THESIS

FORECASTING OF ATLANTIC TROPICAL CYCLONES USING A KILO-MEMBER ENSEMBLE

Submitted by

Jonathan L. Vigh

Department of Atmospheric Science

In partial fulfillment of the requirements

for the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2004

COLORADO STATE UNIVERSITY

April 16, 2004

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY JONATHAN L. VIGH ENTITLED FORECASTING OF ATLANTIC TROPICAL CYCLONES USING A KILO-MEMBER ENSEMBLE BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

Advisor

Department Head

ABSTRACT OF THESIS

FORECASTING OF ATLANTIC TROPICAL CYCLONES USING A KILO-MEMBER ENSEMBLE

The past 30 years have witnessed steady improvements in the skill of tropical cyclone track forecasts. These increases have been largely driven by improved numerical weather prediction models and increased surveillance of the storm environment through aircraft reconnaissance and satellite remote sensing. The skill of deterministic track forecasts from full-physics models is gradually approaching the theoretical limit of predictability that arises due to the atmosphere's chaotic nature and limitations in determining the initial state. To make further progress, it is necessary to treat the uncertainty of the initial condition. One practical approach is to sample this uncertainty by perturbing the initial state. The resulting suite of forecasts that result from integrating such perturbations is known as an ensemble.

This thesis describes the design, implementation, and evaluation of a semi-operational ensemble forecasting system using an efficient multigrid barotropic vorticity equation model (MBAR). Five perturbation classes are used to simulate uncertainties in the storm environment and vortex structure. Uncertainties in the storm environment are simulated by using the background environmental flow evolutions provided by the NCEP Global Forecasting System (GFS) ensemble forecasts. Several deep layer-mean wind averages account for uncertainty in the depth of the storm steering layer. Uncertainties in the decomposition of the tropical atmosphere's vertical modes are simulated by varying the model equivalent phase speed. Finally, uncertainties in the vortex structure are simulated by varying the vortex

size and storm motion vector. Each perturbation in a given class is cross-multiplied with all other perturbations of other classes to obtain an ensemble with 1980 members. One of the fundamental questions addressed by this research is whether such cross-multiplication increases the degrees of freedom in the ensemble.

The ensemble is run for 294 cases from the 2001-2003 Atlantic hurricane seasons. Theory dictates that a properly-perturbed ensemble should, on average, be more accurate than any single ensemble member, but it was found that the kilo-ensemble mean forecast did not demonstrate substantial improvement over the control forecast. However, the ensemble mean did show substantial skill relative to the five-day climatology and persistence model (CLP5) throughout the 120-h forecast period. The ensemble mean spread (the mean distance of the individual members from the ensemble mean), x-bias, and y-bias statistics are also evaluated.

Probabilistic interpretations are valid with an ensemble of this size, so cumulative strike probabilities are calculated explicitly from the kilo-ensemble output. In a related possibilistic interpretation, the ensemble can be looked upon as mapping out the subspace of all possible storm tracks, so the reliability of this ensemble envelope is examined. Finally, if the ensemble can accurately simulate the uncertainties in the dynamical system, then there should be a positive relationship between ensemble mean spread and the error of the ensemble mean forecast. A strong relationship allows useful forecasts of forecast skill to be made at the time of the forecast. The kilo-member ensemble was found to have a weak spread-error relationship that peaks at 60 h.

Jonathan L. Vigh
Department of Atmospheric Science
Colorado State University
Fort Collins, Colorado 80523-1371
Summer 2004

ACKNOWLEDGEMENTS

This work no doubt has many faults, for which I take full responsibility. But if there is anything worthwhile in this work, I attribute it to the following individuals:

I am deeply grateful for the valuable guidance of my adviser, Dr. Wayne Schubert. He has shown great patience and given wide-ranging freedom in choosing and researching this topic. His enthusiasm for dynamics has taught me to see the richness and beauty of our atmosphere.

I also thank my committee members, Dr. Mark DeMaria, Dr. William Gray, and Dr. Gerald Taylor, for their valuable discussions and helpful suggestions. Special thanks goes to Dr. Mark DeMaria for providing the original idea for this research, for providing the LBAR model, and for spending many hours conversing with me about barotropic modeling and the ensemble design.

