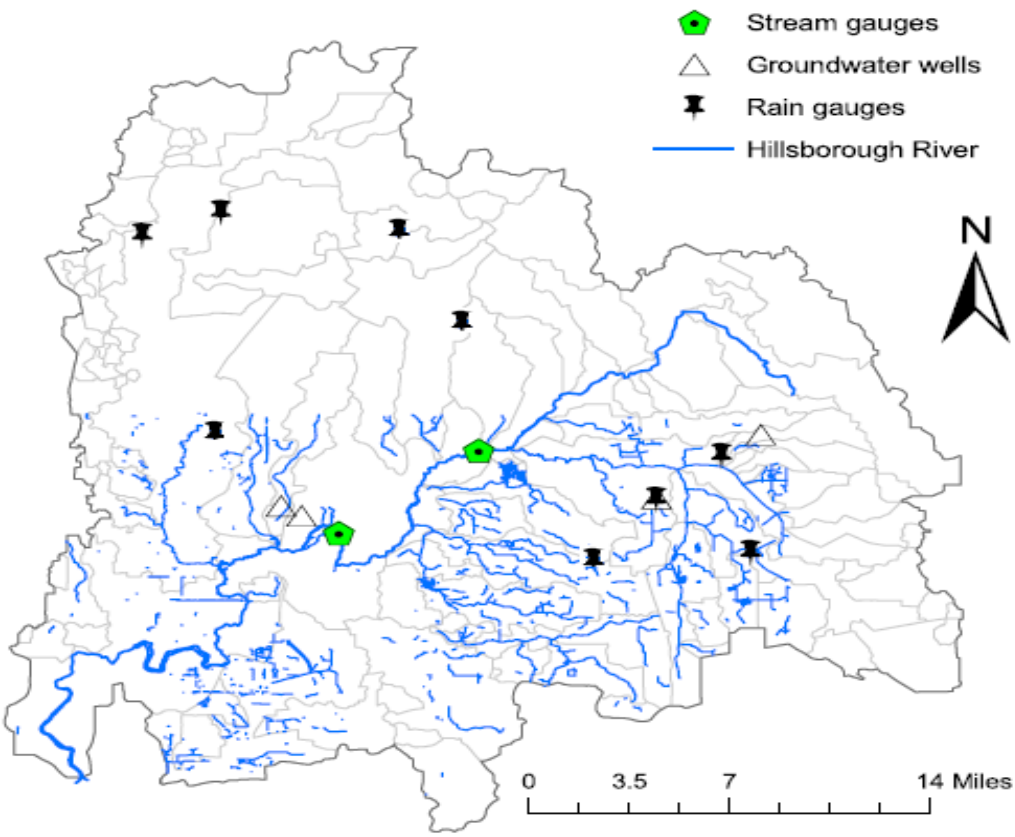


Figure 3-2. Locations of the 9 rain-gauges, 2 stream gauges and 4 groundwater monitoring sites used by SWANN.

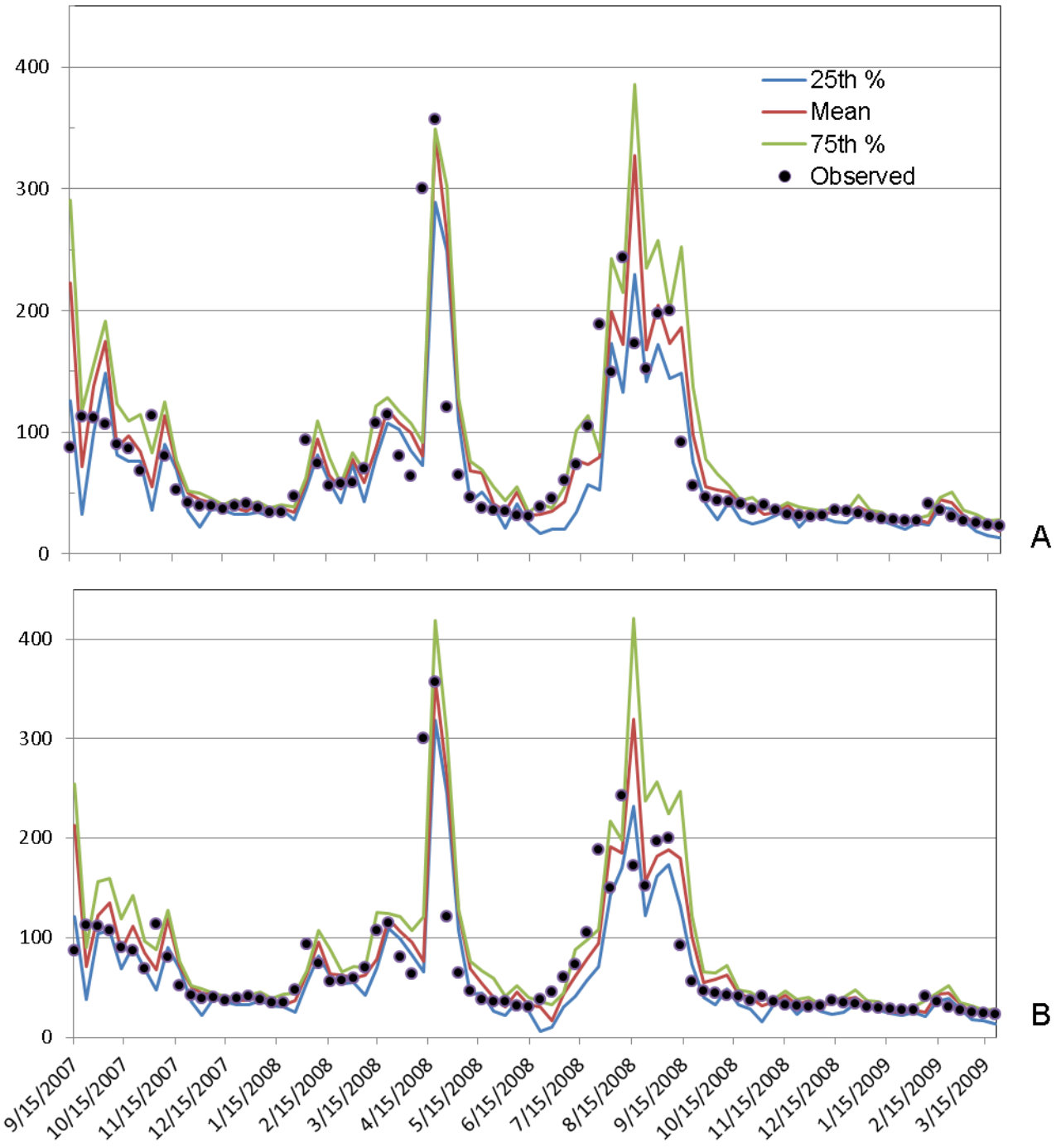


Figure 3-3. Stream-flow forecast results compared to observed data for Morris Road Bridge. A) Persistence forecast. B). Constructed analogs forecasts.

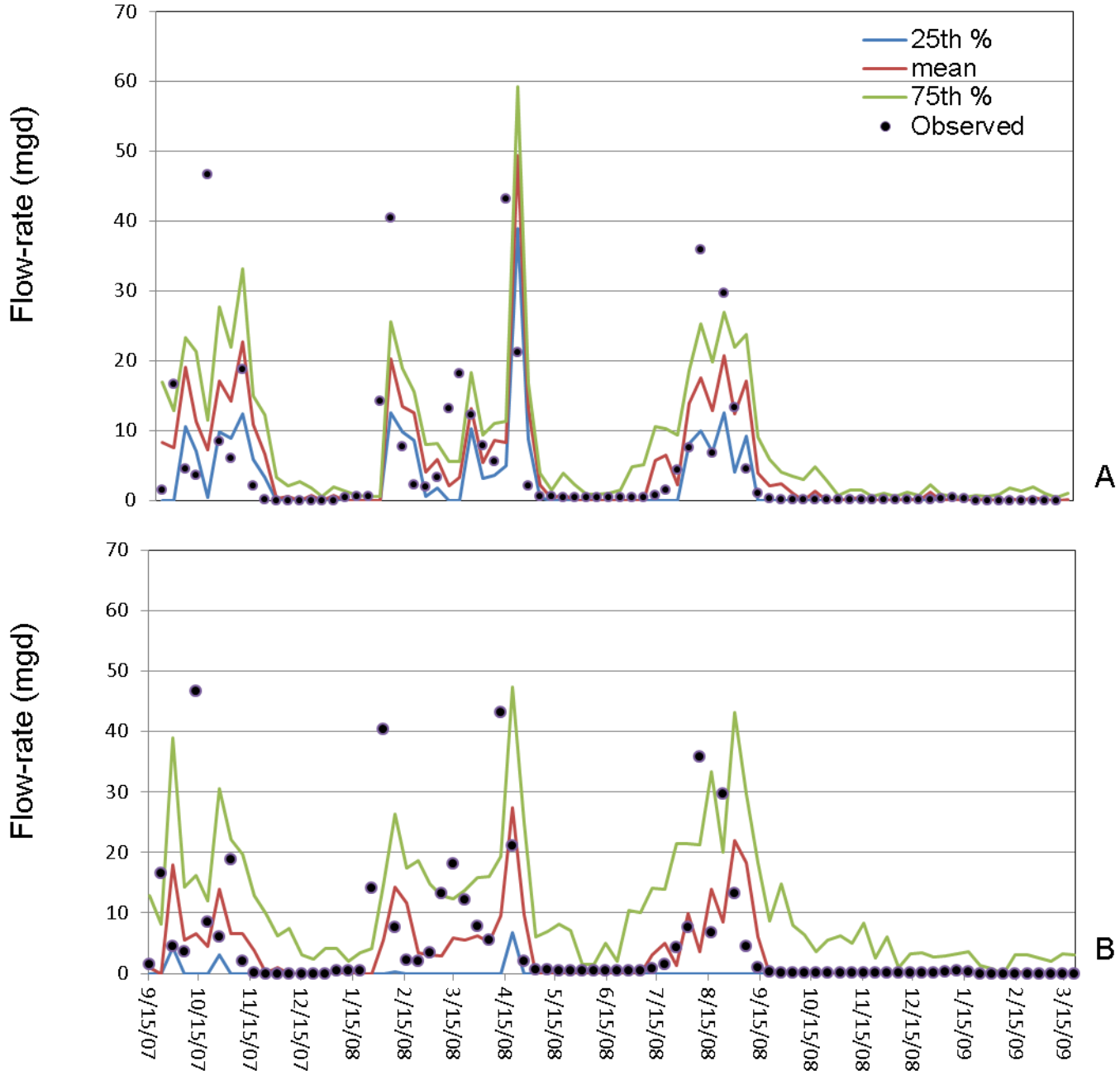


Figure 3-4. Stream-flow forecast results compared to observed data Trout Creek. A) Persistence forecast. B). Constructed analogs forecasts.

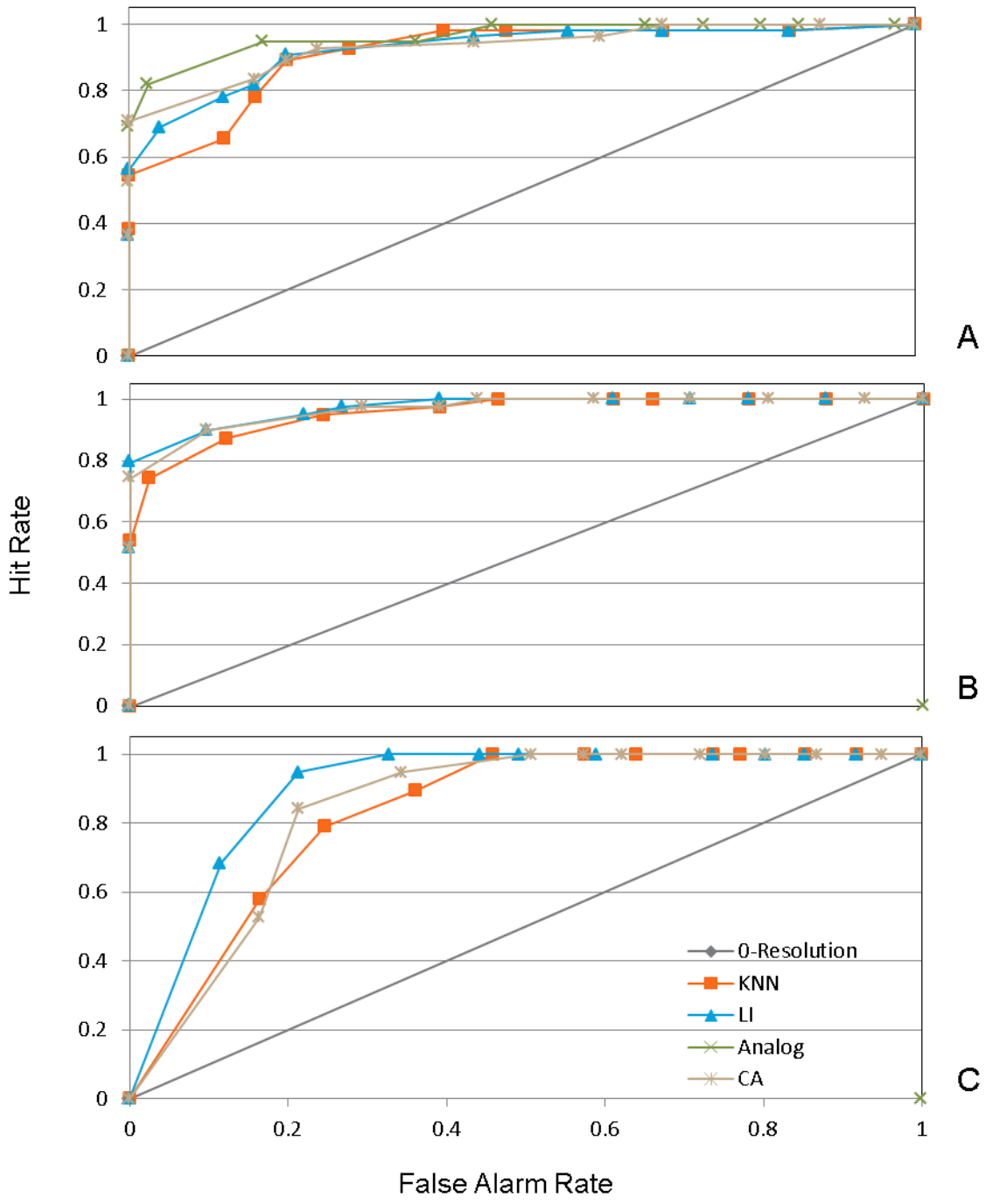


Figure 3-5. Relative operating characteristic plots for Morris Road Bridge stream-flow forecasts. A) Bottom tercile. B) Middle tercile. C) Upper tercile of observed stream-flow.

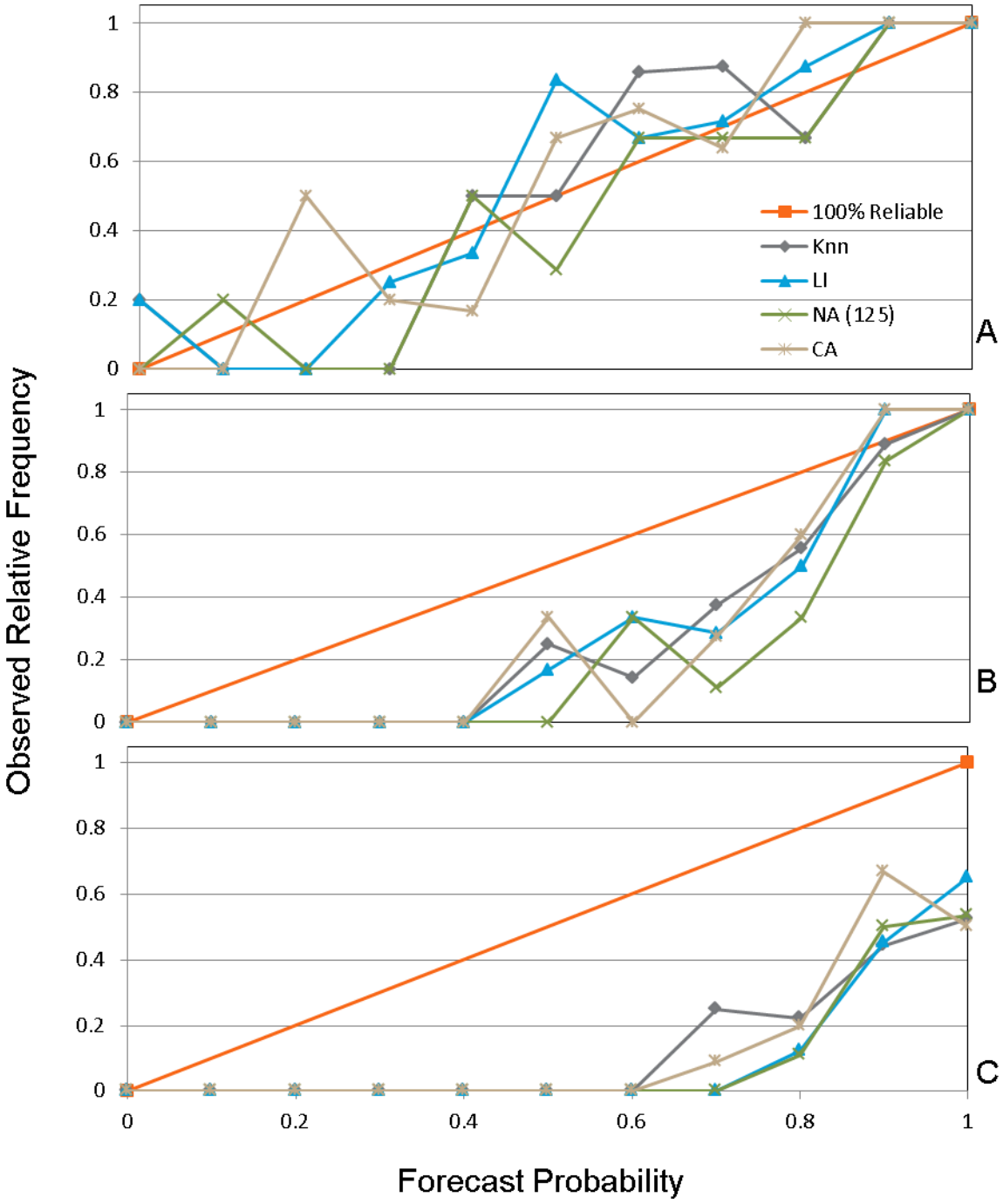


Figure 3-6. Reliability diagrams for Morris Road Bridge stream-flow forecasts. A) Bottom tercile. B) Middle tercile. C) Upper tercile of observed stream-flow.

CHAPTER 4
EXPLORING THE VIABILITY OF USING MULTIPLE PREDICTORS FOR ANALOG SELECTION USING A GLOBAL SENSITIVITY SCREENING METHOD

Background

A qualitative input sensitivity screening is performed on the Analog Method (AM) to assess the viability of expanding the direct analog approach, discussed in the previous study (Chapter 2), to analog searches using multiple predictor fields.

Precipitation forecasts at a local or regional scale are vitally important in their application to hydrologic models that are used in a wide number of decision making processes including agriculture, water management, and flood control. However, the state of the art in Numerical Weather Prediction Simulations (NWPMs) are based on gridded general circulation models (GCMs) on coarse scale that cannot accurately represent local scale circulation patterns or small scale climate variable spatial variability (Giorgi et al., 2001; Jones et al., 2004). In order to successfully use NWPM forecast information on these local areas, it is necessary for the coarse scale output to be downscaled and bias corrected.

While dynamic downscaling has shown ability to capture local scale atmospheric physics and climate variable spatial variability, the limitations of computational expense, lack of understanding of physical phenomena and propagation of NWPM bias has kept statistical downscaling forecast skill on an equal level. The AM is an atmospheric pattern matching approach that has shown, that with a limited number of degrees of freedom, can produce precipitation forecast skill on a spatial point (rain-gauge) basis. The use of the AM to downscale and bias corrects NWPM forecasts is a two-step process; 1) identify similar atmospheric patterns (analog), compared to a current forecast in a historic archive of forecasts, and 2) obtain the local observed precipitation

on the dates of the analogs and use the ensemble as a probabilistic or deterministic forecast. The identification of similar suitable analogs in the historic archive requires a base field of predictor(s) as well as a similarity measure.

The previous study (Chapter 2) has shown that using the AM with a direct analog approach, precipitation as a predictor and predictand, resulted in forecast skill for the Tampa Bay region of Florida with seasonal variation. Those results posited the question; if the predictor field was increased to include multiple predictors; what would be the effect on the temporally based AM output? One approach to answering this question would be to perform a Global Sensitivity Analysis (GSA) on the method.

The selection of the appropriate GSA method to use on a model is predicated on the goal of the assessment. Applicable goals consist of factor fixing, factor prioritization, variance cutting and factor mapping (Cacuci et al., 2003; Reusser et al., 2011; Saltelli et al., 2006). In the case of this study GSA is used to identify the importance of input factors based on the effect that their uncertainty impacts output variability. For this approach a quantitative SA technique would be preferred in that the resulting prioritization could be assessed based on the magnitude of the sensitivity index. However this form of GSA is extremely computationally expensive. As an example, for this study using 8 input factors, the Sobol method (1993) requires a minimum of 4608 simulations. The AM tested in this study requires 8 to 15 hours per simulation, therefore the cost is prohibitive.

Prior to performing a computationally expensive Monte-Carlo based SA a systematic screening sensitivity analysis methodology proposed by Morris (1991) is used to qualitatively identify important input factors based on their relative impact on

output variability. The method of Morris uses the mean and standard deviation of a sensitivity index, known as an elementary effect, to identify the relative input factor importance and potential interactive effects with low computational cost. This information can then be used to assess whether a factor is important or negligible. If negligible the factor can then be removed from more computationally expensive quantitative SA techniques. A secondary goal of this study is to evaluate if there are variations in sensitivity based on seasonal variation, lead-day propagation, and the magnitude of the precipitation forecast event threshold.

Objectives

The objective of this work is to qualitatively assess the use of weighted multiple predictors for analog selection with regard to input factor effects on output variability. The hypothesis is that not all of the six weighting predictors used will have an effect on output variability and can subsequently be discarded. A secondary hypothesis is that there are variations in the importance of input factors that can be attributed to lead-day variation, event magnitude and seasonality.

Study Area and Data Collection

The contributing watershed to Tampa Bay Waters area of responsibility is 6,583 km^2 and intersects 3 counties; Hillsborough, Pasco, and Pinellas. This study was performed using a 9 point GFS reforecast grid ranging from a latitude of 25-30^{0}N and a longitude of 80-85^{0}W (Figure 2-1) with a spatial resolution of 2.5^0 or approximately 200 km.

The AM requires the use of two different sources of data; simulated gridded NWPM forecast data and observed rain-gauge data, both spanning the same temporal range. The NWPM data is used in the analog selection process and the rain-gauge

data for subsequent statistical analysis of corresponding observational data. The forecast data archive used during the selection process is the reforecast data set produced by the National Center for Environmental Protection using a fixed 1998 operational version of the Global Forecasting System (GFS) model. The archive has a temporal range of over 30 years, 1/1/1979 – present, and a spatial resolution of 2.5 degrees or approximately 200km. A 15 member ensemble forecast was produced every day from January 1979 to present resulting in a 30 year data set. The ensemble initial conditions consist of a control initialized with NCEP-NCAR reanalysis and a set of seven bred pairs. The breeding method was consistent with the 1998 operational model. The forecast extends to 15 days and was archived every 12 hours (Hamill et al., 2006b). The reforecast archive can be downloaded at: http://www.esrl.noaa.gov/psd/forecasts/reforecast/data.html.

Three rain-gauges from the Tampa Bay Water watershed (St Leo, CYCC3, and Tarpon rain-gauges), with the most complete data over the temporal range of the reforecast archive, were used in the study.

Methodology

The sensitivity analysis method used for AM input parameter screening is the method of Morris that follows a four step procedure: 1) select input parameters that are to be investigated and determine their appropriate pdf and range; 2) generate simulation input values from the defined input space using the proposed methodology by Morris (1991), and three optimized schemes proposed by Campolongo (2007), Ruano (2012) and Khare (pending); 3) simulate AM outputs using inputs created in step two 4) produce relative sensitivity index rankings according to the modified method of Morris.

AM Model Inputs and Outputs for Screening

The principle components of the AM are analog selection and statistical analysis of corresponding observed precipitation data. Analog identification is accomplished by comparing the current predictor field to potential analog predictor fields in the reforecast archive. Comparisons are made based on a gridded similarity measure. The method used in this study is a static version of the methodology described in Hamill and Whitaker (2006a). Instead of using a moving spatial window, a single coarse scale window of 9 grid-points surrounding the Tampa Bay area is used. The analogs are initially selected by comparing the current days coarse scale precipitation forecast (Z_{GFS}) and comparing it to an archive of historical forecasts within +/- 30 days (as a starting point) for all years in the archive (maintaining seasonality), while omitting the current year. The omission of the current year in the identification of potential analogs is used to cross validate the model. The similarity measure used is the RMSE calculated at each of the nine grid-points. The 125 coarse scale patterns ($Z_{analogs}$) with the lowest RMSE are then selected and the corresponding observed precipitation from individual rain-gauges is then used to create the forecast ensemble ($P_{analogs}$). The vector, $P_{analogs,}$ can then be evaluated as a deterministic or probabilistic forecast.