I am also grateful to Dr. Scott Fulton for the use of the MUDBAR model. He provided helpful guidance and made modifications that allow MUDBAR to use background wind fields. I would also like to thank the members of the Schubert Research Group, especially Rick Taft and Brian McNoldy, for patiently answering my questions and providing encouragement and humor. I acknowledge and thank the National Centers for Environmental Prediction for the use of the Global Forecasting System model data and the National Hurricane Center for the operational guidance and best track files. I am grateful to Mary Haley and the NCAR Command Language (NCL) developers for helping me to unlock the power of that data processing and graphics language. I am deeply appreciative of the love

and support of my friends, family, and church. And last but not least in any way, I thank God for the guidance and grace He has given me in this undertaking. True wisdom comes from knowing Him.

This research was funded by the Significant Opportunities in Atmospheric Research and Science Program (SOARS) through the University Corporation for Atmospheric Research and the National Science Foundation, through fellowship support from the American Meteorological Society, and through the following grants: NSF Grant ATM-0087072, NSF Grant ATM-0332197, NASA/CAMEX Grant NAG5-11010, and NOAA Grant NA17RJ1228.

DEDICATION

To the continued success of Native Americans in atmospheric science.

CONTENTS

FIGURES

TABLES

Chapter 1

INTRODUCTION

1.1 Motivation: The hurricane problem

Hurricanes[1] pose a great threat to society. Their violent winds, flooding rains, coastal storm surges, and waves have caused economic upheaval and human tragedy throughout recorded history. Hurricane impacts on society are increasing due to population growth and several economic and societal factors. Recent disasters and near misses further underscore the need for better forecasts so that further tragedy can be avoided.

Research on multidecadal atmosphere-ocean circulation regimes suggests that the lull in Atlantic tropical cyclone activity from 1971–1994, reflected most strikingly by the scarcity of "major" hurricanes,[2] has given way to a period of increased activity that may last for the next 10–40 years (Goldenberg et al. 2001). The 23 major hurricanes observed from 1995–2000 make this 6-year period the most active on record.[3] In contrast, only 36 major hurricanes were observed during the previous 24-year period. During the quiescent period of the past couple decades, coastal populations have swelled. Combined with economic growth and social trends, more wealth and property are now at risk from hurricanes than ever before. Since major hurricanes are responsible for 83 percent of all United States

[1] A hurricane has a 1-min surface wind speed of at least 64 kt. In this chapter, the term 'hurricane' is used generically for tropical cyclones (TCs) of all intensities.

[2] A major hurricane has a 1-min surface wind speed of at least 100 kt.

[3] Reliable statistics on the numbers and intensity of Atlantic tropical cyclones only extend back to 1944 when aircraft reconnaissance began. Further care must be taken in considering basin wide statistics before the era of continuous satellite surveillance, which began in 1966 (Jarvinen et al. 1984).

hurricane damage, we are likely entering a period of substantial economic losses (Pielke and Landsea 1998).

While hurricane-related economic losses increased substantially during the twentieth century, U.S. death tolls have shown a marked downward trend. Since much of the coastal population growth occurred during a relative lull in hurricane activity, many coastal residents have never experienced the core of an intense hurricane. At the same time, a social trend toward urbanization has concentrated large numbers of people in vulnerable cities along the East and Gulf Coasts. Transportation infrastructure has not kept pace with population growth in most cities, resulting in congestion and decreased mobility. As a result, the lead times required for evacuation of some cities have increased faster than the rate of improvement in hurricane forecasts, raising the potential for future storms to exact large human tolls (Sheets 1990). Societal vulnerabilities have also increased in the developing nations of the Caribbean and Central America, but for different reasons. Exploding populations with associated poverty and deforestation have lead to staggering death tolls, primarily from inland flooding. In 1998, Hurricane Georges hit the Dominican Republic and Haiti, killing 602. Several months later, Hurricane Mitch pounded Honduras and Nicaragua, causing 9086 direct deaths (with an additional 9191 persons reported missing), making it one of the deadliest Atlantic hurricanes in history (Pasch et al. 2001). If the trend toward increasing vulnerability continues, future storms may cause unprecedented death tolls both in the United States and abroad.