The AM being tested is an expansion of that assessed and optimized in Chapter 2. Rather than using direct analogs, precipitation as the predictor and as the predictand, this section of the study examines the value of using multiple predictors to identify historical analogs. The predictors consist of precipitation (APCP), precipitable water (PWAT), relative humidity (RHUM), mean sea level pressure (PRMSL), and two 10 meter wind vectors (U10m and V10m).

Implementing any kind of sensitivity analysis requires a determination of what factors should be tested and what are their potential distributions. In most hydrologic models, the distribution of input factors is based on performing a literature search confined to the temporal and spatial area represented. In the case of the AM a different approach is necessary. The first factor, search window size, is bounded at the high and low ends due to the desire to adhere to seasonal climatic physics and the need for a sufficient analog pool to find a good match. The range chosen was based on a temporal range of 1 month to 6 months or windows of +/- 15 days to +/- 90 days. It was also a discrete uniform distribution due to the necessity of using whole days as an input for data manipulation and model performance. The second input factor, number of analogs used to create the forecast ensemble, is based on previous literature values proposed by Hammil (2006) and refined by the optimization study reported in Chapter 2. The results of Hammil's study recommended the use of 75 analogs throughout the Continental United States. With this number as a starting point a crude one-at-a-time (OAT) evaluation of the number of analogs used was performed resulting in an optimum of 125 analogs. This was identified by an asymptotic plateau of the lead-day 1 BSS and a slight decline in the BSS for longer lead days with an increased number of analogs. The range chosen for this screening evaluation is 15 to 150 analogs and a discrete uniform PDF.

The final 6 input factors are analog predictor weighting coefficients that are used in the identification of the best analogs using RMSE as a similarity measure. Recall that we are using 6 potential predictors provided by the GFS reforecast data set, i.e. APCP, PWAT etc. All of these predictors consist of different units and different ranges of

numerical values. The process of determining likely analogs using multiple predictors consisted of several steps. Initially each predictor is treated separately and the 9 point gridded RMSE is calculated in the same way that it was performed in the direct analog study in Chapter 2. The second step is the normalization of each respective RMSE. This is done simply by producing the z statistic (Equation 4-1) where:

$$SM_i^{std} = \frac{SM_i - \mu_{SM}}{\sigma_{SM}} \quad (4\text{-}1)$$

SM_i^{std} is the standardized value of a single analog similarity measure (using gridded RMSE) calculation for a single predictor, SM_i is the value to be standardized, and μ and σ are the mean and standard deviation of all of the analog SM values for a single predictor. The SM, μ and σ all have the same units therefore the resulting z statistic is dimensionless. The standardization process has the effect of transforming the distribution of the single predictor SM population to a distribution with a mean of zero and a standard deviation of 1.

After the individual predictor SMs were standardized a weighted equation (Equation 4-2) was developed to determine a total multi-predictor similarity measure;

$$SM_{tot} = C1 * SM_{pwat}^{std} + C2 * SM_{rhum}^{std} + \ldots \ldots + C6 * SM_{v10m}^{std} \quad (4\text{-}2)$$

where SM_{tot} is the weighted multi-predictor root mean square error, and C1 through C6 are weighting coefficients corresponding to each individual predictor 9 grid point analog RMSE calculation. The vector SM_{tot} is then sorted with the lowest value indicating the best or closest analog match. The inputs used for sensitivity screening analysis are the weighting factors. The weighting factors (C1 through C6) are given a uniform

distribution from 0 to 1. In this way each sensitivity screening will be testing output variability based on the strength of the predictor in the similarity measure.

The output that is tested for variability consist of several skill measures described in detail in Chapter 2, The Brier Skill Score is analyzed over 6 precipitation thresholds due to its ability to evaluate relative forecast quality with respect to climatology in regard to the size of a precipitation event. The thresholds assessed were 1, 2.5, 5, 10, 25 and 50 mm of rain in a 24 hour period. The second skill measure assessed was the Cumulative Rank Probability Skill Score. This measure shows comparative skill across the entire distribution of potential precipitation events. The screening procedure was also analyzed for 1, 3, and 5 lead days. For purposes of clarity the daily results were grouped into monthly averages in order to assess if there is a difference in input factor ranking based on seasonal variation.

Morris Method

The guiding philosophy of the method of Morris is to determine which factors may be considered to have effects on output variability which are (a) negligible, (b) linear and additive, or (c) nonlinear or interactive (Morris, 1991). While using OAT methods are limited in that the information obtained cannot be extrapolated outside of the specific spatial domain of the calculation, Morris proposed an adaptation of the OAT method that would retain the benefit of limited computational expense while calculating a sensitivity measure over the global input domain. The sampling strategy proposed is the creation of multiple (r) trajectories in input space, with a recommendation of 10 to 50 trajectories being sufficient for screening analysis. Each trajectory has a random starting point and proceeds to the end point by randomly moving one factor at a time. Therefore each step in the trajectory varies by a single factor. In this way each input

factor can be evaluated individually once per trajectory using the change in model output corresponding to the variation in the input factor. Morris proposed a sensitivity index (Equation 4-3) for these individual calculations called the elementary effect (EE).

$$EE_i(X) = \frac{[y(x_i,\cdots\cdots x_{i-1},x_i+\Delta,x_{i+1},\cdots\cdots,x_k)-y(X)]}{\Delta} \quad (4\text{-}3)$$

EE_i (X) is the elementary effect for a given factor i, Δ defines the jump in input factor X_i between two levels (defined below) and is a value in {0,1/(p-1),…,1-1/(p-1),1} (Lins et al., 1997), and p is the number of levels or discrete intervals in input space.

While EEs can be considered an OAT or local sensitivity analysis approach Morris suggested that analyzing the EEs of a single factor across all trajectories using a statistical measure would give the method the ability for global assessment. The sensitivity measure used is the mean of the EEs or μ, and once calculated can be used in relation to the other inputs as a relative importance factor and can be seen in Equation 4-4 (Morris 1991).

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i(X^j) \quad (4\text{-}4)$$

In Equation 3-2 μ_i is the mean of the elementary effects for factor i, r is the number of trajectories, and $EE_i(X^j)$ is the elementary effect for input factor i trajectory j. The input factor that has the highest value would correspondingly have the greatest “relative” effect on output variation. An arbitrary value was chosen whereby Input factors with μ values less than 5% of the highest value are analyzed for a no effect consideration. If it is determined, based on future computational expense and Morris results, that the factor has no effect it can then be eliminated from a subsequent, and more computationally expensive quantitative sensitivity analysis. It does not necessarily mean that the input

can be removed from the model but it does suggest that the value of the input can be fixed with limited regard to input uncertainty propagating to the output.

A modification was made to the Morris method by Campolongo (2005; Campolongo et al., 1999; Saltelli et al., 2000) in that the mean of the EEs is calculated using the absolute values. The thought was that if you had positive and negative EEs (due to non-monotonic behavior) then there is potential that μ could artificially shift towards zero, while using the absolute values of the EEs in the calculation of μ would take into account output variation in either the positive or negative direction. This measure is denoted as μ^* (Equation 4-5).

$$\mu_i * = \frac{1}{r}\sum_{j=1}^{r} |EE_i(X^j)| \quad (4\text{-}5)$$

A secondary EE statistic (Equation 4-6) that is used for screening purposes is the standard deviation, σ, of the EEs

$$\sigma_i = \sqrt{\frac{1}{(r-1)}\sum_{j=1}^{r} EE_i(X^j)} \quad (4\text{-}6)$$

In effect the standard deviation is a measure of the spread of the EE distribution for a given input factor. A large standard deviation relative to μ^* indicates there is a significant interactive effect with other parameters or that the effect is non-linear. Conceptually if the input factor has a linear effect on output variation and Δ is the same for all trajectory calculations then the EE from any trajectory would be the same. However the static input factors that correspond to the EE calculation are different for each trajectory, therefore if there is a large difference in the μ_i^* values for each trajectory then it is possible there is an interactive or second order effect at work or that there is nonlinear behavior. Conversely a small standard deviation indicates little to no

interactive effects on the output. It is worth noting that while μ* is a preferential evaluation of the mean of the EEs, it is incumbent on the modeler to show the relationship of σ with regard to μ. It has been suggested and widely reported that the use of μ* and σ is efficient in describing the relative importance and potential interactive effects of input factors (Campolongo et al., 1999), but this writer disagrees. If σ is virtually zero then there is no issue, however assessing the magnitude of σ relative to μ* is non-informative and somewhat misleading. While σ is calculated directly from the distribution of EEs, μ* is calculated from a transformed distribution of EEs, the absolute value. Therefore the standard μ* vs σ plot commonly used should plot both μ and μ* vs. σ. An example of this plot can be seen in Figure 4-1.

One of the reasons for performing a screening method is in the ability to obtain the relative importance of an input parameter in regards to its effect on output variability using a limited amount of model realizations. The number of simulations required to perform the method of Morris, (N) is calculated according to Equation 4-7.

$$N = r(k + 1) \quad (4\text{-}7)$$

It has been shown good results have been achieved with r = 10 although Morris suggested 10-50 trajectories (Campolongo et al., 1999; Saltelli et al., 2000). For the case of this study using 8 inputs, 90 simulations was required for successful analysis.

Input Trajectory Generation

The original trajectory generation methodology proposed by Morris consists of five steps;

- Discretize the multidimensional input space (unit hypercube) into an equidistant gridded region with p levels.
- Randomly select a starting node at any location within the input space.

- Increment a randomly selected input factor positive or negative Δ where Δ is a predetermined multiple of 1/(p-1)
- Repeat for the remaining factors, in a random order until all factors have been varied.
- Repeat the entire process until the required number of trajectories has been produced.

Once the trajectories are determined the unit hypercube or probability based values are mapped to the pdfs of the model input factors being evaluated. Figure 4-2 shows the Simlab sampling strategy based on p=4 and Δ = 2/(p-1)=2/3. It is worth noting that the input space is truncated in order to map to model pdfs without dealing with singularities at the extremes of 1 and 0 as is the case with several potential distributions.

There is concern that the random methodology used to create the trajectories could lead to non-optimal coverage of the input space especially for models with a large number of input factors (Campolongo et al., 2007). Several methods for improved sample generation have been proposed with regard to trajectory spread, speed of implementation and similitude to the uniform distribution including the Campolongo (2007), Ruano (2012), and Khare (pending) methods and this study will perform a cursory examination of the effectiveness of these methods.

Campolongo Method

The Campolongo (2007) method of trajectory generation is an oversampling method where the optimum trajectories are chosen based on a pairwise brute-force distance calculation between all trajectories. There are three steps to this procedure: 1) create M trajectories (r*10 to r*100) using the original method of Morris, 2) determine the pairwise distance between every possible pair of trajectories, and 3) calculate the

sum of the pairwise trajectory distances for all possible r combinations. Once this is complete the combination of r trajectories with the maximum distance is selected. This measure would account for the greatest spread in input space. The pairwise trajectory distance calculation can be seen in Equation 4-8.

$$d_{ml} = \begin{cases} \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \sqrt{\sum_{z=1}^{k} \left[x_i^m(z) - x_j^l(z) \right]^2} & for\ m \neq l \\ 0 & Otherwise \end{cases} \quad (4\text{-}8)$$

where; d_{ml} is the pairwise distance between trajectory m and l, k is the number of input factors, $x_i^m(z)$ is the z^{th} coordinate of the i^{th} point of the m^{th} trajectory. This can be thought of as the sum of the geometric distances between all of the pairs of points between two trajectories.

Once these values have been calculated in an [M,M] triangular matrix, the r optimum combination of trajectories are calculated according to Equation 4-10 (for r=4 and M=10).

$$D_{4,6,7,9} = \sqrt{d_{4,6}^2 + d_{4,7}^2 + d_{4,9}^2 + d_{6,7}^2 + d_{6,9}^2 + d_{7,9}^2} \quad (4\text{-}10)$$

$D_{4,6,7,9}$ is the sum of the distances for the trajectory combination consisting of the fourth, sixth, seventh and ninth trajectories. Upon completion of distance measures for all possible combinations of r=4 trajectories, the combination with the greatest value of D is chosen.

While the brute force methodology is effective in locating well disbursed input trajectories, the combinatorial approach shows a nonlinear relationship in the number of oversampled trajectories and the amount of computational time required. Ruano (2012)

proposed a method that potentially reduces the computational time while maintaining adequate spread in input space.

Ruano Method

Ruano's enhanced method follows the Campolongo method in that M oversampled trajectories are created as well as the calculation of a pairwise trajectory distance matrix. From this point the optimum trajectories are selected in an iterative fashion initially selecting the trajectory pair with the highest separation (D value) and subsequently through distance calculations selecting the third trajectory. This process is repeated until r optimum trajectories are identified. A detailed description of this process can be viewed in Ruano (2012). In this method it can be seen that the calculation of the optimum combinatorial distance does not require the calculation of every possible combination available thereby saving computational expense.

Khare Method

The Khare method does not follow the same process in trajectory optimization. The theory behind this method is that the uniform distribution of the input samples should be maintained in the input selection process. In this way there is no need to oversample. The process begins with the random generation of trajectory starting and ending points with the stipulation that each factors population at a given level remains uniform. The path of each trajectory is created by randomly changing one factor at a time again while maintaining uniform populations within each level or bin. While this procedure is computationally inexpensive, the process is simulated numerous times in order to compare the optimized combination of trajectories distance measure. The combination produced that has the greatest distance is chosen as the optimum r trajectories.

Results and Discussion

AM Method of Morris Screening

There are difficulties in analyzing the results of a sensitivity screening analysis that uses multiple outputs. Individual screening results are produced whereby each input factor is relatively ranked based on individual output variation. A question arises in which results, if contradictory should take precedence. The answer to that question is to simulate and interpret the results based on the information that the user requires. In the case of this study there are 7 forecast skill outputs used for each of three lead-days. There is also the issue with temporal analysis in that there are 12 monthly outputs for each of the measures previously listed. A graphical representation of the outputs available for analysis can be seen in Figure 4-3. For this study there are 218 potential screening results that require assessment therefore it is necessary to perform some form of grouping.

The primary hypothesis of this study is that input factors related to multiple predictors will have an effect on input variability indicating that multiple predictors can potentially produce greater skill in local precipitation forecasts. There are however secondary considerations. In the case of this study there is a question as to differences in input factor rankings based on seasonal variation, the size of the precipitation event, and the forecast lead-day. An example of sensitivity analysis results for lead day 1 using the CRPSS as output can be seen in Figure 4-1. The y axis shows the value of σ and the x axis shows the value of μ and μ^*. The low values of σ indicate minimal interactive effects for all three lead days. From this graphic it can be seen that input factor sensitivity index ranking for 218 outputs can become significantly unwieldy. In order to present this information in a comprehensible format, individual sensitivity

screening results were grouped and each input factors rankings (1 has the highest effect and 8 has the lowest) are collected into an ensemble. Box plots are then created showing the mean, median, 25th and 75th percentiles of the factor input rankings. In this way not only can the mean be used to show overall relative importance, but comparisons can be made as to the spread of the rankings in the ensemble for example between seasonal variation.

Grouping of the AM outputs is based on temporal (monthly), lead-day, and skill score. Figure 4-4 shows the relative ranking ensembles of all months and all skill scores combined for lead day 1 and lead day 5. It can be seen that the coefficient for precipitation (C1) as well as the number of analogs (NA) and search window size (SW) have the most significant effect on output variation. These input factors coincide with the direct analog approach, precipitation as predictor and predictand, analyzed in Chapter 2. The other 5 analog predictor coefficients have significantly less impact with C3, corresponding to mean sea level pressure consistently had the lowest ranking. , For lead-day 1, predictors coefficients C2 and C4 through C6, are clustered together with C6 (vertical 10 meter wind vector) having the highest mean ranking of the 4. However, the spread of the C6 ranking ensemble overlaps the coefficients for PWAT, RHUM, and the U10m (C2, C4 and C5 respectively) showing that its effects fall in the same range.

The shift in the ranking ensembles from lead day 1 to lead day 5 indicate that NA increases in importance over C1 and SW and these three factors show an increase in dominance over the remaining factors. This can be seen by the separation of ensemble distributions. It makes sense that as you propagate through forecast lead-days the

number of analogs used would have an increased effect in that increased lead days lead to more uncertainty in the original forecasts produced therefore a potentially larger ensemble of analogs would aid in limiting this uncertainty. The other notable shift is that the C6 sensitivity index ranking ensemble mean loses its superior position over the middle cluster of predictand coefficients.

In order to assess the seasonal variation of forecast input factor sensitivity rankings all skill score screening results were grouped into two three month windows, Dec-Feb for winter and Jun-Aug for summer. Figure 4-5 graphically displays these rank order ensembles for lead day 1. The importance rankings for winter show the same three dominant input factors as the overall analysis (C1, SW and NmAn) however C1, the coefficient for precipitation has the highest ranking followed by NmaN and then SW. The coefficient for mean sea level pressure remains as the least important factor with the other four weighting coefficients clustering in the middle with C4, corresponding to RHUM, having the greatest value. It does make sense that the coefficient for APCP would have a more dominant effect in winter as this is the dry season in the study area and precipitation is due to frontal and not convective storms.

In the summer months there is a shift in both the dominant three inputs (C1, SW, NA) as well as the middle cluster. The number of analogs dominates the effect on variability and C1 as well as SW follow with equal footing. Of the cluster, the wind vector (U10m) coefficient replaces the RHUM input as the most dominant with C2 and C4 falling. Again the PRMSL weighting coefficient shows the least effect on model output variability.

The reason for the seasonal shift of NmAn ranking above C1 is based on the climate of the study area. In the summer it is the wet season with convective storms occurring on a regular basis. This form of precipitation event is the hardest to forecast, which can be seen by the reduced skill scores in Chapter 2. Due to the uncertainty in the climatic physics, a larger ensemble of analogs would mute the event variance. In other words using an ensemble on a signal with high variability mutes the signal. It would have the tendency to under-forecast large events and produce a drizzle effect on small events. As for the relative humidity loosing importance from winter to summer, the study area maintains high humidity throughout the summer with little variability; therefore it is potentially easier for a RHUM signal to propagate in the winter.

The third aspect of precipitation forecast model screening results that are of interest is if there is a shift in importance ranking order based on the size of the precipitation events. This assessment is accomplished by comparing screening result ensembles from BSS outputs from low threshold levels (1, 2.5 and 5mm) to that of high thresholds (10, 25 and 50mm). Recall that BSS is based on exceedance thresholds, or in other words the forecasted probability that the precipitation on a given date will be greater than a predefined threshold level. Figure 4-6 shows the rank order comparison of input factors based on lead-day 1 ensembles from all months and high and low BSS output grouping. Both the high and low threshold results follow the same trend as the previous analyses and have similar rank orders and distributions with SW, NmAn and C1 having similar ensemble means. The coefficient for APCP (C1) does show a slight increase in rank over that of SW and NmAn for the low threshold events with NmAn, while maintaining the same ensemble mean value, has a shift in distribution to lower

rank order. The results indicate that there is minimal change in input factor importance rankings due to precipitation threshold levels.

Review of screening results with respect to the elimination of input factors from a more extensive global sensitivity analysis showed that all of the factors should be included. Although the weighting coefficient for PRMSL consistently was ranked last as to its level of impact on output variation, in all but one individual result did the μ* value fall below 5% of the highest input factor ranking value. It can be argued that it should be removed due to its consistent proximity to the removal threshold, however due to the limited number of input factors being tested (8) the removal of just one factor would not make much difference in terms of computational expense.

Input Selection Methods Comparison

The comparison of the four input trajectory generation methods; Morris, Campalongo, Ruano and Khare are evaluated based on computational speed, ability to maintain uniform distributions of input factors in the unit hypercube, spread of the optimized trajectories (evaluated by distance measure D; Equation4-4), and potential differences in sensitivity importance rank order. The comparison of methods is based on 8 input factors (k), 4 levels (p), 10 optimized trajectories (r), and oversampled trajectories. Figure 4-7 and Table 4-2 compare the generated input distributions for the four sample generation schemes and their chi-square failure rate respectively. It can be seen that the Campolongo and method of Morris (implemented by Simlab) methods have the same chi square failure rate, however the seventh input factor produced from the original method of Morris has the greatest deviation from a uniform distribution of any factor (Figure 4-7). The Ruano method shows improvements in the failure rate while the Khare method shows no failures at all. The lack of failures in the Khare

method is expected as this is the only method of the four that is programmed to require uniform sample generation.

The relative spread (D) and the computational speed of executing the generation methods can be seen in Table 4-2. As expected, all three of the optimized sampling schemes show a greater spread in trajectories than the original method with the brute force Campolongo approach achieving the best results. However the Campolongo method also has the second highest computational cost. As this study is performed with a limited number of input factors, computational expense is not an issue, but if these schemes were used with a large number of factors this expense would be compounded. The Khare method time of execution is similar to Campolongo but this value may be misleading. In this case, for the Campolongo and Ruano methods M oversampled trajectories are used, to aide in comparative value the Khare method was executed M times and the most distributed set of trajectories were selected. This was an arbitrary selection. An optimization of the Khare method is underway whereby the number of simulations required should be reduced without reducing performance (Khare, pending).

The four methodologies were subsequently analyzed based on screening results of the AM. Figure 4-8 presents the lead-day 1input factor ensemble rank order boxplots for all four generation methods using all of the skill score outputs. It can be seen that all four methods show similar results with a slight increase in the importance of SW for the three optimized methods. While this shows that there is potential improvement in rank order based on optimized input sampling, further studies would be required for validation. All that can be taken from these results is proof of concept. The original

Morris method is widely used in research and the similarity of results indicates that the new strategies do not adversely affect the technique. Combine that with the intuitive notion that in an input space with unknown behavior, smaller gaps between tested locations improve the likelihood that localized variations will be missed.

Chapter Summary

The method of Morris was used as a screening sensitivity analysis method in order to assess the viability of expanding the direct AM from a single predictor search routine to a multi-predictor algorithm. The method was chosen for its low computational expense and its ability to identify relative input factor effects on output variance. This information can then be used to identify important as well as inconsequential input factors. Two sensitivity measures were used in this assessment; μ^* which is the mean of the absolute values of the elementary effects (importance) and σ which is the standard deviation (interactive effects).

By implementing the method of Morris on a multiple predictor AM, several implications came to light. It was determined that the three inputs of the single predictor method, number of analogs search window size and precipitation (weighting coefficient C1) showed the greatest impact on output variability. While C3, the weighting coefficient for mean sea level pressure consistently showed the least effect, it did not fall below a relative threshold that would indicate its removal from further sensitivity analysis. The remaining four inputs, weighting coefficients corresponding to PWAT, RHUM, U10m and V10m, are clustered in the middle of the rank order.

The study also showed that there is variation in the relative sensitivity based on season as well as forecast lead-day. Results indicated that in winter of the three dominant input factors the importance of C1 increases over that of SW and NmAn. This

could be due to winter being the dry season in the study area and precipitation is due to frontal and not convective more temporally chaotic storms. In the summer as well as at longer lead times the screening results indicate that the number of analogs used show the highest effect on output variation. This coincides with times of larger forecast model uncertainty and decreased skill in precipitation forecasts. The decrease in reforecast skill, in the case of lead time, can be attributed to the forecasts being further removed from the initial conditions. In the summer, model uncertainty is increased due to the chaotic variability of convective storms as well as the modeling limitations of cloud formation. With these limitations the value of the size of the ensemble of analogs becomes more important.

Comparisons of the input sample generation methods showed that the three enhanced methods outperformed the original method proposed by Morris. The Campolongo method had the highest trajectory spread in input space but had the second highest computational time required for execution. While the computational time for this study was not prohibitive, an increase in the number of input factors would show a nonlinear increase in execution time. In terms of reproducing a uniform distribution in the input samples the Khare method was the only one that had no failures in the Chi-Square test at any confidence level. The lack of failures is due to the method having uniform generation of samples as a constraint.

Table 4-1. Definition and characteristics of input factors used for screening SA.

Input Parameter	PDF (Range)	Range	Description
C1	Uniform	0-1 dimensionless	Coefficient for Precipitation (APCP)
C2	Uniform	0-1 dimensionless	Coefficient for Precipitable Water (PWAT)
C3	Uniform	0-1 dimensionless	Coefficient for Mean Sea Level Pressure (PRMSL)
C4	Uniform	0-1 dimensionless	Coefficient for Relative Humidity (RHUM)
C5	Uniform	0-1 dimensionless	Coefficient for 10m East-West Wind Vector (V10M)
C6	Uniform	0-1 dimensionless	Coefficient for 10m North-South Wind Vector (U10M)
SW	Discrete Uniform	30-180 Days	Search Window Size
NmAn	Discrete Uniform	15-150 Analogs)	Number of Analogs

Table 4-2. Chi-Square statistical test results for four method of Morris sampling schemes.

	No of Failures Based on Conf. Level				Distance	Time
	0.8	0.9	0.95	0.99	D	(s)
Campo	2	5	5	5	708.2	110.6
Ruano	1	3	3	3	680.6	14.2
Khare	0	0	0	0	609.9	135.7
Morris	2	5	5	5	549.0	0.257

Note: Comparison of four input generation methodologies based on Chi Square Test to assess uniform distribution, using 4 confidence levels, the trajectory spread evaluated by distance measure D. A higher value of D indicates a greater spread in the input space.

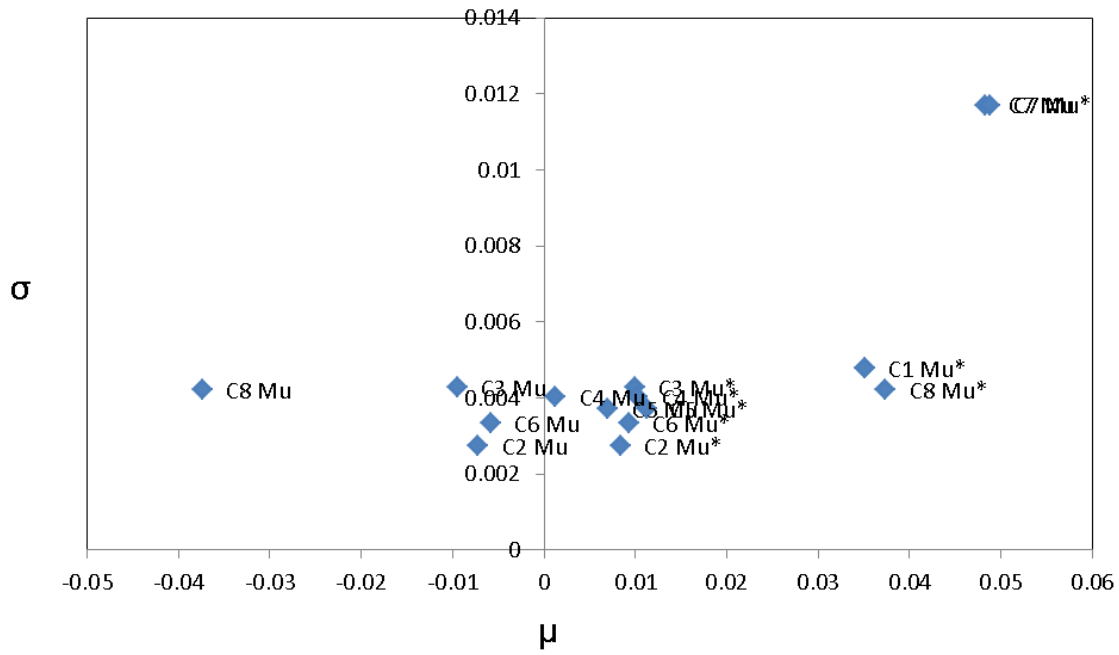


Figure 4-1. Sample Method of Morris results for lead-day 1 forecast results. Both Mu and Mu* are plotted. Normally the y axis has the same range as the x axis to evaluate potential interactive effects (σ is large relative to Mu).

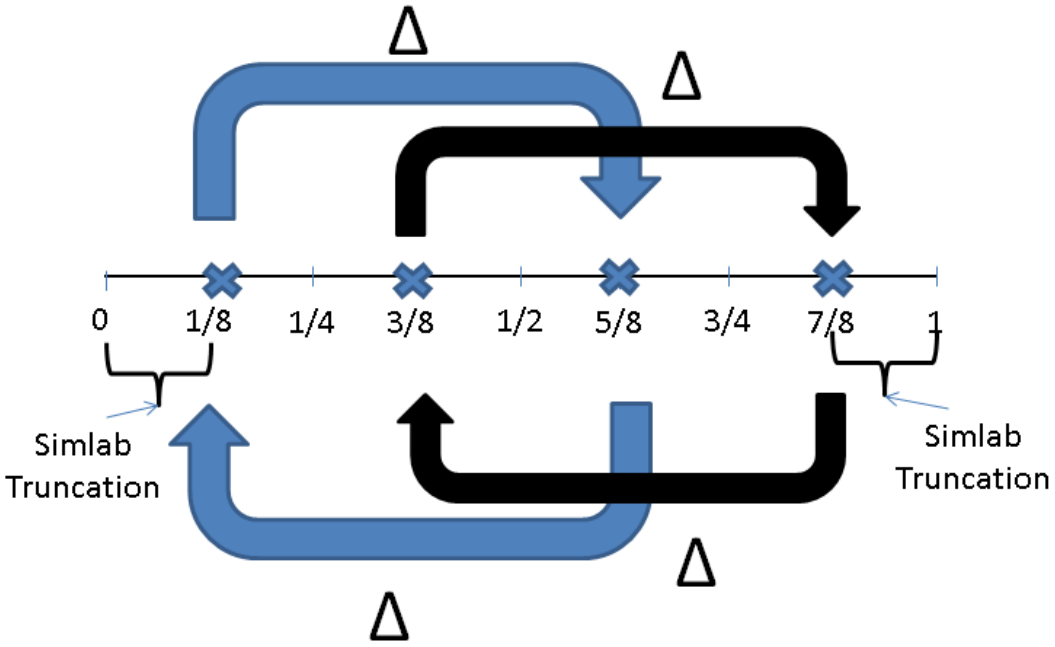


Figure 4-2. Illustration of the Morris sampling strategy for calculating elementary effects of an example input factor. The sampling scheme represented is based on p=4 and Δ=0.5. It can be seen that input values are selected from 4 equidistant locations using jumps of delta.

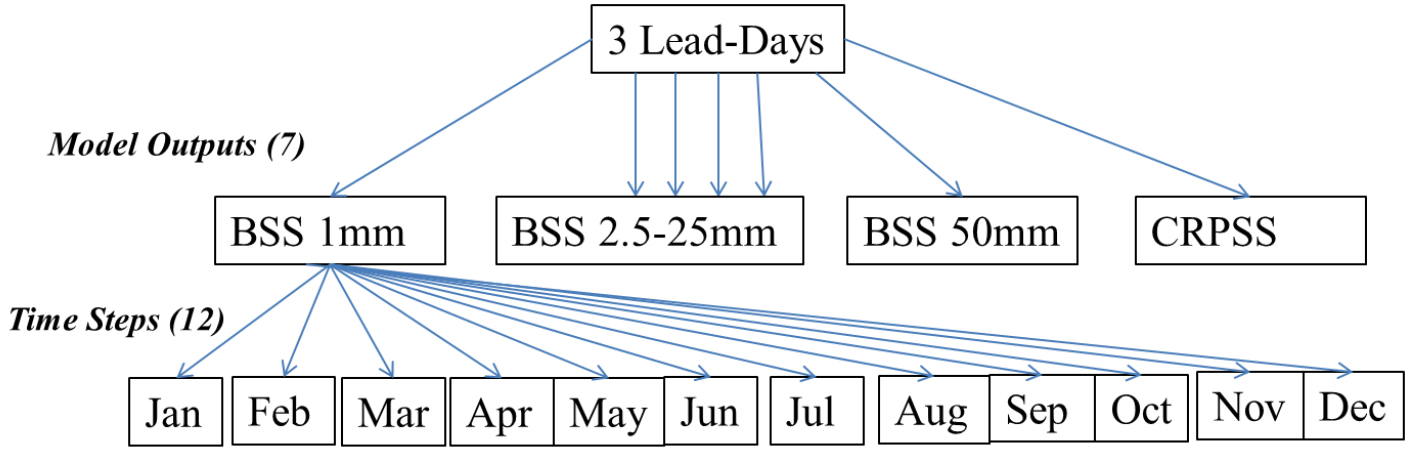


Figure 4-3. Representation of the number of outputs available for sensitivity analysis implementation in this study. The total number of rank order results possible is 254.

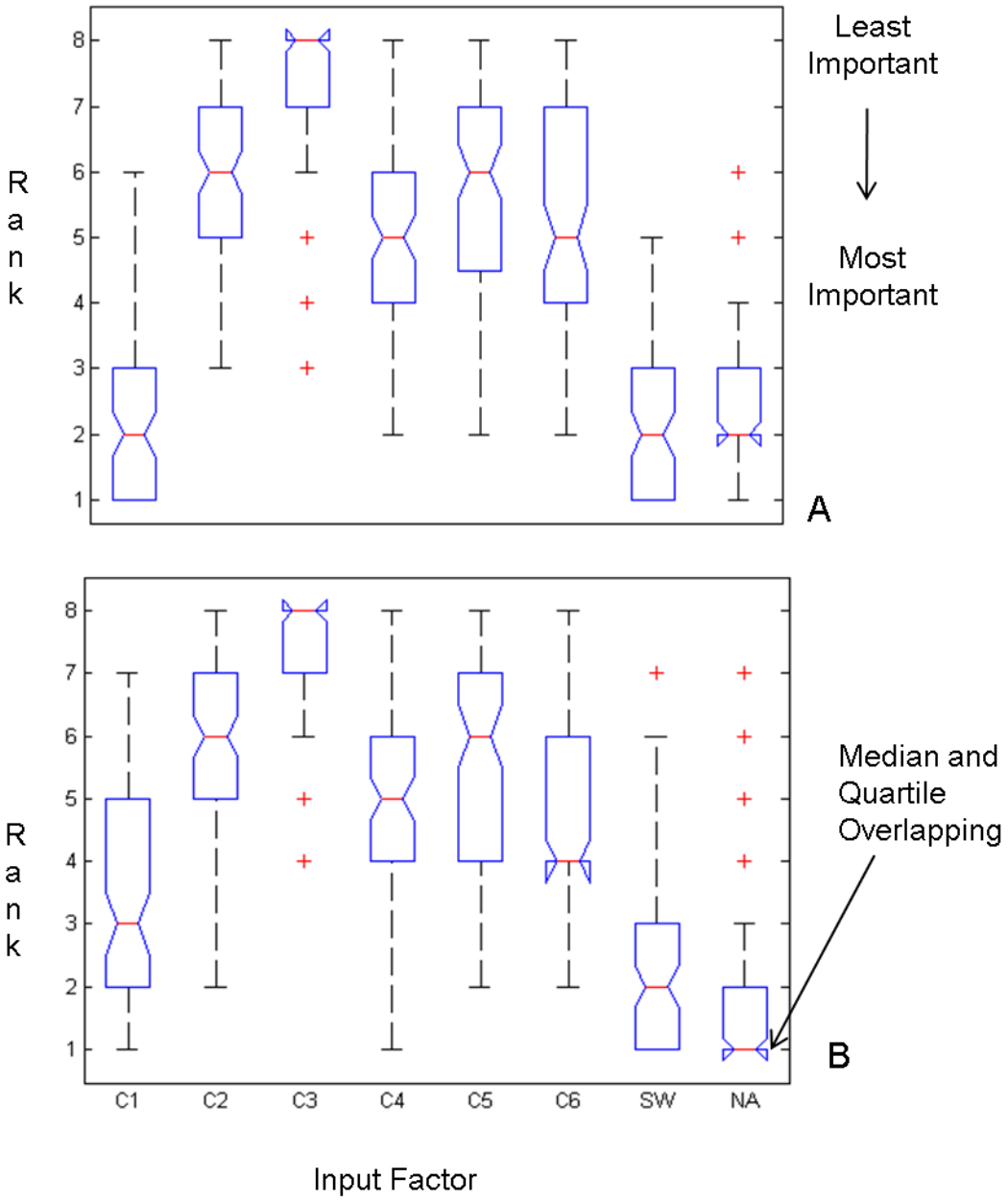


Figure 4-4. Boxplot representing ensembles of the relative sensitivity screening rank order of 8 input factors for all skill scores and all months. A) Lead-day 1. B) Lead-day 5

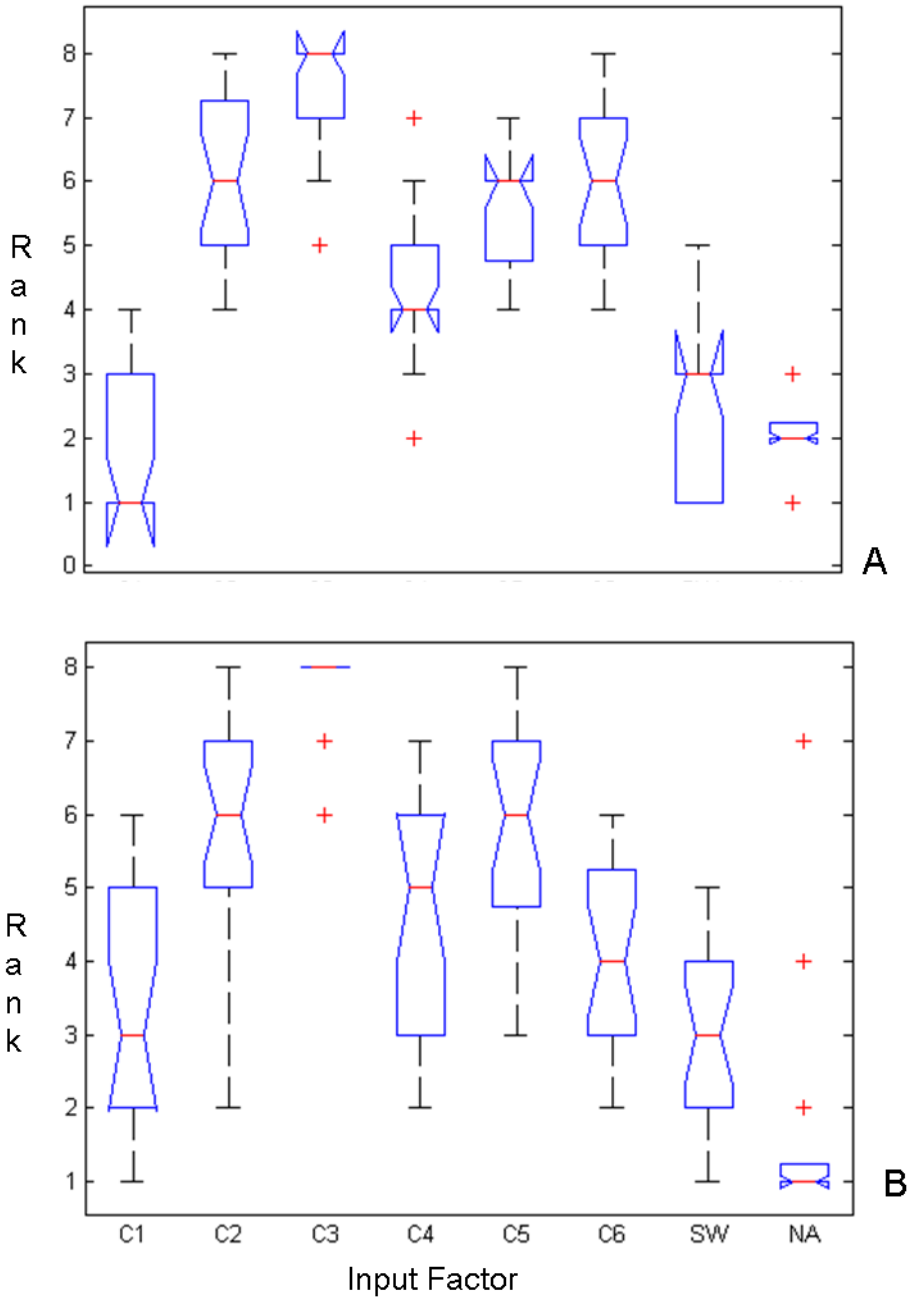


Figure 4-5. Boxplots representing seasonal variation in the ensembles of the relative sensitivity screening rank order of 8 input factors for all skill scores. A) Winter (Dec-Feb). B) Summer (Jun-Aug).

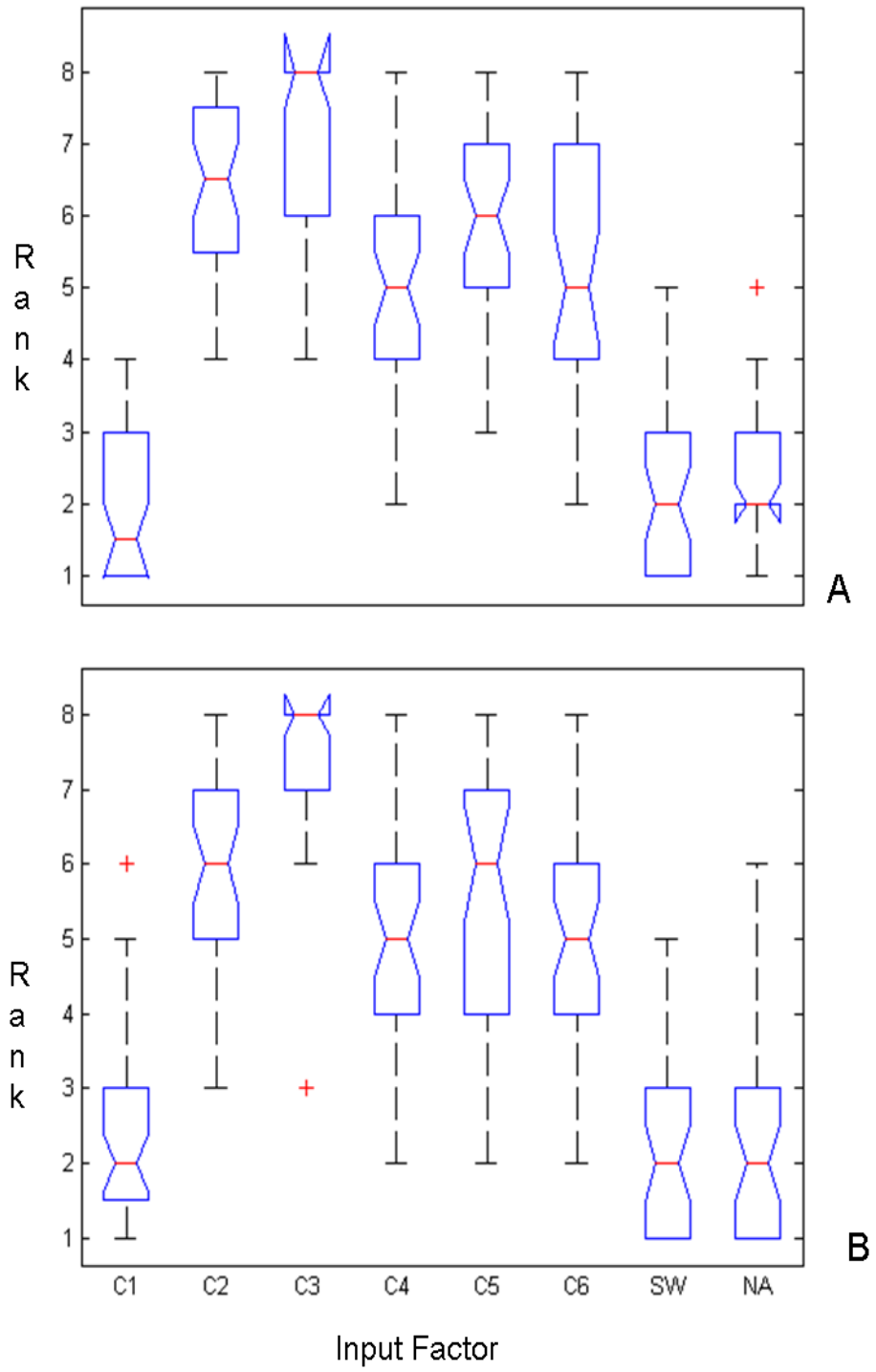


Figure 4-6. Boxplot representing variation in the ensembles of the input factor sensitivity screening rank order based on high and low precipitation event thresholds. A) Outputs from BSSs with thresholds of 1, 2.5 and 5mm. B) Outputs from BSSs with thresholds of 10, 25 and 50mm.

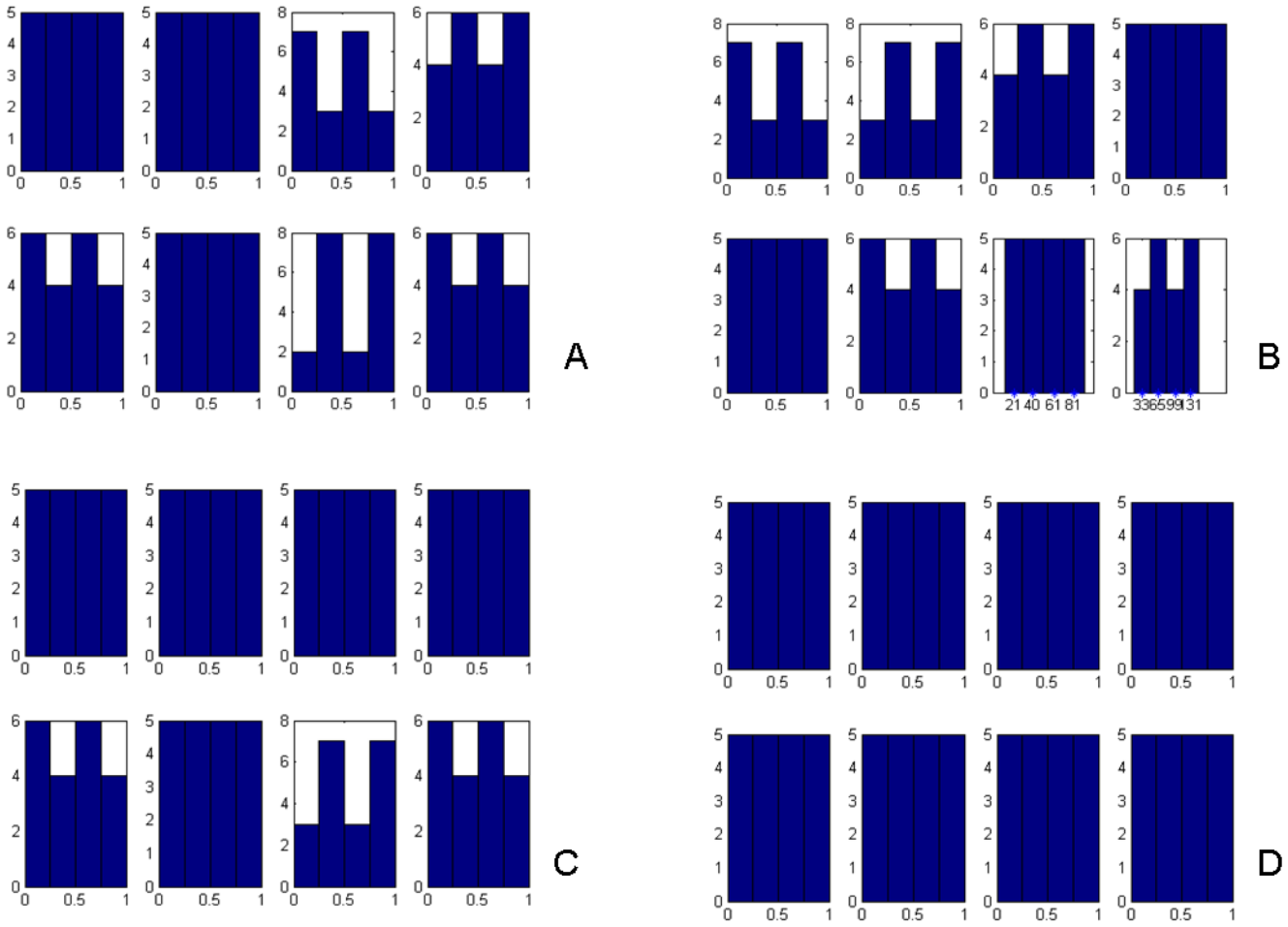


Figure 4-7. Method of Morris input factor distributions for 4 different input generation methodologies. A) Simlab. B) Campolongo. C) Ruano. D) Khare.