In order to avoid such an undesirable outcome, better forecasts are needed to provide enough warning time for evacuations and other mitigation activities such as structure hardening.[4] To be useful for mitigation purposes, such forecasts should be timely and accurate. Foreknowledge of the intensity at landfall, and details of the wind field, rainfall, and surge effects are important, but the future storm track is probably the most basic and critical

[4] Long term mitigation strategies are also vital in preventing massive casualties in the future. Such strategies include tougher building codes, restrictive zoning to prevent additional development in flood- and surge-prone areas, increased transportation capacity where evacuation choke points currently exist, and greater public education on what actions to take for various storm scenarios.

component of a hurricane forecast.

1.2 Limits of predictability

Track errors of the official National Hurricane Center/Tropical Prediction Center (NHC/TPC) forecasts have shown a slow but steady decrease over the past three decades. According to McAdie and Lawrence (2000), official track errors (adjusted for forecast difficulty) have decreased each year by an average of 1.0% at 24 h, by 1.7% at 48 h, and by 1.9% at 72 h for the period 1970-1998. The rate of improvement has accelerated during the most recent 5-year period in their study (1994-1998), with yearly decreases of 2.1%, 3.1%, and 3.5%, respectively. They suggest that the reductions in official forecast track errors are due primarily to the steady increase in skill of global numerical weather prediction (NWP) models, coupled with the improved ability of forecasters to correctly identify the initial storm position using satellite imagery and aircraft reconnaissance data. It is unknown how long this beneficial trend can be sustained through incremental model improvements, better remote sensing platforms, and increased forecaster sophistication. At some point in the future, the practical limit of predictability (as evidenced by the state-of-the-art numerical models at that time) will approach the atmosphere's inherent predictability limit, which arises from nonlinear scale interactions and instability mechanisms coupled with an incomplete knowledge of the initial state (i.e. chaos).

The inherent predictability limit can be defined as the performance of a perfect model with perfect initial conditions. Leslie et al. (1998) have estimated the inherent and practical (simply the performance achieved by current NWP models) predictability limits and shown that there is still plenty of room for further improvement in NWP models. Nevertheless, it behooves the research community to develop methods that can improve forecast skill by recapturing skill normally lost to the inherent predictability limit. As the models become closer to "perfect", further sophistication of the model physics, increased resolution, and better data assimilation will no longer lead to incremental improvements in forecast skill.

Brooks and Doswell (1993) suggest that future developments in computer technology may be more effectively put to use in an ensemble framework, where many model realizations are conducted based on perturbations that try to simulate the uncertainties in the dynamical system of interest. In this manner the ensemble approach uses the uncertainties (both in the initial state and in formulation of the model) to its advantage, thus providing information on the possible range of future atmospheric states, including low probability events. In application to the problem of TC track forecasting, the ensemble information could provide the forecaster with the geographical range of probable (or possible) track forecasts and allow estimation of the uncertainty in the forecast. In contrast, a single model realization only provides one forecast track, which may be right or wrong.

1.3 Research goals

This research aims to develop an ensemble method for predicting hurricane tracks using a simple nondivergent barotropic model. The recent development of an extremely efficient nondivergent barotropic model using multigrid methods (Fulton 2001) has opened the door to forecast methods that utilize very large ensembles, but are computationally inexpensive enough to run on a personal computer (PC). Various metrics are employed to interpret the results and gauge the efficacy of such methods. Some of the questions we seek to answer are:

Can a well-perturbed ensemble give a better forecast than any single model realization?

How many ensemble members are necessary to get the "right" answer?

Is there a relationship between ensemble spread and forecast error?

Can information about the ensemble spread be used to provide a meaningful forecast of forecast skill?

How accurately does the ensemble envelope of all track possibilities encompass the actual tracks observed?

Can a barotropic model provide a useful framework for ensemble forecasts of TC tracks, or is it necessary to include baroclinic dynamics?