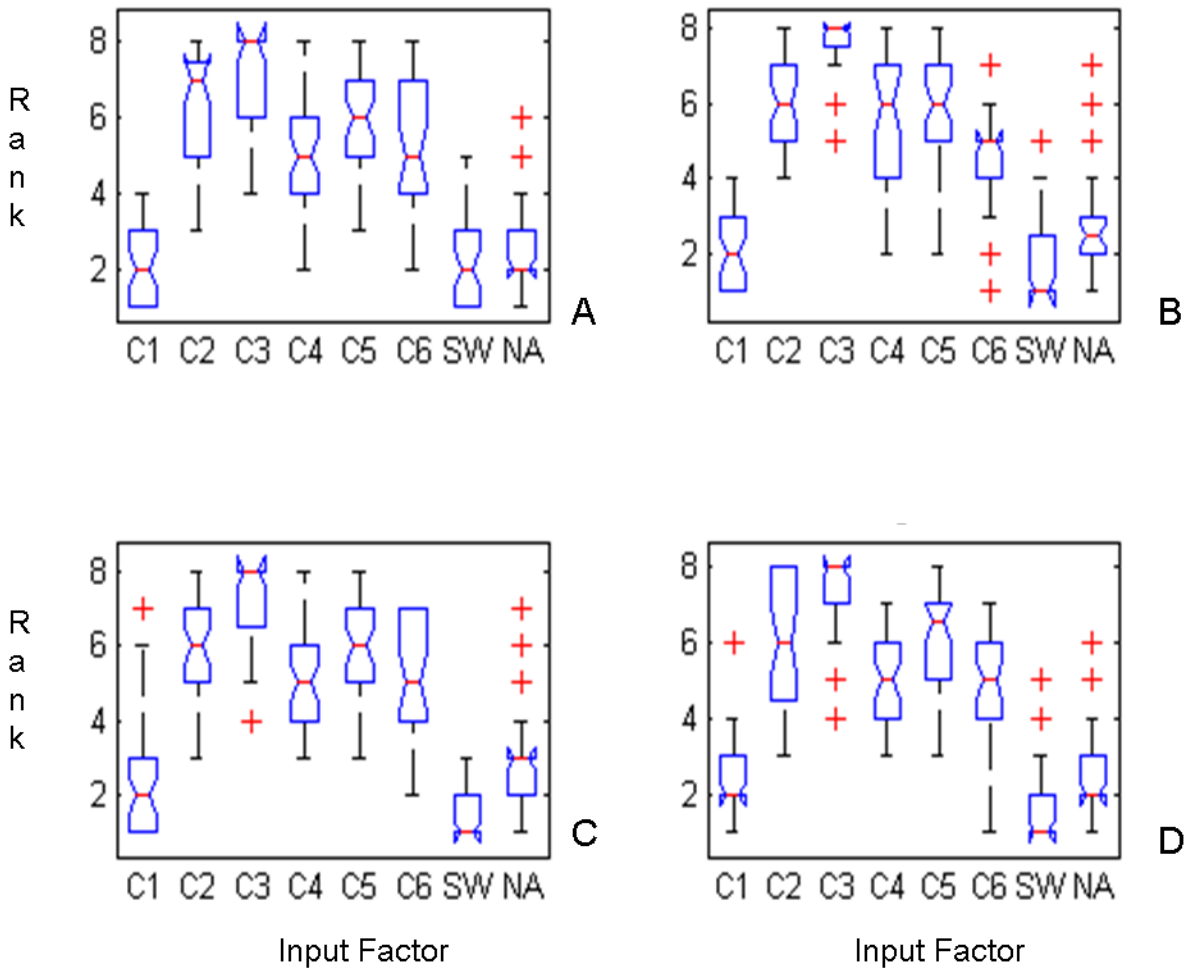


Figure 4-8. A comparison of input factor rank order based on four different sample generation schemes. A) Simlab. B) Campolongo. C) Ruano. D) Khare.

CHAPTER 5
QUANTITATIVE GLOBAL SENSITIVITY AND UNCERTAINTY ANALYSIS OF THE ANALOG METHOD TO ASSESS THE VALUE OF MULTIPLE PREDICTOR FIELDS AS A FUNCTION OF LEAD-TIME, SEASONAL VARIABILITY, AND THE MAGNITUDE OF FORECASTED PRECIPITATION EVENTS

Background

A Global Sensitivity/Uncertainty Analysis (GSA/UA) methodology, and subsequent Monte-Carlo (MC) filtering, is conducted on the AM method in order to ascertain if there are seasonal variations in input factor settings related to analog selection predictors and physical phenomena. Secondary objectives of this study are to quantitatively prioritize multiple predictor input factors to determine the value of expanding the direct analog selection technique in terms of forecast quality, and determine if there are other input parameterization sets that can be used according to the needs of the user, specifically with regard to precipitation event thresholds and lead–day propagation.

Sensitivity analysis is a technique used by modelers to gain insight into a model, specifically the model behavior induced by uncertain input factor parameterization and how it propagates to the output. Comparison of the variability in model output resulting from input factor fluctuations is essentially the resolution of a derivative, the change in a single input factor over the resulting change in model results. The simplest form of this technique is known as OAT sampling whereby the model undergoes a series of simulations changing a single input parameter for each consecutive model realization (Haan, 1995). A significant limitation to this type of technique is that the calculated sensitivity is localized to a single position (base point) and does not give an assessment of model sensitivity throughout the input parametric space (Saltelli 2006, 2008). Local assessment of sensitivity can be effective for linear models, where linear extrapolation

can be used to produce results away from the base point; however, it is not effective in non-monotonic models (Saltelli, 2004). GSA, a methodology which is based on input sampling throughout the parametric space with subsequent analysis of variations in the resulting model output, is capable of overcoming this limitation.

Unlike local sensitivity techniques, GSA assesses changes in model output as each input factor varies throughout the range of all possible values in the input domain. In this way the behavior of model output variability can be quantitatively attributed to the propagation of individual and, depending on the technique used, combined input factor uncertainty. Numerous techniques are used for GSA and they can be grouped into three main categories: screening, regression, and variance decomposition. There are other techniques such as meta-modeling (a simplified model based on analyzed model's behavior) which do not fall into these three categories, but the most widely used techniques do. Screening methods such as the method of Morris are an extension of local sensitivity analysis in that local sensitivities from random points throughout input parametric space are used to calculate an averaged global sensitivity (Morris, 1991). While these techniques have the benefit of low computational cost, the results provide a qualitative and not quantitative assessment of relative input factor effects. For this reason, these methods are used prior to more computationally expensive GSA techniques in order to discover factors that have little to no effect on output variability. Once identified, the factor can be removed from subsequent testing.

Regression methods, as the name implies, perform sensitivity analysis with the use of regression techniques to estimate model sensitivities (Saltelli 2000). A major limitation to these techniques is that they are not effective for nonlinear or non-

monotonic models (Saltelli 2005). Saltelli also showed concern in that GSA is designed for use in uncertainty propagation while in general, regression techniques remove variability from the input parameter thereby giving an indicator value that does not incorporate the full range of input and output uncertainty (Saltelli 2006). The use of variance based methods can effectively overcome these limitations.

Variance based methods, such as the method of Sobol' (1993), Fourier Amplitude Sensitivity Test (FAST) (Cukier, 1973) and extended Fourier Amplitude Sensitivity Test (eFAST) (Saltelli, 1999) are based on an ANOVA-like decomposition of the total output variance using MC simulations (see Sobol Method Background section). The MC simulations are produced from inputs values selected using a predefined sampling strategy that accounts for the range of input factor uncertainty. Analysis is performed by calculation and review of input factor sensitivity indices. For a given input factor, the sensitivity index is a value that corresponds to the proportional contribution or relative effect that factor has on the total model output variability. Calculation of sensitivity indices is performed using an ANOVA-like decomposition of the total output variance in terms of a summation of individual input effects (first order) as well as higher order interactive effects (Sobol' 1993).

With the selection of methods available, it is important to define the goal of analysis before selecting an individual or combination of techniques for use. Applicable goals consist of factor fixing, factor prioritization, variance cutting and factor mapping (Saltelli 2006; Ruesser 2011; and Cacuci 2003). Factor fixing is based on the identification of input factors that have little or no effect on output variability. In these instances a modeler's efficiency can be improved by fixing the factor to a singular value,

within its particular range of uncertainty, and not spend time researching and calibrating an optimum value. Factor prioritization assesses the level of impact an input factor has on output variability in relation to the other input factors. Variance cutting identifies the minimal set of input factors that can be fixed in order to achieve a predefined reduction in uncertainty. Factor mapping identifies input factors that have the greatest impact on model output for a specific region of input space. While all can be used to gain insight into model output uncertainty, factor mapping is based on MC filtering while the other three base their analysis on model variance (Saltelli, 2004). The analysis objectives, once understood can drive the determination of GSA technique(s) to use based on their inherent strengths and weaknesses.

While GSA effectively apportions output variations to individual and combined effects resulting from individual input factor uncertainty, it is only one piece of the puzzle. Without identifying the range of variation, knowing the contributing proportions has limited impact. Uncertainty Analysis (UA) can be used to quantify the total unconditional variation in model output resulting from recursive simulations produced from uncertain input parameters. This is accomplished simply by constructing probability distributions from ensembles of model output. The mean and more importantly the spread of the probability distributions give a valuation of the range and scalar magnitude of the model output variability.

The implementation of GSA/UA using MC simulations have the added benefit of a large archive of model simulation results based on input factor uncertainty. This archive affords the opportunity to assess optimum input factor settings using an MC filtering technique to decrease uncertainty as well as calibrate a model (Rose, 1999;

Saltelli, 2008). The basis of this technique is to determine a model output threshold that effectively splits the results into acceptable and unacceptable outputs based on risk. Input factor distributions that lead to acceptable output results are termed “behavioral” and conversely distributions that lead to unacceptable results are considered non-behavioral (Saltelli, 2004). These two distributions, for each input factor, are then statistically evaluated using the Smirnov test. The null hypothesis that is evaluated with this test is that the two distributions, behavioral and non-behavioral, are representative of the same cumulative probability distribution. The alternative hypothesis is that they are from different distributions.

The Sobol’ method for GSA is a variance based technique whereby model output variance is decomposed in an ANOVA-like fashion into fraction of output variability attributed to individual input factors as well as those induced by their interactive effects. The generic decomposition of the unconditional output variance V(Y) can be seen in Equation 5-1.

$$V(Y) = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<m} V_{ijm} + \ldots + V_{12\ldots k} \quad (5\text{-}1)$$

where; V_i is the output variance attributed to input factor i, V_{ij} is the variance attributed to the interactive effects of inputs i and j, and k is the number of inputs factors. The number of terms in ANOVA decomposition is based on the number of factors and increases at a rate of 2^k, therefore two sensitivity indices are commonly used to assess the results (Saltelli, 2000).

Sobol proposed using sensitivity measures ST_i and S_i for total and first order sensitivities respectively. S_i is the ratio of the fraction of the total output variance due to the input factor I (Equation 5-2), to the total unconditional variance of the output;

$$S_i = \frac{V_i}{V(Y)} \quad (5\text{-}2)$$

V_i can be written in terms of conditional variance transforming Equation 5-2 into Equation 5-3.

$$S_i = \frac{V(E[Y|X_i])}{V(Y)} \quad (5\text{-}3)$$

where; X_i represents input factor i. In simple terms the sensitivity index S_i can be viewed as the fraction of output variance directly attributed to input factor i. Values of S_i can fall anywhere within the range [0, 1]. When there are no interactive effects present in the model, then all output variance can be described by the conditional variance term of the RHS of Equation 5-3. Correspondingly the total unconditional variance would be equal to the sum of input factors first order conditional variance. Such a model is called a purely additive model.

The second sensitivity index, ST_i, accounts for first order effects as well as variations in model output due to higher order interactive effects. In order to calculate ST_i a second conditional variance is used, $V(E(Y|X_{-i})$ (Homma, 1996). This is the closed contribution to the output variance due to the remaining k-1 factors (Saltelli, 2004). This conditional variance can then be subtracted from the total variance leaving the sum of the variance contributions from all the terms in Equation 5-1 that include X_i. Therefore ST_i can be calculated according to Equation 5-4.

$$ST_i = \frac{V(Y) - V(E(Y|X_{-i}))}{V(Y)} = 1 - \frac{V(E(Y|X_{-i}))}{V(Y)} \quad (5\text{-}4)$$

For i=1 and k=3, Equation 5-5 shows all direct and interactive sensitivity effects for input factor 1.

$$ST_1 = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3} \quad (5\text{-}5)$$

where; S_1 is the sensitivity index for the first order effects and the remaining terms ($S_{1,2}$, $S_{1,3,}$ and $S_{1,2,3}$) correspond to the interactive effects for input factor 1. A quantitative assessment of a factor's interactive effects, in relation to first order effects, can be made by subtracting S_i from ST_i. Input factors with a S_i approaching the magnitude of ST_i indicate very little interactive effects, while small values of S_i with corresponding large values of ST_i are indicative of a factor whose influence on output variability is based on interactions with other factors. Due to the computational expense required to calculate interactive effects for all input factor combinations, the individual interactions are typically not parsed out.

Objectives

The objective (goal) of this study is threefold; 1) to identify the value of incorporating multiple predictors into the direct analog method (AM) by varying the relative influence of the predictor during the analog search algorithm, 2) to identify changes in input factor importance due to seasonal variability, lead-day propagation and the magnitude a forecasted events, and 3) find a relationship between the behavior of this statistically based data driven model and physical climatic phenomena. The first goal can be achieved by identifying the range of output variability indirectly introduced by the individual predictors through variations in their corresponding weighting coefficients. Results for this will come from two aspects of the study: 1) uncertainty analysis, which is used to show the probabilistic range of an output based on uncertain inputs or the total variability, and 2) GSA, which gives an indication as to the fraction of the total variability that can be attributed to individual factors. From this information

modelers can focus their efforts on reducing the uncertainty of predictors that have the largest effect on output variability. A comparison can also be made from a deterministic baseline model simulation to the UA Probability Distribution Function (PDF). In the case of this study the baseline was derived in Chapter 2, the one predictor model.

The second objective can be considered a factor prioritization goal. By quantifying the relative effect individual input factors have under different circumstances (i.e. season or lead-day) and subsequently assessing their change (if any), it is possible to develop multiple input parameterization schemes for differing conditions. The third goal is related to model transparency. Statistical models can be somewhat removed from observable phenomena and when these linkages are lost users tend to be apprehensive in their implementation regarding decision making. If variations in factor prioritization are observed and they can be related to physical phenomena it is easier to instill trust in the model. An arbitrary example would be if the weighting coefficient for precipitable water showed low impact in winter but high impact in the summer, and we identify a similar change in the atmosphere either direct or inverse, it can be inferred that there is a signal propagating through the forecast model. Based on the goals of this study, the method proposed by Sobol' is used for GSA with subsequent MC filtering for input factor mapping.

Study Area and Data Collection

Tampa Bay Water's area of responsibility includes a 6,583 km^2 watershed which intersects 5 counties; Hillsborough, Pasco, and Pinellas. This study was performed using a 9 point GFS reforecast grid ranging from a latitude of 25-30^{0}N and a longitude of 80-85^{0}W (Figure 2-1) with a spatial resolution of 2.5^0 or approximately 200 km.

Two sources of data are required for the implementation of the AM; a simulated gridded Numerical Weather Prediction Model (NWPM) forecast data set and a temporally corresponding observed rain-gauge dataset that spatially corresponds to the NWPM grid. Analog selection is based upon the NWPM data while the observed rain gauge data is used for subsequent bias correction, downscaling and statistical interpretation of local forecasts. The forecast data archive used during the selection process is the reforecast data set produced by the National Center for Environmental Protection using a fixed 1998 operational version of the Global Forecasting System (GFS) model (Hammil, 2006). The archive has a temporal range spanning 30+ years, 1/1/1979 – present, and a spatial resolution of 2.5 degrees or approximately 200km. A 15 member ensemble forecast was produced every day. The ensemble initial conditions consist of a control initialized with NCEP-NCAR reanalysis and a set of seven bred pairs. The breeding method was consistent with the 1998 operational model. The forecast extends to 15 days and was archived every 12 hours (Hamill, 2006b). The reforecast archive can be downloaded at: http://www.esrl.noaa.gov/psd/forecasts/reforecast/data.html.

For this study, rain-gauge data was obtained from the National Climate Data Center and can be accessed at; http://www.ncdc.noaa.gov/cdo-web/. Three rain-gauges from the Tampa Bay Water watershed (St Leo, CYCC3, and Tarpon Springs rain-gauges), with the most complete data over the temporal range of the reforecast archive, were used in this study.

Methodology

AM Input and Output Factors for GSA/UA

The main components of the AM are analog selection and statistical analysis of corresponding observed precipitation data. Given an archive of NWPM forecasts with sufficient temporal range, analogs can be identified using a similarity measure (Sm) to compare the current forecast to an ensemble of historic forecasts. The method used in this study is a static version of the methodology described in Hamill (2006a) and Hamill and Whitaker (2006b). A 9 grid point static spatial window encompassing the study area was used as opposed to the moving window used in the aforementioned studies. Analog selection was implemented by comparing the current days gridded coarse scale precipitation forecast to an archive of historical forecasts. A predefined annual temporal window is used to limit the ensemble of potential forecasts to those that occur during times of similar climatic physics. Forecasts from the current year are omitted from the ensemble of potential analogs in order cross-validate the model. The analog section similarity measure used in this study was the RMSE calculated using nine grid-points. Once the similarity measure for all potential analogs is calculated a predefined ensemble of the closest matches are then selected and the observed rain-gauge precipitation, corresponding to the analog dates, are then used to create a bias corrected, downscaled forecast ensemble. The downscaled forecast ensemble can then be used to produce a probabilistic or a deterministic local forecast.

The AM being tested is an expansion of that assessed and optimized in Chapter 2. Rather than using direct analogs, with precipitation as both the predictor and predictand, this section of the study examines the value of using multiple predictors to identify historical analogs. The predictors consist of precipitation (APCP), precipitable

water (PWAT), relative humidity (RHUM), mean sea level pressure (PRMSL), and two (zonal and meridional) 10 meter wind vectors (U10m and V10m).

In most process-based models, the distribution of input factors is based on performing a literature search confined to the temporal and spatial area represented. In the case of the AM a different approach is necessary. The first factor, search window size, is bounded at the high and low ends due to the desire to adhere to seasonal climatic physics and need for a sufficient analog pool to find a good match. The range chosen was based on a temporal range of 1 month to 6 months or windows of +/- 15 days to +/- 90 days. It was also a discrete uniform distribution due to the necessity of using whole days as an input for data manipulation and evaluating model performance. The second input factor, number of analogs used to create the forecast ensemble, is based on previous literature values proposed by Hamill (2006a) and refined by the optimization study reported in Chapter 2. The results of Hammil's study recommended the use of 75 analogs throughout the Continental United States. With this number as a starting point a crude OAT evaluation of the number of analogs used was performed resulting in an optimum of 125 analogs. This value was chosen based on results that showed for lead day 1, the lowest CRPSS is produced when using the smallest number, 10 analogs, with increasing skill until it plateaus at 45 analogs. From this point the skill remains consistent with a marginal drop all the way to 200 analogs. For the subsequent lead days, the skill shows marginal improvement after 75 analogs all the way to 200 analogs, with no drop. The CRPSS for lead day 7 shows negative values when less than 45 analogs were used. The range chosen for this evaluation was 15 to 150 analogs.

The final 6 input factors are paired to predictors chosen from the reforecast archive for use in analog selection. All of the selected predictors consist of different units and different ranges of numerical values. This needs to be taken into account during the comparison step of determining likely analogs. Initially each predictor is treated separately whereby an individual sensitivity measure Sm_i is calculated in the same manner that it was performed in the direct analog study in Chapter 2 (9pt spatial RMSE between current and potential analog forecasts). The second step is the normalization of each respective similarity measure. This is done, for the case of using RMSE, simply by producing the z-statistic (Equation 5-6).

$$\mathrm{Sm_i^{std}} = \frac{\mathrm{RMSE_i} - \mu_{\mathrm{RMSE}}}{\sigma_{\mathrm{RMSE}}} \tag{5-6}$$

Sm_i^{std} is the standardized value of a single analog RMSE calculation for a single predictor, $RMSE_i$ is the value to be standardized, and μ and σ are the mean and standard deviation of all of the analog RMSE values for a single predictor. The RMSE, μ and σ all have the same units therefore the resulting z statistic is dimensionless. The standardization process has the effect of transforming the distribution to one with a mean of zero and a standard deviation of 1.

After the individual predictor RMSEs were standardized, a weighted Equation 5-7 was developed to determine a multi-predictor similarity measure.

$$Sm_{tot} = C1 * Sm_{APCP}^{std} + C2 * Sm_{PWAT}^{std} + \cdots \ldots + C6 * Sm_{V10m}^{std} \tag{5-7}$$

where; SM_{tot} is the weighted multi-predictor similarity measure, C1 through C6 (Table 4-1) are weighting coefficients corresponding to each individual predictor 9 grid point analog RMSE. The vector Sm_{tot} is then sorted with the lowest value indicating the best analog match. The inputs used for sensitivity screening analysis are the weighting

factors. The weighting factors are given a uniform distribution from 0 to 1. In this way the sensitivity screening will be testing output variability based on the strength of the predictor in the similarity measure.

The output that is tested for variability consists of several skill measures described in detail in Chapter 2. The Brier Skill Score (for a detailed description of BSS, refer to Chapter 1) is analyzed over 6 precipitation thresholds due to its ability to evaluate relative forecast quality with respect to climatology in regard to the size of a precipitation event. The thresholds assessed were 1, 2.5, 5, 10, 25 and 50 mm of rain in a 24 hour period. The second skill measure assessed was the Cumulative Rank Probability Skill Score. This measure shows comparative skill across the entire distribution of potential precipitation events. The GSA procedure was also analyzed for 1, 3, and 5 lead days. For purposes of clarity the daily results were grouped into 12 monthly averages in order to assess if there is a difference in input factor ranking based on seasonal variation. This leads to a GSA based on 254 model outputs (Figure 4-3). A refinement of the model output sensitivity results presented is discussed in the results section.

Sobol' Method

In order to determine the values of S_i and ST_i, it is necessary to perform integrations to resolve the components of the function derived from the ANOVA decomposition of the objective function under analysis. Saltelli (2002) proposed an efficient Monte Carlo based approach that is capable of calculating the Sobol' sensitivity indices at a cost of N(k+2) simulations. In this procedure the first step is to generate two sample matrices A and B (Equation 5-8).

$$A = \begin{bmatrix} X_{1,1} & \cdots & X_{1,i} & \cdots & X_{1,k} \\ X_{2,1} & \cdots & X_{2,i} & \cdots & X_{2,k} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,i} & \cdots & X_{N,k} \end{bmatrix} B = \begin{bmatrix} X'_{1,1} & \cdots & X'_{1,i} & \cdots & X'_{1,k} \\ X'_{2,1} & \cdots & X'_{2,i} & \cdots & X'_{2,k} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ X'_{N,1} & \cdots & X'_{N,i} & \cdots & X'_{N,k} \end{bmatrix} \quad (5\text{-}8)$$

where k is equal to the number of factors and n is equal to the sample size used for the MC estimate. Matrix A can be considered the sample matrix and B the resample matrix. Setting the function U_i equal to the integral in Equation 5-9,

$$U_i = \int E^2(Y|X_i = \widetilde{X}_\iota) p_i\left(\widetilde{X}_\iota\right) d\widetilde{X}_\iota \quad (5\text{-}9)$$

Equation 5-3 can be rewritten as Equation 5-10.

$$S_i = \frac{V(E[Y|X_i])}{V(Y)} = \frac{\left(U_i - E^2(Y)\right)}{V(Y)} \quad (5\text{-}10)$$

In Equation 5-9 $\tilde{X}_j$ is one of the input factors X_i fixed to a generic value. U_i as an MC estimate ($\widehat{U}_\iota$) can be determined from model output Y obtained from output obtained using matrices A and a new set of matrices BA (Equation 5-11);

$$BA_i = \begin{bmatrix} X'_{1,1} & \cdots & X_{1,i} & \cdots & X'_{1,k} \\ X'_{2,1} & \cdots & X_{2,i} & \cdots & X'_{2,k} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ X'_{N,1} & \cdots & X_{N,i} & \cdots & X'_{N,k} \end{bmatrix} \quad (5\text{-}11)$$

BA_i is a matrix where all factors except X_i are resampled. $\widehat{U}_\iota$ can then be calculated according to Equation 5-12.

$$\widehat{U}_\iota = \frac{1}{N} \sum_{j=1}^{N} f\left(X_{j1}, \ldots, X_{jk}\right) f\left(X'_{j1}, \ldots, X'_{j(i-1)}, X_{ji}, X'_{j(i+1)}, \ldots, X'_{jk}\right) \quad (5\text{-}12)$$

The second term, $E^2(Y)$, can be calculated as the mean of results derived from sample matrix A. Rewritten in terms of matrix results it becomes Equation 5-13.

$$S_i = \frac{\frac{1}{N} \sum_{j=1}^{N} f(A^j) f(BA_i^j) - f_0^2}{\frac{1}{N} \sum_{j=1}^{N} [f\left(A^j\right)]^2 - f_0^2} \quad (5\text{-}13)$$

where the component f_0^2 can be calculated according to Equation 5-14.

$$f_0^2 = \left[\frac{1}{N}\sum_{j=1}^{N} f(A^j)\right]^2 \quad (5\text{-}14)$$

For the MC calculation of ST_i a similar procedure is performed but rather than use U_j to rewrite the conditional variance term in S_i, U_{-i} is used for the conditional variance term. The MC estimate ($\widehat{U}_{-i}$) can be calculated according to Equation 5-15.

$$\widehat{U}_{-i} = \frac{1}{N}\sum_{j=1}^{N} f\left(X_{j1}, \dots, X_{jk}\right) f\left(X_{j1}, \dots, X_{j(i-1)}, X'_{ji}, X_{j(i+1)}, \dots, X_{jk}\right) \quad (5\text{-}15)$$

In order to perform this calculation a set of input sample matrices AB_i (Equation 5-16)is required.

$$AB_i = \begin{bmatrix} X_{1,1} & \cdots & X'_{1,i} & \cdots & X_{1,k} \\ X_{2,1} & \cdots & X'_{2,i} & \cdots & X_{2,k} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X'_{N,i} & \cdots & X_{N,k} \end{bmatrix} \quad (5\text{-}16)$$

AB_i is a matrix where all factors sampled except X_i which is from the resampled matrix B. Substituting the MC estimate of U_{-i} into Equation 5-10 results in Equation 5-17.

$$ST_i = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} f(A^j) f(AB_i^j) - f_0^2}{\frac{1}{N}\sum_{j=1}^{N} [f(A^j)]^2 - f_0^2} \quad (5\text{-}17)$$

Performing this MC algorithm for the N model executions for sample matrix A and B and N*k simulations for AB_i and BA_i resulting in a total of N*(2k+2) runs. In order to adequately resolve the MC estimations Saltelli (2005) recommended a range for N of 500-1000. It has been recognized that computationally expensive models make that number of simulations prohibitive and adequate results have been achieved with N values as low as 30. It should be noted however that with higher values of N the more

precise the sensitivity estimates and complex non-linear models require large N for stability purposes (Saltelli, 2001, Crosetto, 2001, Liburne, 2009).