It is our hope that this research will verify that ensemble methods have great utility in maximizing the benefits of increasing computer power, and lead to more accurate forecasts that may prevent devastating loss of life in the future. The rest of this paper is arranged as follows: in Chapter 2, a literature review on ensemble methods is presented. Chapter 3 reviews the details of the multigrid barotropic model used in the ensemble, discusses the process by which an optimum configuration is obtained, and compares its performance to that of a current operational barotropic model. Chapter 4 discusses the design philosophy of the ensemble, the choice of perturbation classes and magnitudes, and the implementation of the ensemble in a semi-operational framework. Methods of verification and results are presented in Chapter 5, while several case studies are given in Chapter 6. Chapter 7 closes with conclusions and interpretation of the results.

Chapter 2

LITERATURE REVIEW

2.1 Forecasting and predictability

The application of forecasting the future state of the atmosphere has a rich and varied history. The earliest examples of forecasting the weather relied on rules of thumb (e.g., "Red sky in the morning, sailors take warning.") derived from empirical observation. The invention of the telegraph led to routine synoptic observations of the atmosphere, which were plotted onto daily weather maps. From these maps, forecasts were made based on pattern recognition and conceptual knowledge of the atmosphere's behavior (e.g. advection, modification of airmasses, and storm motion).

Vilhelm Bjerknes (1904) conceived the notion of numerical weather prediction, by which the physical laws governing the atmosphere could be used to predict its future state. The first attempt at numerical weather prediction was made by Lewis Fry Richardson (1922). Although Richardson's single 6-h integration failed due to an incomplete understanding of what was required for a successful forecast (e.g. the importance of geostrophic balance between the mass and motion fields in the initial condition), he correctly foresaw the tasks that would be necessary in order to make a numerical weather prediction, namely data acquisition, processing procedures, and forecast dissemination.

The invention of the electronic computer in the mid-1940s spurred further efforts in numerical weather prediction. In 1948, Jule G. Charney established a Meteorology Group for this purpose within the Electronic Computer Project at the Institute for Advanced Study in Princeton, New Jersey. By 1955, operational numerical weather prediction (NWP)

commenced. A more complete history of NWP is chronicled by Shuman (1989).

The practical experience gained through operational NWP spurred rapid gains in knowledge that led to improved models, so that by 1960, the model skill for some products surpassed that of human forecasters. However, some weaknesses of the deterministic method (i.e., integrating forward in time from an initial state) soon became apparent. Lorenz (1963) noticed that a forecast carried on from a slightly perturbed intermediate state of a previous forecast diverged from the original forecast after some time, eventually losing any likeness to the original. Because of the limit in numerical precision in storing the data, the intermediate state used to initialize the second forecast was slightly different than the state represented in computer memory. It was recognized that imperceptible differences between initial states eventually lead to completely different evolutions of the flow. This sensitivity to initial conditions couples with various instability mechanisms to cause chaotic behavior – small errors (the difference between the forecast state and the observed state) grow rapidly in time. At some point, the error saturates and all predictability is lost. Thus, the deterministic method is limited by our inability to completely and accurately describe the initial state of the atmosphere and to model the atmosphere's physics (due to approximations made in the model equations and in the numerical methods used for solving them). Even with a perfect model, the imprecise specification of the initial state sets an inherent limit on the atmosphere's predictability, as mentioned in Chapter 1.

Several alternatives to the deterministic method have been proposed to take the uncertainty of the initial state into account. Each seeks to provide information on the time development of the distribution of the variables in the dynamical system. The uncertainty for a variable (which tends to be small at the initial time) can be expressed in terms of a probability density function (PDF), wherein the likelihood that the variable takes on the actual value is given in terms of probability. Figure 2.1 shows an example of two hypothetical PDFs for a variable. The first one is strongly peaked, indicating a large probability that the variable is close to the actual value: a case of small uncertainty. The second PDF is

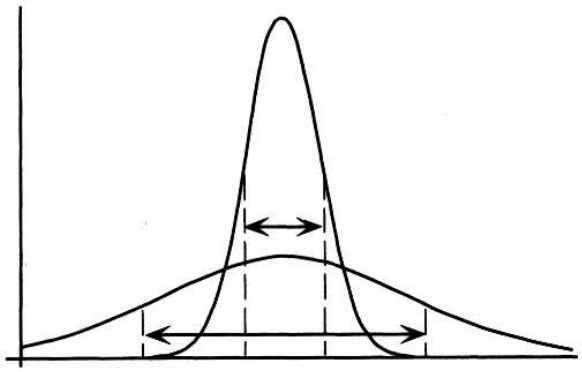


Figure 2.1: Two hypothetical probability distributions shown as probability density functions (PDFs). The two distributions have the same location (expected value), but reflect different degrees of uncertainty. The tall, narrow distribution represents less uncertainty, while the broader distribution represents more uncertainty. Arrows delineate the 75% central credible intervals in each case. From Wilks (1995).