Sobol' quasi random sequences were used to generate input sample matrices A, B, AB_i and BA_i. This form of random number generation is derived from primitive polynomials using binary bit comparisons and exclusions (Sobol' 1967). The Sobol GSA technique uses random samples generated from LPt Sobol sequencing. This form of pseudo random number generation has proven to be efficient in providing locations in multivariate input space with relative uniform density (Sobol, 1993).

Monte-Carlo Filtration

This study uses MC filtering to comparatively assess input factor settings based upon seasonal, lead-day and event magnitude considerations. These comparisons are not necessarily behaviorally based but rather optimization based. Additionally, a definitive risk based acceptance threshold is not available. The model output results are in the form of skill scores which have a potential range of one (perfect score) to negative infinity where any value greater than zero is a passing result showing the AM forecast is of higher quality than the reference forecast. For these reasons, a modification in the MC filtering technique is implemented whereby MC thresholds are determined by analysis of multiple MC filtration results using percentage based behavioral and non-behavioral splits as well as the skill score corresponding to the point of division. In this way comparative value can be maintained during subsequent analysis of optimum input setting variability. The reason for this approach will become apparent in the presentation of results which identifies the distinct and different ranges of skill scores with respect to season, lead-day, and event magnitude.

As a result of performing GSA, each output (i.e. CRPSS lead-day 1 and 5) has 4608 model results available for MC filtering. MC filtration was performed seasonally and annually on CRPSS, 2.5mmnBSS, and 25mm BSS model output. Potential skill score thresholds were determined using the top 50, 60, 70, and 80 percentile of the total ensemble of model results. MC filtration using the 2 sample Smirnov test was performed using the aforementioned divisional splits. The Smirnov test is used as a statistical measure to evaluate the individual input factor's behavioral and non-behavioral cumulative probability distributions based on hypotheses described in Equation 5-18 and Equation 5-19.

$$H_0: f_m(X_i|B) = f_n\left(X_i|\overline{B}\right) \quad (5\text{-}18)$$

$$H_1: f_m(X_i|B) \neq f_n\left(X_i|\overline{B}\right) \quad (5\text{-}19)$$

Where, $f_m(X_i|B)$ is the behavioral cumulative distribution of input factor X_i, and $f_n\left(X_i|\overline{B}\right)$ is the non-behavioral cumulative distribution. The Smirnov test statistic is:

$$d_{n,m}(X_i) = sup_y\left\|F_m(X_i|B) - F_m\left(X_i|\overline{B}\right)\right\| \quad (5\text{-}20)$$

Where sup_y is the maximum vertical distance between the behavioral and non-behavioral cumulative distributions and || indicates absolute value. It is worth noting that unlike the one sample test, the two sample Smirnov test does not use the test statistic in relation to a critical value to determine if the null hypothesis is accepted or rejected. Instead the relationship between the p-value and the confidence level (CI) is used. If the p-value is less than the CI, then a test decision of 1 is returned rejecting the null hypothesis, and a 0 indicating acceptance of the null hypothesis otherwise. Prior to the performance of the Smirnov test, both the behavioral and non-behavioral data sets

are converted to an empirical cumulative distribution function using the ecdf command in Matlab.

Once input factor Smirnov test results were produced for the multiple percentile thresholds and outputs, they were analyzed along with the corresponding skill score thresholds to determine an acceptable percentile to use as a threshold. A comparative assessment of the behavioral input factor ensembles could then be conducted with regard to seasonality, lead-day propagation and forecast event magnitude. These comparisons were visually as well as statistically evaluated using the same Smirnov test described previously.

As can be seen in Equation 5-7, input factors C1 through C6 are coded to introduce relative weights to the individual predictor similarity measures. In this way the overall similarity measure used to identify multiple predictor analogs can be adjusted according to situational needs. While the uncertainty range and distribution of these factors is uniform from zero to one, it is their relative weight that creates the influence. For this reason, all of the MC filtration tests were performed on input factor distributions using normalized values. Normalization was calculated simply by dividing an input factors value by the sum of all of the corresponding factors, itself included. If this step was not performed, it can be seen that there is potential for different combinations of input factor settings would produce identical model results. An example would be a setting with all input factors set to 1 producing the same exact analogs, based on the similarity measure Sm_{tot} , as a setting with all weighting coefficients set to 0.5.

Presentation of Results

In this study the AM has a single run time of 6 to 13 hours which makes it a computationally expensive model. A value for N of 256 was used resulting in a

requirement of 4608 simulations. For a conservative run time of 6 hours, if this analysis were to be performed on a single computer it would take over 1000 days to finish. Due to the extensive nature of computational time required, the simulations were run using the University of Florida High Performance Computer network (UF-HPC). In this way 100 model runs were simulated in batch mode on 100 computational nodes thereby reducing the time by a factor of 100 thus making it more manageable. Sample generation as well as sensitivity index calculation was performed using Simlab V2.2 (Saltelli, 2004).

There are difficulties in analyzing the results of a sensitivity analysis that uses multiple outputs. Individual results are produced whereby each input factor is relatively ranked based on variations in individual outputs. A question arises in terms of which results, if contradictory, should take precedence. Previously the screening method of Morris was performed on the AM using the same model structure with regard to inputs and available outputs. In that study, individual factor prioritization results were grouped in order to answer questions of seasonality, precipitation event magnitude and lead-day propagation. While that method of representing and analyzing results is appropriate for an assessment of the relative, *qualitative,* rank of importance, this is not necessarily a good approach for a *quantitative* assessment as provided by the Sobol' GSA method. Numerical results (Si and STi) based on the magnitude of output variation may be skewed if combined with results from variations that have a significantly different distribution. In order to assess if grouping was a tractable approach, combined UA results were analyzed. If a PDF produced from a combination of model outputs resulted in multimodal distributions, then grouping was considered unwarranted and potentially

misleading. Using the same result combinations that were implemented in the screening method of Morris (Chapter 4), it was apparent that grouping would not work for Sobol' quantitative results. In all cases the resulting probability distributions were multimodal. Therefore it was necessary to select individual sensitivity results that are representative of the types of information pursued by the modeler and develop a concise way of graphically introducing the information to the reader.

With these goals in mind a representative selection of GSA results were chosen for presentation. In terms of skill scores, the CRPSS was the first model output selected as it is a cumulative based skill score that captures forecast quality for all magnitudes of precipitation events (Chapter 1). Unlike the CRPSS, the BSS is not cumulative. It is a probabilistic forecast quality measure based on precipitation event exceedance thresholds (Chapter 1 and Chapter 2), consequently using multiple BSSs is a common practice in evaluating a forecast's ability to capture events of differing magnitude. This study used 24 hour precipitation event thresholds of 1, 2.5, 5, 10, 25 and 50mm. The BSS with a threshold of 2.5mm was selected to represent rain occurrence. While in practice, many meteorologists use 1mm, over a 24 hour period, as an indicator of a precipitation event the World Meteorological Association (WMO) uses 2.5mm as the lower limit for precipitation event identification. On the upper end of precipitation event thresholds, 25mm was chosen. The highest available threshold was not chosen for the following reason. The AM is based on pattern matching which by its nature has an inherent weakness, difficulty in forecasting extreme events. This is due to the limited number of extreme events in the reforecast archive. While a 50mm event in 24 hours is not considered extreme, it still follows that there are fewer good matches

in the archive. The studies reported in Chapter 2 and Chapter 3 indicated there is an inverse relationship in the skill score and precipitation event magnitude.

Previous studies (Chapter 2 and Chapter 4) have shown that, of the three lead-days available (1, 3, and 5), lead-day one consistently has the highest forecast accuracy with declining skill scores as forecast lead-day propagated forward. When it is possible to present results for all three lead-days without inundating the reader, all three are used, otherwise lead-day 3 is omitted in order to preserve the extremes of the output range. In the spatial domain of this study, Tampa Bay Florida, there are two climatic extremes with regard to precipitation. The winter is the dry season where precipitation events are infrequent and usually driven by large scale weather fronts, and the summer is the wet season, where precipitation is temporally common while being spatially chaotic, and are usually driven by convection. In this study six representative months were chosen; two from winter (Dec, Jan, and Feb), and two from summer (Jun, Jul and Aug). Three in each season were selected in order to evaluate seasonality with replication.

Results and Discussion

Prior to analyzing the GSA results an UA was performed on model outputs in order to identify the overall range of output variability. When analyzing multiple GSA results to determine potential ways to improve model implementation, it is worthwhile to know these ranges. If, for example, a GSA result produced from model output with negligible variability indicates factor “k” is overwhelmingly important, and a second result derived from output with significant variability contradicted that indication, knowledge of the respective uncertainty ranges would aid the modeler in deciding how to proceed. Remember that the sensitivity index, Si is calculated by dividing an input

factors conditional contribution to variability by the unconditional total output variability therefore it is standardized to its particular output distribution.

Uncertainty Analysis Results

Probability distributions were created for every model output using the 4608 simulations generated for GSA. A statistical summary of the PDFs for all representative outputs cited earlier are listed in Table 5-1 and graphical representations can be seen in Figure 5-1, Figure 5-2 and Figure 5-3. The mean and median give a representation of where the output variability is located on a quantitative scale while the 95% confidence (bold) interval gives an indication of the range of variability. This information is of value in that skill score output is considered good if it is non-negative and has an upper limit of 1 (perfect score). Therefore, where the distribution lies on this scale identifies the value of the forecasted output. Additionally, the 95% CI gives an indication of the total output variance. In a case of model output variability with little to no range, then uncertainty in model input factors does not significantly propagate to the output. This could be an issue with the model's lack of sensitivity, or due to the selection of input factors tested did not cast a wide enough net and potentially important factors were omitted.

The UA results in Table 5-1 are partitioned by lead-day forecast, summer and winter months, and skill score. Indications are that the skill scores for lead-day 3 and 5 forecasts (winter and summer) as well as lead-day 1 summer forecasts have a relatively low range of output variability in comparison to the winter lead-day 1 skill scores. The 2.5mm BSS of lead-day 1 forecasts in June has a 95% CI of 0.06 (0.15-0.21) indicating that changes in input parameterization consistent with GSA sampling only changed the output by 0.06. While this appears to be marginal it should be noted that the value of a skill score is in its comparison to a reference forecast. The skill score range for a good

precipitation forecast is 0-1, and a ‘highly skillful” forecast can be achieved at 0.4 (Hamill, 2006a). With this in mind a range in output variability of 0.06 encompasses 24% of a highly skill-full skill score. For this reason identification of a CI being low should be accompanied by the mean or median value of the range to assess its value in terms of forecast quality improvement.

The most significant range of output variability was for all lead-day 1 skill scores in the winter months, Dec. and Jan. The 95% CIs have values from 0.1 to 0.15. Interestingly, this coincides with the highest means skill scores. However there does not appear to be a causal relationship between the size of the spread and the magnitude of the mean. The fact that the AM shows greater sensitivity for these forecasts is twofold. Although the AM doesn’t have an issue with lead-days (compares lead-day 2 forecasts to lead-day 2 analogs) the NWPM that the reforecast archive is based on does. A NWPM uses current observed physical phenomena as inputs to a Navier-Stokes based model and propagates the model forward in time. The further this type of model gets from the initial conditions, the greater the chance that numerical methods, limited physical representations or the inherent chaotic behavior of natural phenomena on this scale skews the output. For this reason there is logic that the AM would have its greatest sensitivity to input factor variations where the potential for reliable forecasts is at its peak (closest to initial conditions). If this were the only factor influencing the lead-day 1 sensitivity then the variability of lead-day1 forecasts in the summer would also be significant.

On first impression it is surprising the lead-day 1 summer forecasts skill scores did not have elevated variance when compared to lead-day 3 and 5. The summer in

Florida is well known for frequent convective precipitation events. Convective storms are the most difficult to model (Vidale et al, 2003), (mainly due to lack of modeling skill in cloud formation); add to that their temporally and spatially sporadic nature and you have a recipe for uncertainty which would propagate through the AM, yet this is not the case. There may be some validity in this line of reasoning that may provide an answer. Yes, summer forecasts in Florida have less quality due to the aforementioned conditions, but skill scores are based on a comparison to a commonly used reference forecast, climatology. It could be that the potential level of skill score variability is offset by climatology's inability to cope with the same issues. In other words when the forecast misses, that's OK, so does the reference.

Using the UA results, histogram of GSA model outputs, the single predictor optimized AM CRPSS was evaluated to assess its relative performance to that of the multi-predictor model (Figure 5-4). Although the single predictor model would have its own uncertainty domain if it went through GSA the comparison is valid. If, for multiple predictors, all of the weighting coefficients were set to 0 and the precipitation weighting coefficient were set to 1 then they would be the same. Comparison plots were generated for all representative months, Dec, Jan, Jun. and Jul. The vertical line is the single predictor result plotted against the PDF and CDF of lead-day 1 CRPSS GSA ensembles. In all months the single predictor result fell well below the peak of the uncertainty region. One thing that this says is either the crude OAT sensitivity analysis performed in Chapter 2 was insufficient at locating an optimum input set, or the use of multiple predictors has the potential of improving skill scores beyond the level of the single predictor search algorithms.

GSA

Three sets of sensitivity analysis results (CRPSS, 2.5mm BSS and 25mm BSS) are presented in box-plot pairs corresponding to lead-day 1 and lead day 5 (Figure 5-5, Figure 5-6, and Figure 5-7). There is a horizontal line across the bottom of the plots which is only a reference where the first order sensitivity index S_i would account for 5% of the total output variability. The box-plots were developed from individual model outputs using an annual ensemble of sensitivity results consisting of all months. This grouping led to the removal of information regarding seasonal variability, therefore Si data points from three winter and three summer months were included in the plot. The upside-down triangles overlaying the ST_i box plot correspond to S_i measures for June, July and August, and the right-side up triangles overlaying the ST_i boxplot are representative of December, January, and February Si measures. Multiple seasonal data points were used for replication purposes.

Figure 5-5, Si and ST_i results for CRPSS, indicate that for lead-day 1, SW, NmAn and the weighting coefficient corresponding to precipitation as a predictor, C1, have the greatest effect on output variability. The mean of value of Si for SW, NmAn and C1 were 0.31, 0.16 and 0.22, respectively. While a seasonal influence for C1 is not indicated, the data points show a positive winter influence for search window size and conversely, there is evidence summer produces much larger values for Si for NmAN. The remaining input factors' Si values all fall below 5% of the highest value with C3, the coefficient related to mean sea level pressure appearing to have the mean as well as the 25 and 75% quartiles at 0. While it is worthwhile to identify negligible input parameters, it is very important that type II errors are avoided. A type II error is one in which a false negative is reported. In this case the purpose is to identify an input factors

effect on output variability. While Si is an indication of that effect, it only represents first order (direct) effects. The input factor's level of effect can be increased due to interactive behavior; therefore it is necessary to also look at the value of ST_i which is the total combined first and interactive higher order effects. In this case it can be seen that there is a level of interaction that occurs for all parameters. In fact, all of the box-plots, with the exception of C3, rise above the 5% threshold. Figure 5-5 also shows that there are some interactive effects present in that the boxplots for ST_i are of a greater magnitude. For the 5 predictor coefficients that hade Si values below 5%, their corresponding total effects indices, with the exception of C3 (PRMSL), breach the threshold. It should be noted at this time that the arbitrary negligibility threshold of 5%is not relevant to ST_i. For S_i, if a model has no interactive effects, then the sum of all of the factors would equal the unconditional or total model variance. Part of the determination of S_i is accomplished by dividing the first order variance by the total model output variance which constrains Si to a range of 0-1. Therefore it is easy to comprehend the effective magnitude of a 5% S_i score. As for ST_i, it not only encompasses the effects identified by S_i but also combinatorial effects caused by interactive effects. If one was to add all of the total order effects indices for the factors, then the interactive effects would be counted multiple times, once for each factor in the interaction. This leads to the possibility that the sum of the total order effects being larger than the unconditional model output variability (greater than 1). That being said further evaluation prior to a determination of negligibility is warranted.

Figure 5-5 B indicates behavior based on the propagation of lead-days for CRPSS from 1 to 5. It is immediately apparent that there is a significant shift in input

factor importance. The importance of NmAn overtakes both C1 and search window size as the dominant factor with C1 falling below the 5% reference. The seasonal signal for number of analogs and search window size remains present with summer producing larger effects on output variability by NmAn, and winter strongly influencing SW's effect on output variability.

The seasonal signal for the number of analogs selected and search window size used, is based on the relative difficulty in forecasting the seasonal precipitation regimes. As noted previously, the winter in Florida is a dry season where precipitation is infrequent and is propagated by frontal systems that are usually accompanied by predictable climatic signals. In this case the value of a good analog would have a higher impact. Basically, there is a higher probability that observed precipitation from the date of the matching reforecast climatic pattern would propagate in the same manner as the current forecast/observation relationship. For that reason a higher value is placed on obtaining good analogs, thus the importance of the size of the group of potential forecasts and correspondingly the search window size. In summer, for reasons previously cited, it is more difficult to forecast precipitation. This is due to a more tenuous relationship between forecasted and observed precipitation. Without that strong relationship it becomes less important to get a good match and more important to have a lot of matches that can subsequently statistically evaluated creating a relationship of the model sensitivity to NmAn.

The second pair of graphics analyzed is representative of the AM forecasts ability to predict the occurrence of precipitation (Figure 5-6 A). The 2.5mm BSS model output produced sensitivity index trends not inconsistent with CRPSS. Lead-day 1 again

identifies SW, C1 and NmAn, in order of importance, as the dominant factors by a wide margin. The seasonal data points for SW and NmAn, on the other hand, are no longer as indicative of a seasonal signal. The mean of the Si population for SW and C1 also showed an increase to 0.25 and 0.35, respectively. The rest of the predictor coefficient input factors remain below the subjective 5% reference. Input factors used to produce low precipitation events show a notable difference from the lead-day propagated importance ranking shift seen in the cumulative-based CRPSS results (Figure 5-6 B). While NmAm again overtakes the SW size, the mean of the Si ensemble for C1 (precipitation) is the dominant factor. This is not due to a significant increase in C1 or NmAn but rather a decrease in SWs effect on output variability. It is also noteworthy that the spread of the C1 ensemble tightened while the spread for NmAn increased. The seasonal signal for both SW and NmAn remained through the propagation of lead days. Two of the remaining predictor coefficients S_i box plots penetrated the subjective threshold and became “relevant”. Both C4 and C6, related to relative humidity and the meridional (North-South) 10 meter wind vector increased their level of importance although their mean value remained under 0.05. C4 showed no seasonal signal while C6 seasonal data points indicated that although their level of importance was relatively small its importance was largest during the winter, while the June, July and August Si values overlapped very near to 0. One would expect that if there was a seasonal signal for one of the wind vectors it would be in the winter in which precipitation events are driven by fronts which are by their nature are accompanied by prevailing winds. It may be that this signal is always present but due to the input factors relatively low level of importance it cannot be identified.

The final GSA assessment is based on forecast skill relating to large precipitation events. The model output skill assessment is accomplished by using the AM's 25mm BSS as model output. For lead-day 1, the same three input factors provide the dominant effect on output variability. However the number of analogs shows the highest level of importance for lead-day 1. The increased importance of NmAn could be due to the reduced number of good analog matches present in the archive for larger storm events. With analog matches of lower quality, the number of analogs used for subsequent probabilistic analysis becomes more important. The same seasonal signal is indicated for both NmAn and SW, and for the first time winter shows a preferential effect on C1's ability to effect variability. The zonal and meridional 10 m wind coefficients, C5 and C6, have the highest value of the remaining 5 predictors, indicated by the 75% quartile reaching the 5% threshold. There is also a seasonal component to C5's S_i value with more importance produced in the winter.

Similar to the cumulative CRPSS evaluation, C1 falls below the 5% threshold for lead-day 5 and NmAn relative importance increases to a mean of 0.65 becoming by far the most dominant factor. There is also an indication of a large amount of interactive effects when compared to their low S_i. Figure 5-7 also shows larger interactive affects indicated by increased distance between the means of S_i and ST_i. Due to the difficulty of forecasting precipitation at long lead times, finding lower grade analogs become more important. When these lower grade matches are selected it is plausible that there is increased dependence on the use of multiple predictors interactively in order to help in their evaluation thereby increasing the sensitivity of the model to these inputs. These

conditions would also justify the extreme importance of the number of analogs used for probabilistic evaluation.

A direct look at interactive effects for all model results (Table 5-2) was performed by subtracting input factor S_i indices from their corresponding ST_i index. The evaluation resulted in 576 individual input factor interactive indices. Of these, 124 had values above 0.1, 20 above 0.2, and 4 above 0.3. The first characteristic of the results noted was that the majority of interactive effects over 0.1 (82) populated the lead day 5 and high precipitation threshold forecasts. The second result was that the largest interactive effects were attributable to NmAn and SW. The second result makes sense related to scale. NmAn, SW and C1 consistently have the highest Si. Therefore it follows that if interactive effects are present they would be larger relative to the remaining parameters. It is difficult, although not impossible, to envision a scenario where a parameter has a first order sensitivity index less than 0.05 but shows interactive effects of 0.31. In effect this would mean that while the input factor directly attributes to less than 5% of the total output variability, the same factor accounts for 31% of the total variability through interactive effects.

The first result is harder to explain because there are multiple potential causes. It is possible that when forecasting becomes more difficult for the AM, such as at long lead times or high precipitation thresholds, the reliance on analogs of less value become more important. In this case, the discrimination of analogs with marginal differences in quality may require the interactive contribution of multiple predictors. The grouping may also be due to the PDF distributions of the output results. Aside from high precipitation in lead day 1, the range of variation identified by the 95% CI is small.

It is plausible that at the times when the total variation is at its lowest the indication of interaction would become more visible. This is based on the assumption that the interactive effects in the model are always present to some degree, based on the model structure, and are not linearly related to the total model variance. However, this assumption is difficult to support.

The model was also tested for its additive nature by summing each model result Si's (Table 5-3). If the model is totally additive then they would sum to 1. Cukier (1973) showed, using the FAST method that if the sum of the first order sensitivity effects was greater than 0.6 then the model could be considered mostly additive in nature and treated as such. Results from this study showed that, of the 72 sets of model results, the summation of the first order sensitivity indices fell below 0.7 six times and below 0.6 twice. The majority of results indicated that the model is indeed additive in nature.

Monte-Carlo Filtration

MC filtration tests were performed separately on CRPSS, BSS 2.5mm and BSS 25mm AM model output initially using all monthly results in a single ensemble as well as average annual values. The division between behavioral and non-behavioral conditions was set to 50, 60, 70, and 80% of the highest skill scores for each output. Using the Smirnov test it was found that there was no statistical difference between the distributions of behavioral and non-behavioral input factor settings. However, when the output was tested using seasonal three month ensembles, statistical differences in distributions were identified. A summary of the Smirnov test results for a 70% split can be seen in Table 5-4. The top 30% of the model output results was chosen due to the similarity in Smirnov test results as well the summer and winter skill score thresholds corresponding to 50, 60, 70 and 80% ensemble splits (Table. 5-4). It can be seen in

Table 5-4, while the change in skill score for winter is marginal from 50% to 80%, there is a significant jump in the summer at 70%.

The second aspect of MC filtration used in this study is a comparison of the behavioral input factor ensembles in order based on seasonality, lead-day propagation and event threshold. Figure 5-8, Figure 5-9 and Figure 5-10 give a comparative visual representation of the winter and summer behavioral input factor frequency and cdf for CRPSS, 2.5mm BSS and 25mm BSS, respectively. The corresponding Smirnov tests statistics as well as Smirnov tests for lead-day and event magnitude variation can be seen in Table 5-5. It can be seen that SW and NmAm are the only input factors that show a significant difference in the behavioral and non-behavioral distributions during the summer months. In the winter months, for the low event threshold 2.5mm BSS model results, all input factors show statistical differences in distributions. The high event threshold (25mm BSS) winter results indicate the zonal wind vector coefficient as well as C1, NmAn, and SW are distinct. For CRPSS winter results, again C1, NmAn, and SW are statistically different distributions as well as C2 which corresponds to precipitable water.

The statistical comparisons of the behavioral distributions with regard to seasonality, lead-day propagation and event magnitude, indicate that while the NmAn, SW and C1 are statistically different distributions, other input factors also show seasonal variation. C3 (PRMSL) is again the only factor that shows no difference in any instance. Low event threshold seasonal comparison has the most input factors showing statistical significance. These results reinforce the GSA results in that their is some seasonal significance to the selection of input factor parameter values. Conversely the

event magnitude and lead-day propagation results indicate that only C1, NmAn and SW are of importance when identifying variations in input settings.

Chapter Summary

In this study the multi-predictor AM underwent a GSA/UA using the Sobol method in an attempt to;

- Ascertain if there are seasonal variations in input factor settings related to analog selection predictors and physical phenomena
- Quantitatively prioritize multiple predictor input factors to determine the value of expanding the direct analog selection technique in terms of forecast quality
- Determine if there are other input parameterization sets that can be used according to the needs of the user, specifically with regard to precipitation event thresholds and lead-day propagation.

The method of Sobol' was selected for its ability to quantify individual input factor direct and interactive effects on output variability. This is done by calculating the Sobol' indices, S_i and ST_i, using an ensemble of results produced from a pool of quasi random samples. From these two indices, three pieces of information about an individual input factor are available for assessment; input factor first order effects on output variability (S_i), total order effects (ST_i), and by using the difference, the interactive effects (ST_i-S_i). Combined with the quantification of total output variability with provided by UA this becomes a powerful tool.

Based on UA results, the largest range output variability occurs for lead day 1 forecasts in the winter regardless of the precipitation event forecasted. The cumulative-based CRPSS had a range of 0.10-0.12 (95% of the Confidence Interval), the low threshold BSS was 0.13-0.15, and the high threshold was 0.12-0.13. All other output results (lead-day 1 for summer and lead-day 3 and 5 for both seasons) ranged from 0.03-0.08. This information can aid in a value judgment as to whether a particular input

factor's fractional contribution and its potential improvement holds value. The distributions providing this information are based on skill scores that have a range for 1 to -∞, where any positive value indicates that the model's forecast is good in relation to a reference forecast with one being a perfect score. Negative values conversely indicate the reference forecast is of better quality and therefore more value to the user. In the literature a Brier skill score of 0.4 has been presented as "Highly skillfull" (Hammil, 2006). A second descriptive statistic of the UA distributions relating to total output variability is the mean or median of the range. This indicates the scale of the distribution that correlates to the quality of model forecast.

A comparison was performed on the optimized single predictor AM (Chapter 2) to the UA probability distributions for CRPSS lead day one results. In all four months evaluated the single predictor results fell well below the peak of the distributions. This implies one of two things. Either the OAT optimization method used in Chapter 2 did not find the single predictor optimum settings, or some form of multiple predictor strategy is warranted.

The GSA results are presented as ensembles consisting of 12 monthly values. Individual monthly results are also reported to determine seasonality. The results were reported in terms of lead day 1 and 5 for CRPSS, low precipitation threshold BSS, and high precipitation threshold BSS. For all lead-day 1 results the sensitivity indices indicated the contribution to model output total variability was dominated by three input factors, the number of analogs (NmAn), the search window size (SW) and the weighting coefficient related to predictor precipitation (C1). Their order of prioritization for CRPSS and low threshold BSS was SW, C1 and NmAn. The high threshold BSS lead-day 1

results changed with NmAn taking the highest position and C1 showing a marked decrease in Si, although it does not fall from its top three ranking. The shift of NmAn to the dominant factor also occurs in all lead-day 5 forecasts.

The increased importance of NmAn for leadDay 1 high threshold events is due to an inherent weakness in the AM. Large precipitation events occur less frequently than smaller ones, the result of which is that there are fewer good analog matches in a reforecast archive. While a 25mm event in a 24 hour period is not extreme, the relationship of storm size to available good analogs holds true. The AM has to suffice with some good matches and some of less quality. In these instances the size of the analog ensemble that can be used to produce the forecasts becomes more important due to the potential necessity of smoothing out bad analogs, thus increasing the models sensitivity to changes in NmAn.

All of the GSA results, with the exception of low precipitation threshold BSS lead-day 1, showed a seasonal influence on the importance of NmAn and SW with respect to their first order effects. In summer, NmAn's effect on output variability was substantially higher than in the winter and the reverse is true for SW. This seasonal influence could be due to the relative seasonal quality of the NWPM forecasts in the reforecast archive. Florida summers are well known for frequent convective precipitation events which by their spatial and temporally sporadic nature, as well as difficulty in modeling cloud formation lead to lower quality forecasts. In the winter dry season precipitation is driven by well-defined frontal events that are easier to predict. This leads to high forecast quality. When the analog method is presented with a pool of potential analogs of low quality, the number of analogs takes on a more significant role

due to the smoothing that naturally occurs when using an ensemble to produce a forecast. When the pool has high quality forecasts available, then the number of analogs is less important and obtaining the best match becomes more important. In this case the size of the search window which effectively changes the number of potential analogs in the pool becomes more important.

There is an opposing seasonal influence on SW, the winter increases SWs importance. The winter in Florida is the dry season where precipitation is infrequent and is propagated by frontal systems that are usually accompanied by predictable climatic signals. In this case the value of a good analog would have a higher impact and potentially higher skill. Basically the is a higher probability that observed precipitation from the date of matching reforecast climatic pattern would propagate in the same manner as the current forecast observation relationship. For that reason a higher value is placed on obtaining good analogs, thus the importance of the size of the group of potential forecasts and correspondingly the search window size.

The predictor coefficient input factors showed limited effect on variability with their first order sensitivity indices, in point of fact their boxplots only reached the subjective 5% threshold of importance three times. This would normally draw one to the conclusion that they should all be discarded. However, using this information alone could lead to reporting a false negative. When looking at the total effect sensitivity index, all five of the predictor coefficient ST_i means, with the exception of C3, fall within a range 0.05 to 0.1. A review of the interactive effects calculation (ST_i- S_i) for all factors from and model outputs, indicated that there are interactive effects present, with the majority falling within the range of 0-0.1. As a result of this information combined with

the comparison of the UA PDFs to the optimized single predictor results the recommendation is to keep all predictors with the exception of C3. The benefit may only be slightly higher than marginal, but for precipitation forecasting, any improvement is warranted.

Due to the identification of interactive effects the question arises as to whether the model is additive and appropriate for this GSA methodology. The answer is yes. All 72 of the individual model results had their input factor S_i summed and checked to see if they fell below the 0.6 standard for a “mostly additive” model. In only 6 instances did the sum fall below 0.7 and in only 2 did it fall below 0.6. This means the model can be considered additive.

Table 5-1. Summary of NA model output probability distributions.

Output	Statistic	Lead-Day 1				Lead-Day 3				Lead-Day 5			
		Dec	Jan	Jun	Jul	Dec	Jan	Jun	Jul	Dec	Jan	Jun	Jul
CRPSS	Mean	0.23	0.28	0.12	0.04	0.15	0.14	0.04	0.02	0.06	0.10	0.01	-0.01
	Median	0.24	0.28	0.12	0.04	0.15	0.15	0.04	0.02	0.07	0.10	0.01	-0.01
	2.50%	0.17	0.20	0.10	0.02	0.12	0.11	0.02	0.00	0.04	0.07	-0.02	-0.03
	97.50%	0.27	0.32	0.13	0.05	0.17	0.16	0.05	0.03	0.08	0.11	0.02	0.00
	95% CI	0.10	0.12	0.03	0.04	0.05	0.06	0.03	0.04	0.03	0.04	0.04	0.04
BSS 2.5mm Threshold	Mean	0.37	0.39	0.18	0.06	0.23	0.23	0.08	0.02	0.05	0.13	0.03	-0.01
	Median	0.38	0.40	0.18	0.06	0.23	0.24	0.09	0.02	0.05	0.14	0.04	0.00
	2.50%	0.27	0.31	0.15	0.03	0.18	0.19	0.06	0.00	0.03	0.11	0.00	-0.04
	97.50%	0.42	0.44	0.21	0.08	0.25	0.26	0.10	0.04	0.07	0.15	0.05	0.01
	95% CI	0.15	0.13	0.06	0.05	0.07	0.06	0.04	0.04	0.04	0.04	0.04	0.05
BSS 25mm Threshold	Mean	0.17	0.19	0.08	0.02	0.11	0.11	0.02	0.02	0.05	0.06	0.01	-0.01
	Median	0.17	0.20	0.09	0.03	0.11	0.11	0.03	0.02	0.05	0.06	0.01	-0.01
	2.50%	0.10	0.11	0.05	-0.01	0.08	0.07	0.00	0.00	0.03	0.03	-0.03	-0.05
	97.50%	0.23	0.23	0.11	0.04	0.15	0.14	0.04	0.04	0.07	0.09	0.02	0.00
	95% CI	0.13	0.12	0.06	0.05	0.08	0.07	0.04	0.04	0.04	0.06	0.05	0.05

Note: Data is arranged to show seasonal variation using 3 representative months for 3 lead-days.

Table 5-2. Summary of Interactive effects (ST_i - S_i.) for all model results used during GSA/UA.

Range	LD1	LD5	CRPSS	BSS 2.5mm	BSS 25mm	Winter	Summer	C1	C2	C3	C4	C5	C6	SW	NmAn
$0.0 \le (St_i - S_i) \le 0.1$	225	206	161	148	122	87	110	56	69	70	66	62	65	19	24
$0.1 < (St_i - S_i) \le 0.2$	45	66	26	39	46	28	29	14	3	2	6	10	7	35	34
$0.2 < (St_i - S_i) \le 0.3$	16	12	4	4	20	8	5	2	0	0	0	0	0	15	11
$0.3 < (St_i - S_i) \le 0.4$	2	4	1	0	5	5	0	0	0	0	0	0	0	2	4
Total	288	288	192	192	144	144	72	72	72	72	72	72	72	72	72

Note: The data is recorded as the number of occurrences within the specified ranges. The data is categorized by the type of output based on skill score, lead-day, season, and input factor. Lead-day results encompass all months, skill scores and input factors. Skill score results encompass all months, lead-days and input factors. Seasonal results encompass all skill scores, input factors and lead-days and 3 seasonal months. Input factor results encompass all skill scores, months, and lead-days.

Table 5-3. Sum of the first order effects sensitivity index S_i for use in identifying the models additive nature.

	CRPSS		BSS 2.5mm		BSS 25mm	
	Lead-Day 1	Lead-Day 5	Lead-Day 1	Lead-Day 5	Lead-Day 1	Lead-Day 5
Winter (mean)	0.95	0.73	0.89	0.88	0.79	0.70
Summer (mean)	0.91	0.75	0.86	0.70	0.77	0.87

Note: If the sum is equal to one the model is totally additive, and if it is greater than 0.6 the model is considered mostly additive.

Table 5-4. Monte-Carlo filtering threshold and ensemble size comparison for optimum skill score results identification.

		Percent Threshold			
		0.8%	0.7%	0.6%	0.5%
CRPSS	Summer	0.12 (2242)*	0.11 (3630)	0.05 (4968)	0.04 (6434)
	Winter	0.28 (2764)	0.27 (4147)	0.26 (5529)	0.26 (6909)
2.5mm	Summer	0.17 (2794)	0.16 (4124)	0.10 (5367)	0.09 (6668)
	Winter	0.41 (2763)	0.40 (4143)	0.40 (5535)	0.39 (6905)
25mm	Summer	0.08 (2469)	0.07 (3791)	0.03 (5093)	0.03 (6551)
	Winter	0.25 (2764)	0.24 (4149)	0.23 (5532)	0.22 (6917)

Note: The data is based on the best 20, 30, 40, and 50%. This data in combination with corresponding input factor MC filtering Smirnov tests is used to identify a percentile to use as the behavioral/non-behavioral split.

* First number is skill score and the number in parenthesis is the size of the optimum ensemble at the corresponding percent threshold

Table 5-5. Smirnov results comparing behavioral and non-behavioral input factor distributions.

Input	25mm BSS		2.5mm BSS		CRPSS	
Factor	H	p-value	H	p-value	H	p-value
C1*	0	2.1E-01	0	9.9E-01	0	1.9E-01
C2*	0	1.0E+00	0	7.8E-01	0	9.9E-01
C3*	0	1.0E+00	0	1.0E+00	0	1.0E+00
C2*	0	1.0E+00	0	9.7E-01	0	9.8E-01
C5*	0	4.2E-01	0	8.8E-01	0	7.8E-01
C6*	0	1.0E+00	0	9.5E-01	0	1.0E+00
SW*	*1*	4.9E-02	*1*	1.5E-06	*1*	1.5E-06
NnAm*	*1*	9.7E-13	*1*	2.9E-02	*1*	1.8E-13
C1**	*1*	2.6E-20	*1*	2.7E-33	*1*	2.5E-08
C2**	0	2.6E-01	*1*	2.0E-05	*1*	3.5E-02
C3**	0	1.4E-01	*1*	4.1E-02	0	2.2E-01
C4**	0	6.5E-02	*1*	2.6E-03	0	7.5E-01
C5**	*1*	5.5E-03	*1*	5.1E-02	0	8.6E-01
C6**	0	6.2E-01	*1*	1.0E-03	0	2.0E-01
SW**	*1*	7.12E-18	*1*	9.8E-13	*1*	1.0E-10
NnAm**	*1*	3.63E-19	*1*	6.2E-27	*1*	4.4E-08

Note: The logical identifier H identifies the acceptance or rejection of the null hypothesis where 1 indicates that the Null hypothesis is rejected and the distributions are statistically different at a confidence level of 0.05. Rejection occurs when the p-value is less than the confidence interval.

* Summer months Jun., Jul., and Aug

** Winter months Dec., Jan., and Feb.

Table 5-6. Smirnov results comparing behavioral input factor distributions for seasonal, event magnitude and lead-day propagation.

	BSS 2.5mm Seasonal*		BSS 2.5mm Seasonal		CRPSS Seasonal		Event Magnitude		Lead-day Propagation	
	H	p-value	H	p-value	H	p-value	H	p	H	p
C1	1	8.1E-03	1	1.12E-14	1	4.66E-05	1	5.4E-05	1	4.20E-02
C2	1	2.6E-03	0	0.7151836	0	9.88E-01	0	9.0E-01	0	3.50E-01
C3	0	9.7E-01	0	0.7466403	0	9.99E-01	0	1.6E-01	0	2.60E-01
C4	1	1.9E-14	0	0.7514072	0	0.0639441	0	9.5E-01	0	9.30E-02
C5	1	6.2E-02	1	3.66E-05	0	0.2895156	0	9.3E-01	0	6.80E-03
C6	1	1.8E-11	1	0.0240967	1	9.72E-08	0	8.6E-01	0	8.90E-02
SW	1	6.93E-09	1	7.12E-18	1	1.01E-10	1	1.72E-02	1	2.50E-03
Na	1	1.10E-24	1	3.63E-19	1	4.44E-08	1	3.93E-02	1	1.67E-02

Note: The logical identifier H identifies the acceptance or rejection of the null hypothesis where 1 indicates that the Null hypothesis is rejected and the distributions are statistically different at a confidence level of 0.05. Rejection occurs when the p-value is less than the confidence interval.

* Seasonal comparison is performed between Dec. Jan and Feb with Jun., Jul., and Aug.

** Event Magnitude compares 2.5mm BSS with 25mm BSS for all months

*** Lead-day propagation compares CRPSS lead-day 1 to CRPSS lead-day 5.

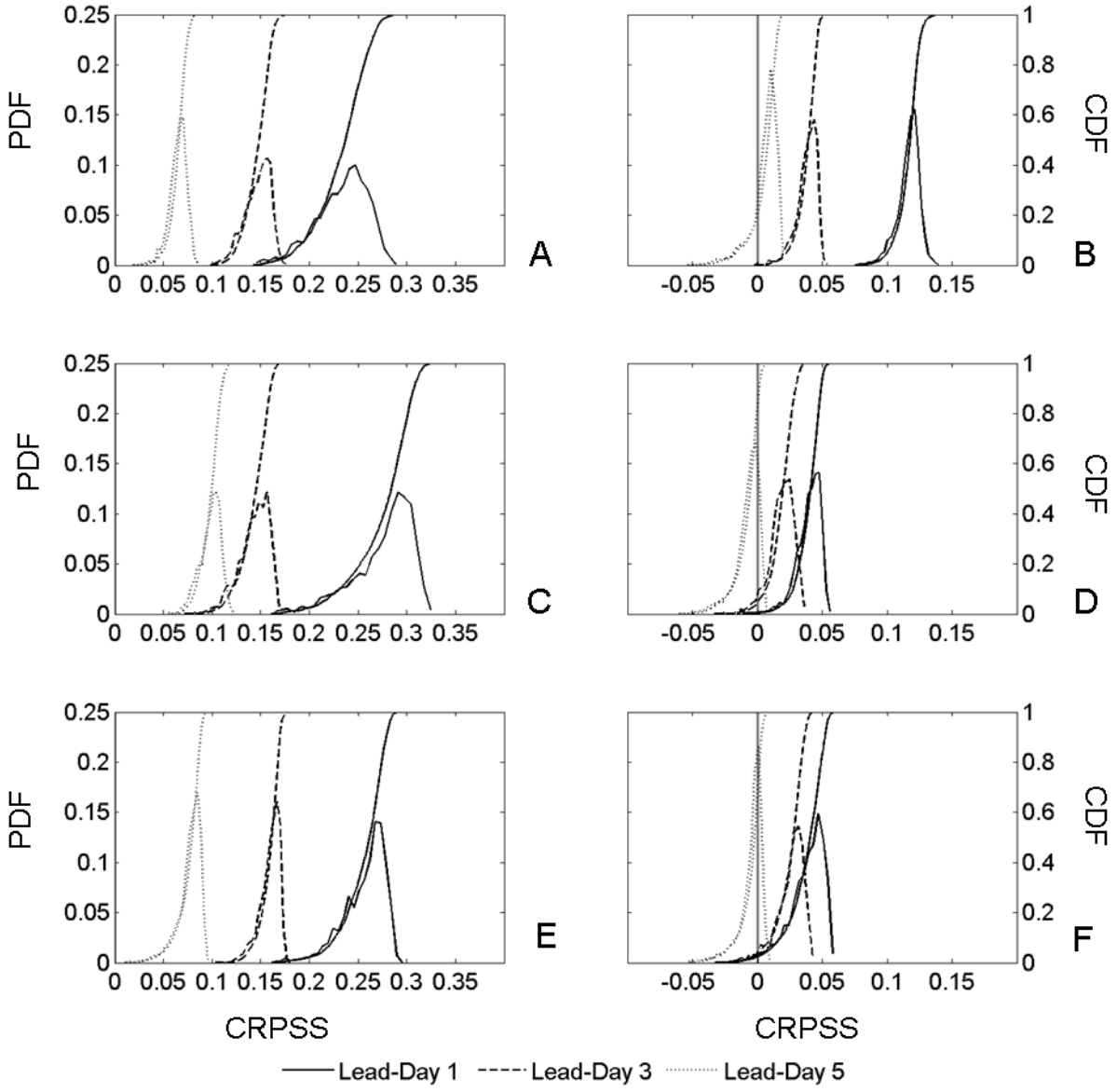


Figure 5-1. PDF and CDF comparative representation of winter and summer months CRPSS model output variability. Distributions created from 4608 GSA simulations. A) Dec. B) Jun. C) Jan. D) Jul. E) Feb. F) Aug.

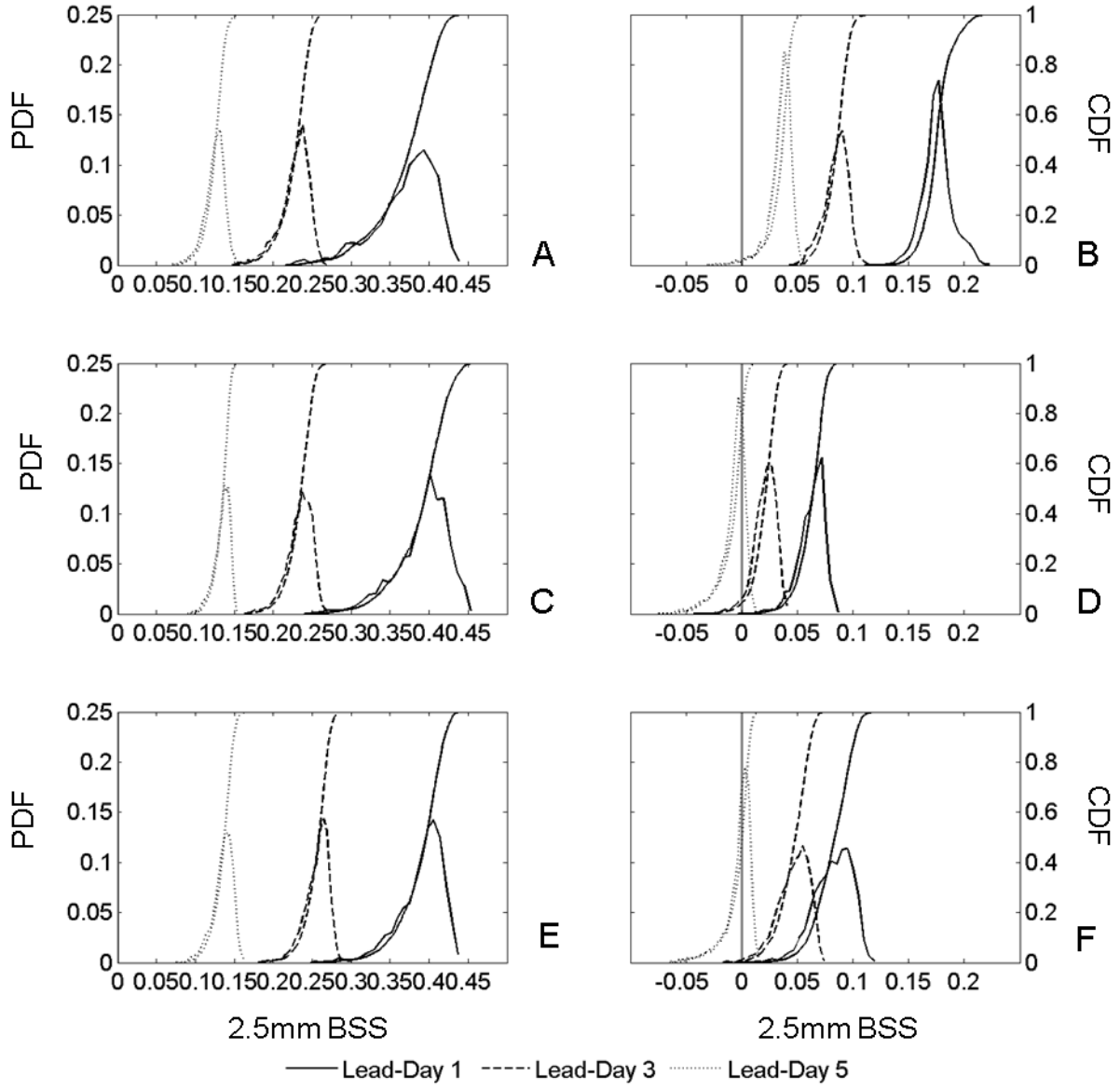


Figure 5-2. PDF and CDF comparative representation of winter and summer months 2.5mm BSS model output variability. Distributions created from 4608 GSA simulations. A) Dec. B) Jun. C) Jan. D) Jul. E) Feb. F) Aug.

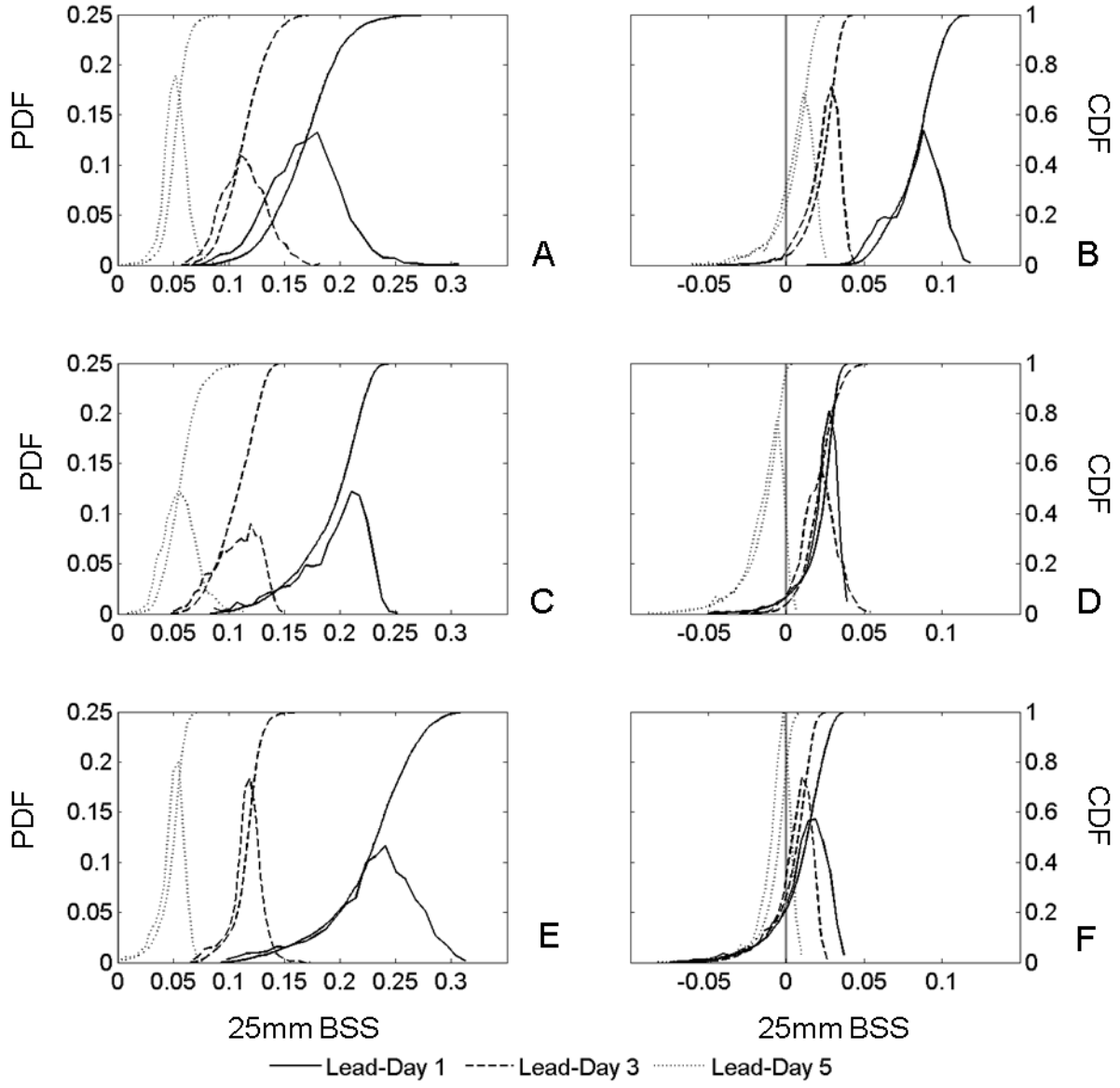
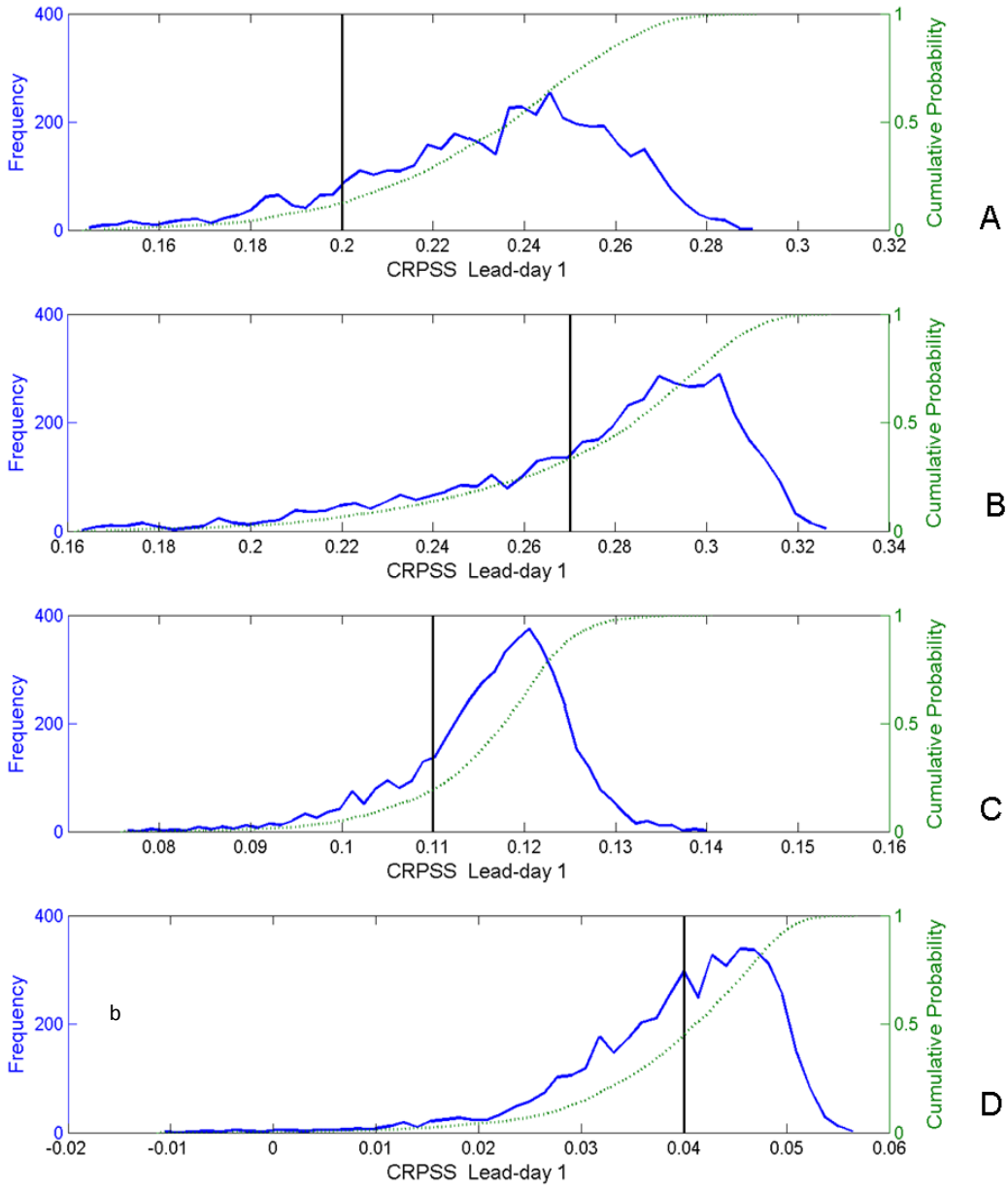


Figure 5-3. PDF and CDF comparative representation of winter and summer months 25mm BSS model output variability. Distributions created from 4608 GSA simulations. A) Dec. B) Jun. C) Jan. D) Jul. E) Feb. F) Aug.



.

Figure 5-4. Comparison of single predictor baseline results to multi-predictor probabilistic lead-day 1 CRPSS results (PDF and CDF). Data is based on GSA simulations for 2 summer and 2 winter months. The vertical line represents a deterministic forecast produced by "optimized" NA input factors. PDF and CDF created from 4608 simulations. A) December. B) January. C) June. D) July.

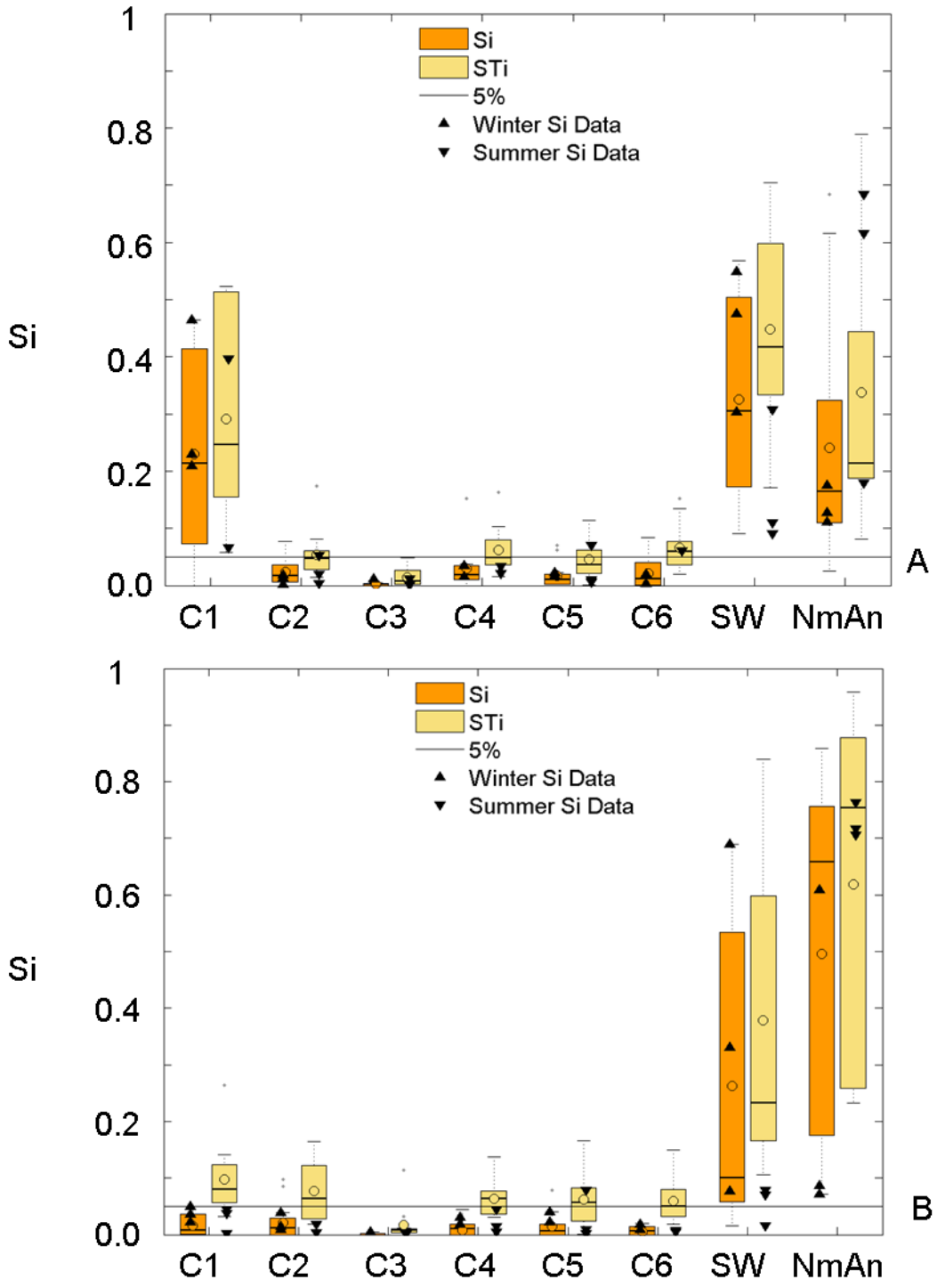


Figure 5-5. Sensitivity index ensembles (Si and STi) using CRPSS model output. The boxplots were assembled from all monthly Si evaluations. The box indicates the 25, 50 and 75% quartiles while the whiskers are 5 and 95 percentile. Δ = Dec., Jan., and Feb. ∇ = Jun., Jul., and Aug. ° = median. A) Lead-day 1. B) Lead-day5.

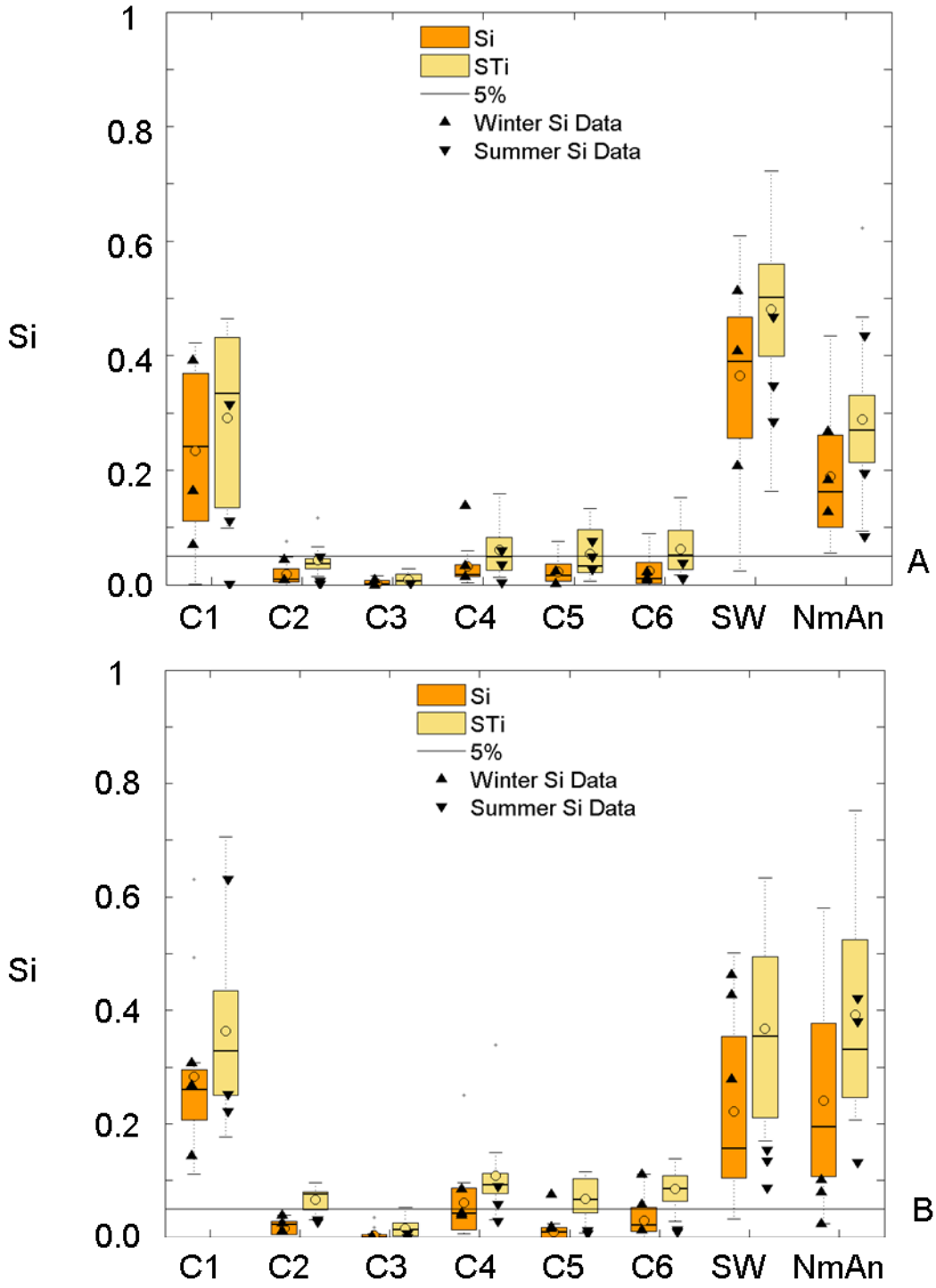


Figure 5-6. Sensitivity index (Si and STi) ensemble boxplots using 2.5mm threshold BSS model output. The boxplots were assembled from all monthly Si evaluations Boxplots indicate the 25, 50 and 75% quartiles, while the whiskers are 5 and 95 percentile. Δ = Dec., Jan., and Feb. ∇ = Jun., Jul., and Aug. ° = median. A) Lead-day 1. B) Lead-day 5.

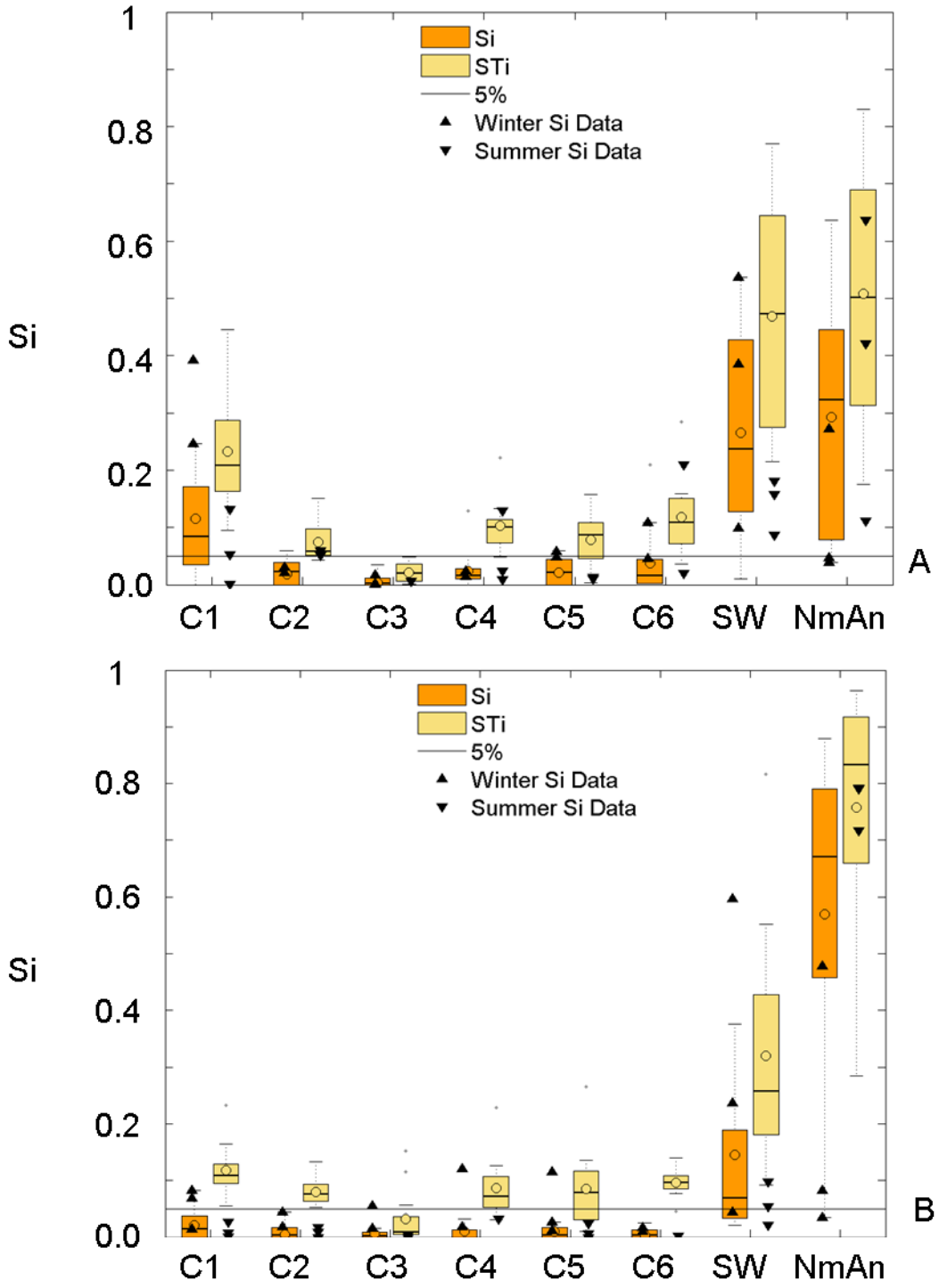


Figure 5-7. Sensitivity index (Si and STi) ensemble boxplots using 25mm threshold BSS model output. The boxplots were assembled from all monthly Si evaluations. The box indicates the 25, 50 and 75% quartiles while the whiskers are the 5th and 95th percentiles. Δ = Dec., Jan., and Feb ∇ = Jun., Jul., and Aug. ° = median. A) Lead-day 1. B) Lead-day 5.

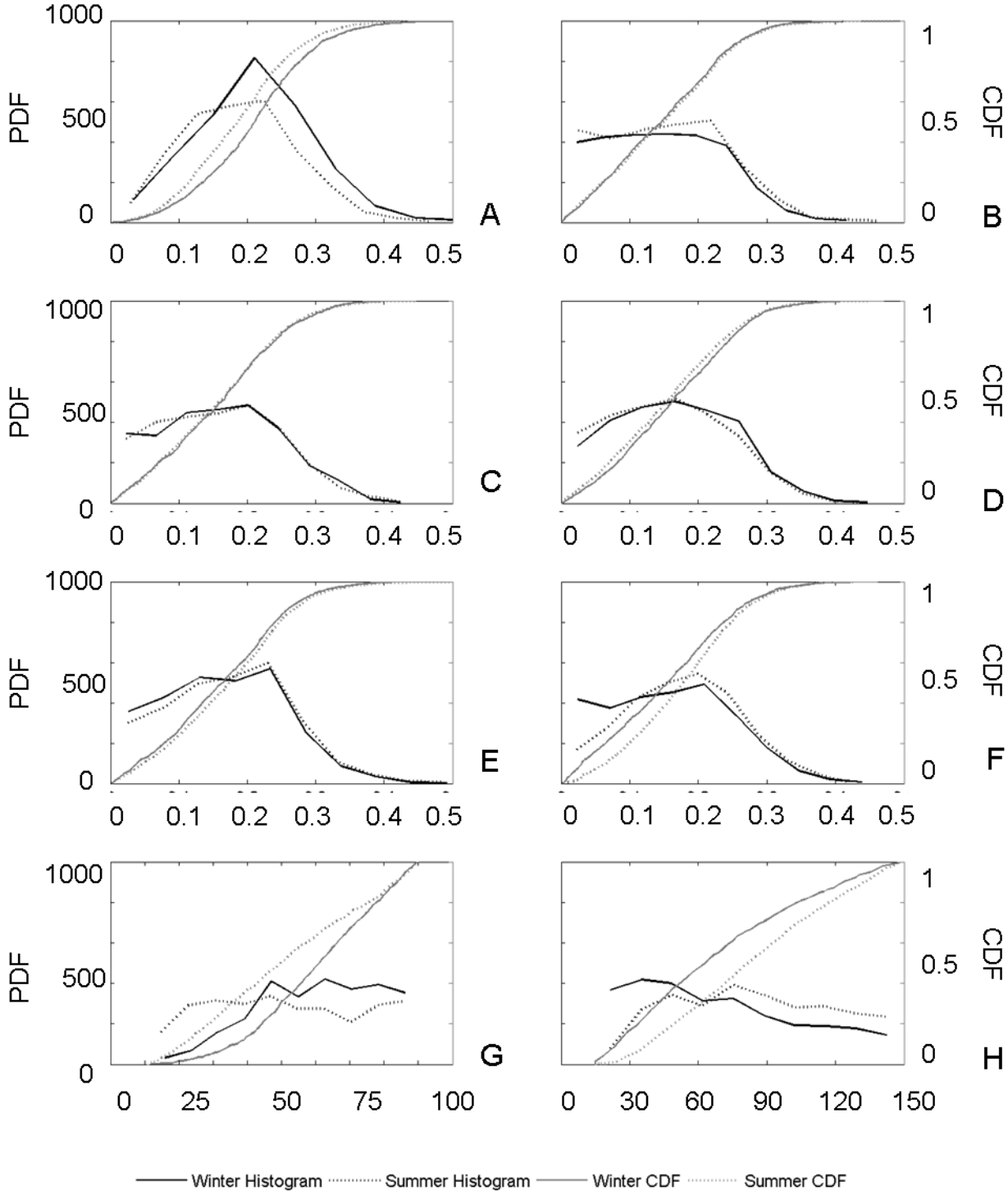


Figure 5-8. Comparison of optimum (behavioral) input factor distributions for summer and winter CRPSS model results. Both the frequency and CDFs are presented. A) C1. B) C2. C) C3. D) C4. E) C5. F) C6. G) SW. and H) NmAn.

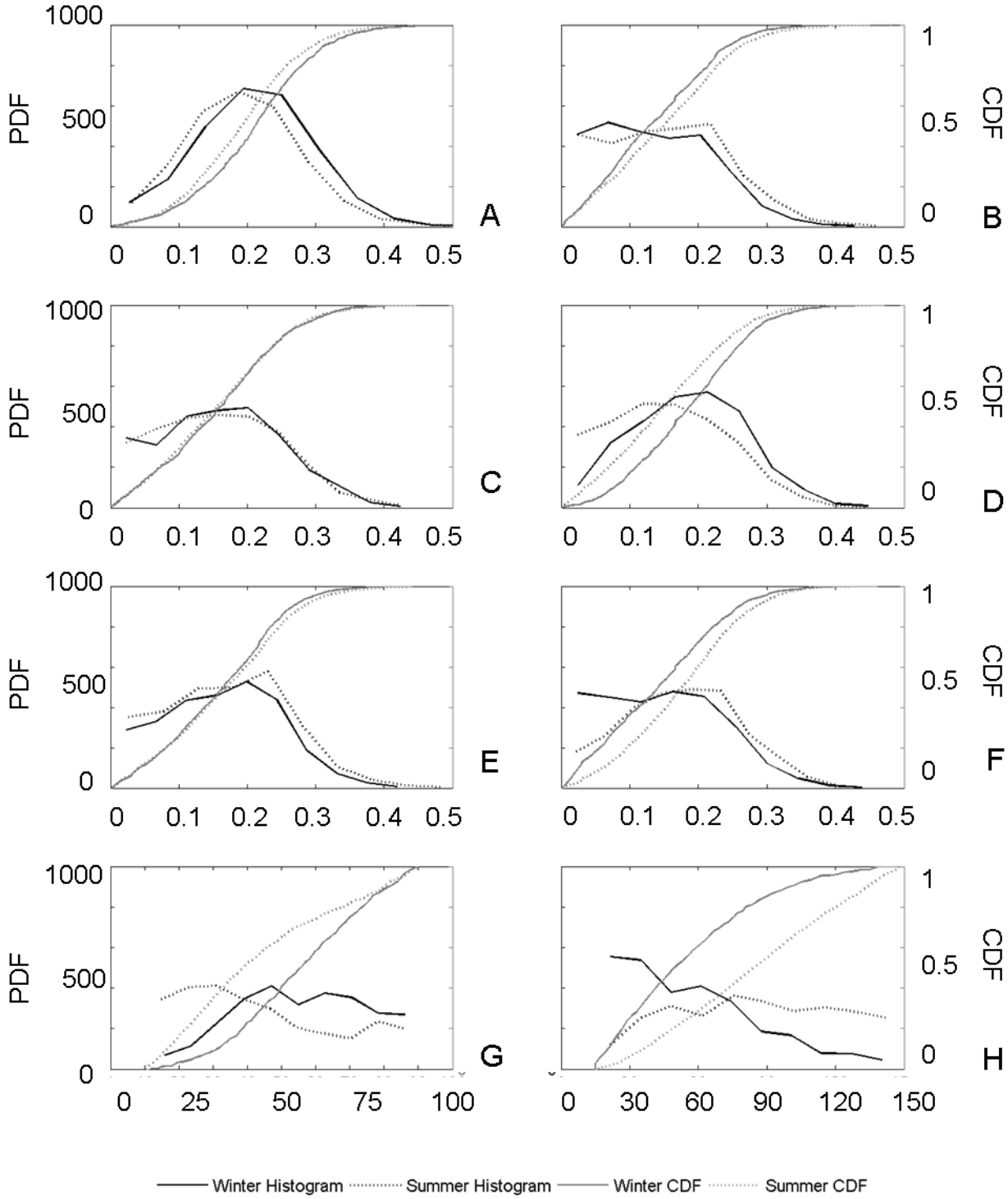


Figure 5-9. Comparison of optimum (behavioral) input factor distributions for summer and winter 2.5mm BSS model results. Both the frequency and CDFs are presented. A) C1. B) C2. C) C3. D) C4. E) C5. F) C6. G) SW. H) NmAn.

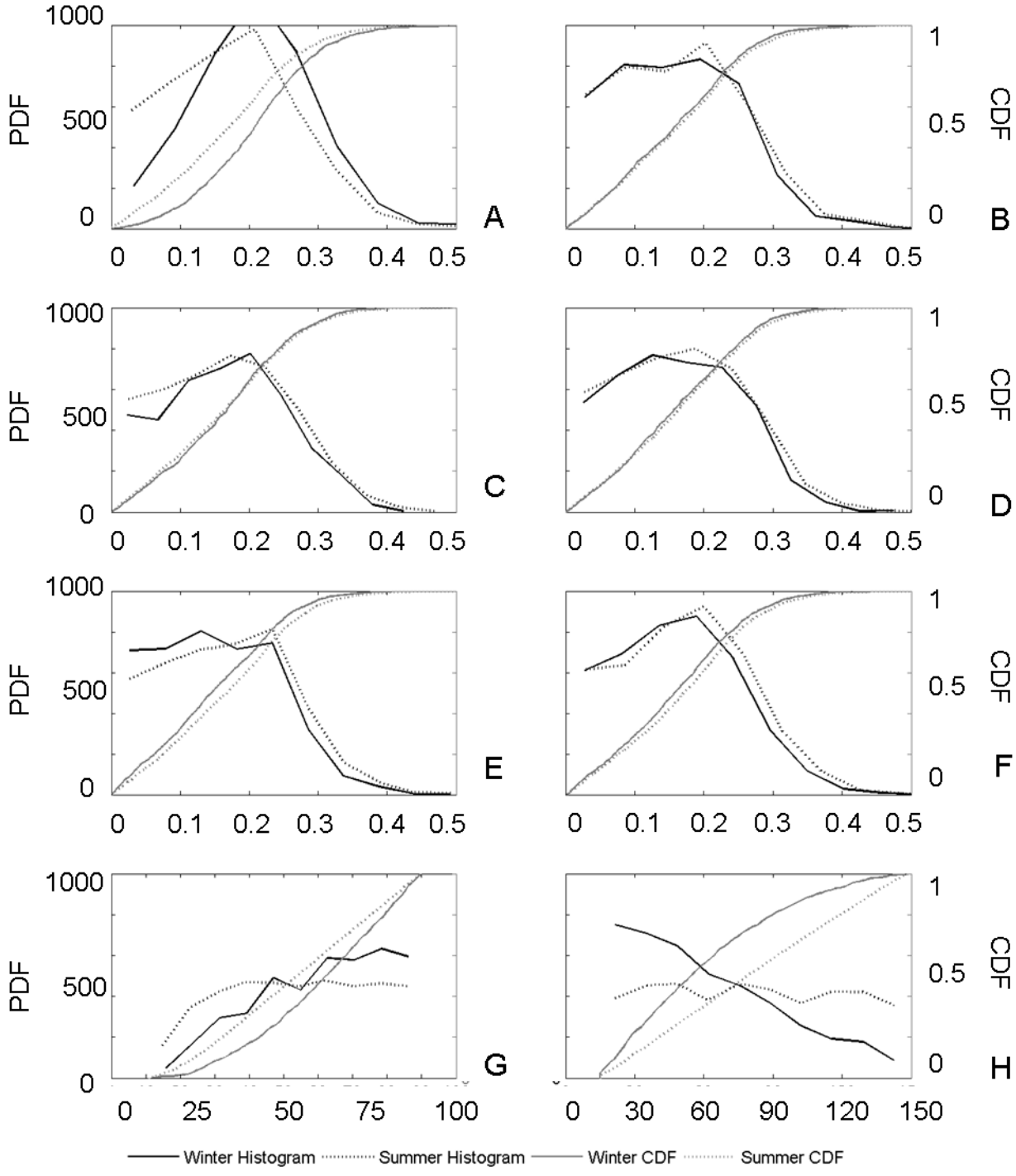


Figure 5-10. Comparison of optimum (behavioral) input factor distributions for summer and winter 25mm BSS model results. Both the frequency and CDFs are presented. A) C1. B) C2. C) C3. D) C4. E) C5. F) C6. G) SW. H) NmAn.

CHAPTER 6
CONCLUSIONS

This study evaluates the use of Analog methodologies for use in downscaling and bias correcting Numerical Weather Prediction Models to improve short term precipitation forecasts in Tampa Bay Florida. The study attempts to identify unique input settings as well as seasonal variability of the methods based on the noted history of difficulties predicting precipitation given Florida's unique geographic and atmospheric conditions. The study initiates using a single predictor approach; precipitation used as the predictor for analog pattern matching and precipitation as the forecasted variable, and expands to multiple predictor algorithms. Global Sensitivity and Uncertainty Analysis techniques are used to evaluate the potential improvement using the multiple predictors in an attempt to; evaluate potential for analog method (AM) improvement, identify input setting seasonal variation, and find physically based climatological links in the statistically and data driven AM. This has not been systematically performed previously. The final aspect of this study is implementation of the Analog method into Tampa Bay Waters proprietary artificial neural network stream flow forecasting model.

The process of achieving the aforementioned goals is documented in the following four chapters; Chapter 2-Evaluation of Analog Methodologies for Use as Downscaling and Bias Correcting Precipitation Forecasts for Florida's Unique Conditions; Chapter 3-Implementation of AM Precipitation Forecasts to Tampa Bay Water's Proprietary Artificial Neural Network Stream-flow Model; Chapter 4-Exploring the Viability of Using Multiple Predictors for Analog Selection Using a Global Sensitivity Screening Method; and Chapter 5-Quantitative Global Sensitivity and Uncertainty

Analysis of the Analog Method to Assess the Value of Multiple Predictor Fields as a Function of Lead-time; Seasonal Variability and the Magnitude of Forecasted Precipitation Events. A summary of the key conclusions of each section is listed below:

Evaluation of single predictor Natural and Constructed Analogs.

- Use of the NA method in Florida, while showing skill with seasonal variability, does not perform at optimum potential using cited input settings prescribed for the Continental United States
- Both the single predictor NA and CA methods show skill in precipitation forecasting with noticeable reduction in skill for longer lead times, precipitation event magnitude, and the summer season
- While CA and NA showed similar forecasting quality (assessed by RMSE, MBE, and MSESS) for lead-day 1 forecasts, NAs outperformed CA at longer lead times (lead-day 5).

Implementation in the SWANN model.

- The SWANN model is incapable of forecasting for low flow streams
- CA and NA produced better results than using the original Knn method
- LI outperformed CA and NA in terms of Bias Error and RMSE for the high flow-rate stream flow model outputs.
- The NA using the mean (as opposed to 125 individual members), had a bias error of approximately double all of the other methods.
- The NA and CA methods showed the highest resolution for low flow conditions with LI having the best for high flow conditions
- Improvement in all instances were marginal
- Based on a simple sensitivity test and subsequent reevaluation of the code, the training and validation of the SWANN model was performed using persistence forecasts (previous week) and not knn.

Qualitative input factor global screening sensitivity analysis for use with multiple predictors.

- The input factors used in the single predictor methods, precipitation, search window (SW) and number of analogs (NmAn) showed the most influence on output variability under all conditions.

- Seasonal variability in SW and NmAn importance has been identified with more impact from SW in winter and more impact from NmAn in summer under all conditions. RHUM also shows seasonal variability for LD5 forecasts with higher effects in Winter

- NmAn shows lead day variability with increased effect at longer led times.

- The predictor coefficient related to mean sea level pressure consistently ranked last in importance, however fell below the threshold for removal only in a small fraction of output result

- An assessment of 4 sample generation schemes produced similar results indicating proof of concept

Quantitative GSA/UA on the multi-predictor AM.

- Uncertainty analysis showed that the optimized NA method produced in section 1 fell below the peak of the Monte-Carlo output PDF. This result is indicative that either the use of multiple predictors increase the potential skill of the methodology, or the coarse optimization of the NA method was not rigorous enough to achieve true optimum values.

- Relative seasonal, lead-day and precipitation event magnitude input factor variability identified by Morris, were validated with identification and quantification using the Sobol method.

- With the exception of precipitation the predictor coefficients showed low levels of influence based on their first order effect, however the index for the total effects (includes interactive) increase to the 0.05-0.15 range (scale goes from 0-1)

- The north south wind vector V10m indicates higher effects in winter than in summer months.

- Although Interactive effects were identified in the model, it did not exceed the threshold for a “mostly additive model”.

- The PRMSL predictor consistently ranked last and had values very near to 0.

In summary, the NA and the CA are effective downscaling and bias correction methodologies for use in Florida. The precipitation forecasting skill achieved, shows seasonal, event threshold, and lead-day variability with declining scores in summer, at higher magnitude precipitation events, and longer lead times. Implementation of AM forecasts into SWANN indicated that NA and CA improved stream-flow forecast skill,

although incremental. Interestingly for LI method also performed well and had the lowest MBE. This was interesting in that this technique is the only one that does not use observed precipitation and therefore is not bias corrected. In the process of evaluating these unexpected results it was identified that the SWANN model, contrary to literature (Asefa, 2009), used persistence precipitation forecasts (previous weeks values) rather than Knn forecasts for training and operation. Therefore the model, when looking for data based relationships during training, identified and incorporated persistence as the most important link. Therefore it is surprising that CA and NA improved the results at all. In retrospect the choice of model chosen for stream-flow evaluation should have underwent sensitivity analysis to ascertain the potential of altering the precipitation forecast input methodology.

GSA/UA and MC filtering indicated that different input settings should be used depending on the conditions (i.e. seasonality, event magnitude and lead-day propagation) in order to optimize skill. GSA also indicated that when conditions exist that increase forecasting difficulty (longer lead time, Florida summer, higher precipitation events) then the quality of the analogs and their potential connection the observed precipitation is decreased, this leads to the importance of the NmAn to increase. The more members you have in an ensemble of questionable value, the more smoothing occurs in the averaging or probabilistic calculations of forecasts. Conversely in conditions of improved analog member quality, it is more important to identify and select the best analogs available. In these cases the search window size, which increases the pool of potential analogs and increases the availability of quality analogs, has the most effect on output variability.

The UA results were indicative of the value of using multiple predictors, by showing the optimized NA forecast skill fell below the peak of the uncertainty PDF. Although this result could be a function of the crude OAT optimization approach used for the direct NA, GSA reinforced the conclusion. GSA indicated that the predictors first order sensitivity effects, with the exception of precipitation, were low (below 5%), the total effects were not marginal. This is due to interactive effects. The predictor PRMSL was the only one that showed marginal effect and should be removed.

BIBLIOGRAPHY

Asefa, T., 2009. Ensemble Streamflow Forecast: A GLUE-Based Neural Network Approach1. J Am Water Resour As, 45(5): 1155-1163.

Astatkie, T., Watt, W.E., 1998. Multiple-input transfer function modeling of daily streamflow series using nonlinear inputs. Water Resources Research, 34(10): 2717-2725.

Cacuci, D.G., Ionescu-Bujor, M., Navon, I.M., 2003. Sensitivity and uncertainty analysis. Chapman & Hall/CRC Press, Boca Raton, v. <1-2> pp.

Callahan, B., Miles, E., Fluharty, D., 1999. Policy implications of climate forecasts for water resources management in the Pacific Northwest. Policy Sci, 32(3): 269-293.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software, 22(10): 1509-1518.

Campolongo, F., Cariboni, J., Saltelli, A., Schoutens, W., 2005. Enhancing the Morris method, Proc. 4th International Conference on Sensitivity Analysis of Model Output (SAMO 2004), pp. 369-379.

Campolongo, F., Saltelli, A., Jensen, N.R., Wilson, J., Hjorth, J., 1999. The role of multiphase chemistry in the oxidation of dimethylsulphide (DMS). A latitude dependent analysis. J Atmos Chem, 32(3): 327-356.

Cannas, B., Fanni, A., See, L., Sias, G., 2006. Data preprocessing for river flow forecasting using neural networks: Wavelet transforms and data partitioning. Phys Chem Earth, 31(18): 1164-1171.

Clark, M.P. et al., 2004. A resampling procedure for generating conditioned daily weather sequences. Water Resources Research, 40(4).

Crosetto, M., Tarantola, S., 2001. Uncertainty and sensitivity analysis: tools for GIS-based model implementation. Int J Geogr Inf Sci, 15(5): 415-437.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients .1. Theory. J Chem Phys, 59(8): 3873-3878.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Prog Phys Geog, 25(1): 80-108.

Enke, W., Spekat, A., 1997. Downscaling climate model outputs into local and regional weather elements by classification and regression. Clim Res, 8(3): 195-207.

Giorgi, F. et al., 2001. Emerging patterns of simulated regional climatic changes for the 21st century due to anthropogenic forcings. Geophys Res Lett, 28(17): 3317-3320.

Govindaraju, R.S., Artific, A.T.C.A., 2000a. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. J Hydrol Eng, 5(2): 115-123.

Govindaraju, R.S., Artific, A.T.C.A., 2000b. Artificial Neural Networks in Hydrology. Ii: Hydrologic Applications. J Hydrol Eng, 5(2): 124-137.

Haan, C.T., Allred, B., Storm, D.E., Sabbagh, G.J., Prabhu, S., 1995. Statistical Procedure for Evaluating Hydrologic Water-Quality Models. T Asae, 38(3): 725-733.

Hagedorn, R., Hamill, T.M., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Monthly Weather Review, 136(7): 2608-2619.

Hamill, T.M., Colucci, S.J., 1998. Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. Monthly Weather Review, 126(3): 711-724.

Hamill, T.M., Hagedorn, R., Whitaker, J.S., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. Monthly Weather Review, 136(7): 2620-2632.

Hamill, T.M., Whitaker, J.S., 2006a. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Monthly Weather Review, 134(11): 3209-3229.

Hamill, T.M., Whitaker, J.S., 2007. Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. Monthly Weather Review, 135(9): 3273-3280.

Hamill, T.M., Whitaker, J.S., Mullen, S.L., 2006b. Reforecasts: An Important Dataset for Improving Weather Predictions. Bulletin of the American Meteorological Society, 87(1): 33-46.

Hamill, T.M., Whitaker, J.S., Wei, X., 2004. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. Monthly Weather Review, 132(6): 1434-1447.

Hartmann, H.C., Bales, R., Sorooshian, S., 2002. Weather, climate, and hydrologic forecasting for the US Southwest: a survey. Clim Res, 21(3): 239-258.

He, X., Asada, H., 1993. A new method for identifying orders of input-output models for nonlinear dynamic systems, American Control Conference, 1993. IEEE, pp. 2520-2523.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5): 559-570.

Hidalgo, H.G., Dettinger, M.D., Cayan, D.R., 2008. Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States. California Climate Change Center: 48.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety, 52(1): 1-17.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast verification : a practitioner's guide in atmospheric science. J. Wiley, Chichester, West Sussex, England ; Hoboken, NJ, xiii, 240 p. pp.

Jones, C., Waliser, D.E., Lau, K.M., Stern, W., 2004. Global occurrences of extreme precipitation and the Madden-Julian oscillation: Observations and predictability. Journal of Climate, 17(23): 4575-4589.

Jung, T., Vitart, F., 2006. Short-range and medium-range weather forecasting in the extratropics during wintertime with and without an interactive ocean. Monthly Weather Review, 134(7): 1972-1986.

Juras, J., 2000. Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system". Weather and Forecasting, 15(3): 365-366.

Lemos, M.C., Finan, T.J., Fox, R.W., Nelson, D.R., Tucker, J., 2002. The use of seasonal climate forecasting in policymaking: Lessons from Northeast Brazil. Climatic Change, 55(4): 479-507.

Lilburne, L., Tarantola, S., 2009. Sensitivity analysis of spatial models. Int J Geogr Inf Sci, 23(2): 151-168.

Lins, H.F., Wolock, D.M., McCabe, G.J., 1997. Scale and modeling issues in water resources planning. Climatic Change, 37(1): 63-88.

Lorenz, E.N., 1969. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. J Atmos Sci, 26(4): 636-&.

Lynch, P., 2006. The emergence of numerical weather prediction : Richardson's dream. Cambridge University Press, Cambridge, UK ; New York, xii, 279 p. pp.

Lynch, P., 2008. The origins of computer weather prediction and climate modeling. J Comput Phys, 227(7): 3431-3444.

Matulla, C. et al., 2008. Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. Clim Dynam, 30(2-3): 133-144.

Maurer, E.P., Hidalgo, H.G., 2008. Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrol Earth Syst Sc, 12(2): 551-563.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4): 115-133.

Morris, M.D., 1991. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics, 33(2): 161-174.

Murphy, J., 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. Journal of Climate, 12(8): 2256-2284.

Nicholls, N., 1999. Cognitive illusions, heuristics, and climate prediction. Bulletin of the American Meteorological Society, 80(7): 1385-1397.

Pagano, T.C., Hartmann, H.C., Sorooshian, S., 2002. Factors affecting seasonal forecast use in Arizona water management: a case study of the 1997-98 El Nino. Clim Res, 21(3): 259-269.

Phillips, N.A., 1956. The General Circulation of the Atmosphere - a Numerical Experiment. Quarterly Journal of the Royal Meteorological Society, 82(352): 123-164.

Rayner, S., Lach, D., Ingram, H., 2005. Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. Climatic Change, 69(2-3): 197-227.

Reusser, D.E., Buytaert, W., Zehe, E., 2011. Temporal dynamics of model parameter sensitivity for computationally expensive models with the Fourier amplitude sensitivity test. Water Resources Research, 47.

Robertson, A.W., Qian, J.H., Tippett, M.K., Moron, V., Lucero, A., 2012. Downscaling of Seasonal Rainfall over the Philippines: Dynamical versus Statistical Approaches. Monthly Weather Review, 140(4): 1204-1218.

Ruano, M., Ribes, J., Seco, A., Ferrer, J., 2012. An improved sampling strategy based on trajectory design for application of the Morris method to systems with many input factors. Environmental Modelling & Software.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput Phys Commun, 145(2): 280-297.

Saltelli, A., 2004. Sensitivity analysis in practice : a guide to assessing scientific models. Wiley, Hoboken, NJ, xi, 219 p. pp.

Saltelli, A., 2008. Global sensitivity analysis : the primer. John Wiley, Chichester, England ; Hoboken, NJ, x, 292 p. pp.

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity analysis. Wiley series in probability and statistics. Wiley, Chichester ; New York, xv, 475 p. pp.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., Commission, E., 2006. Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety, 91(10-11): 1109-1125.

Saltelli, A., Tarantola, S., Chan, K.P.S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics, 41(1): 39-56.

Schmeits, M.J., Kok, K.J., 2010. A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts. Monthly Weather Review, 138(11): 4199-4211.

Schoof, J.T. et al., 2009. Dynamically and statistically downscaled seasonal temperature and precipitation hindcast ensembles for the southeastern USA. Int J Climatol, 29(2): 243-257.

Sobol, I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 7(4): 784-802.

Sobol, I.M., 1993. Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments, 1(4): 407-414.

Tian, D., Martinez, C.J., 2012. Comparison of two analog-based downscaling methods for regional reference evapotranspiration forecasts. Journal of Hydrology, 475: 350-364.

Van den Dool, H., 1994. Searching for analogues, how long must we wait? Tellus A, 46(3): 314-324.

Van den Dool, H., Huang, J., Fan, Y., 2003. Performance and analysis of the constructed analogue method applied to US soil moisture over 1981-2001. J Geophys Res-Atmos, 108(D16).

Van den Dool, H.M., 1989. A New Look at Weather Forecasting through Analogs. Monthly Weather Review, 117(10): 2230-2247.

Vasiliades, L., Loukas, A., Patsonas, G., 2009. Evaluation of a statistical downscaling procedure for the estimation of climate change impacts on droughts. Nat Hazard Earth Sys, 9(3): 879-894.

Vidale, P.L., Luthi, D., Frei, C., Seneviratne, S.I., Schar, C., 2003. Predictability and uncertainty in a regional climate model. J Geophys Res-Atmos, 108(D18).

Voisin, N., Pappenberger, F., Lettenmaier, D.P., Buizza, R., Schaake, J.C., 2011. Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. Weather and Forecasting, 26(4): 425-446.

Vonstorch, H., Zorita, E., Cubasch, U., 1993. Downscaling of Global Climate-Change Estimates to Regional Scales - an Application to Iberian Rainfall in Wintertime. Journal of Climate, 6(6): 1161-1171.

Wang, E.L., Zhang, Y.Q., Luo, J.M., Chiew, F.H.S., Wang, Q.J., 2011. Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data. Water Resources Research, 47.

Whitaker, J.S., Wei, X., Vitart, F., 2006. Improving week-2 forecasts with multimodel reforecast ensembles. Monthly Weather Review, 134(8): 2279-2284.

Wilby, R.L., Wigley, T., 1997. Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geog, 21(4): 530-548.

Wilks, D.S., 2000. On interpretation of probabilistic climate forecasts. Journal of Climate, 13(11): 1965-1971.

Wilks, D.S., 2006. Comparison of ensemble-MOS methods in the Lorenz '96 setting. Meteorological Applications, 13(03): 243.

Wilks, D.S., Hamill, T.M., 2007. Comparison of Ensemble-MOS Methods Using GFS Reforecasts. Monthly Weather Review, 135(6): 2379-2390.

Wood, A.W., Leung, L.R., Sridhar, V., Lettenmaier, D.P., 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Climatic Change, 62(1-3): 189-216.

Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. Journal of Hydrology, 214(1-4): 32-48.

Zhu, Y.H., Day, R.L., 2009. Regression modeling of streamflow, baseflow, and runoff using geographic information systems. J Environ Manage, 90(2): 946-953.

Zorita, E., Hughes, J.P., Lettemaier, D.P., Vonstorch, H., 1995. Stochastic Characterization of Regional Circulation Patterns for Climate Model Diagnosis and Estimation of Local Precipitation. Journal of Climate, 8(5): 1023-1042.

Zorita, E., von Storch, H., 1999. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. Journal of Climate, 12(8): 2474-2489.

BIOGRAPHICAL SKETCH

After finishing his tour of duty in the Air Force, Robert Rooney received a bachelor's degree from the Chemical Engineering Department at Arizona State University. Subsequently, Robert received a master's degree from University of Florida's Department of Environmental Engineering while working as a research assistant. The final leg of the academic journey came as a research assistant in the Agricultural and Biological Engineering Department at the University of Florida where he earned his doctoral degree in engineering, specializing in hydraulic modeling. The final degree was conferred in August 2013.