flatter, expressing an increased uncertainty in the actual value (as is the case for a forecast verifying further out in time). If the shape of the variable's PDF can be predicted, along with the expected value, then forecasts of forecast skill can be made at the time of the forecast, indicating the likelihood that the forecast state encompasses the actual state.

In discussing predictability theory, it is useful to consider the concept of phase space (Gleeson 1970), wherein each dimension corresponds to a model variable. The model atmosphere's initial state in such a multidimensional phase space consists of a number of points, each of which represents a possible instantaneous state of the system. An example of such a phase space is shown in Fig. 2.2. Each state evolves according to the laws governing the dynamical system, resulting in a new state. The dynamical pathway taken by a given state is referred to as a trajectory. The collection of all states for a given time can be interpreted as a multivariate PDF, with the uncertainties of the system variables expressed in terms of probability.

Edward Epstein developed a prediction method (Epstein 1969) based on analogies to statistical physics and probability theory. Known as stochastic dynamical prediction, this method treats the probabilistic nature of initialization by recognizing that it is impossible

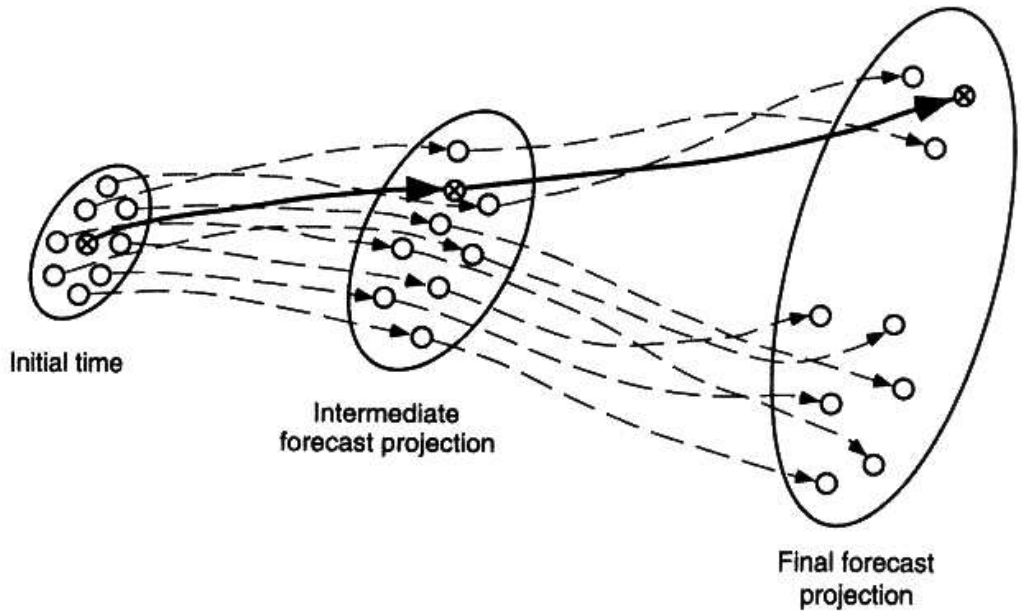


Figure 2.2: Schematic illustration of ensemble forecasting, plotted in terms of a two-dimensional phase space. The heavy line represents the evolution of the single "best" analysis of the initial state of the atmosphere, corresponding to the more traditional single deterministic forecast. Dashed lines represent the evolution of individual ensemble members. The ellipse in which they originate represents the statistical distribution of initial atmospheric states, which are very close to each other. At the intermediate projection, all of the ensemble members are still reasonably similar. By the time of the final projection, some of the ensemble members have undergone a regime change, and thus represent qualitatively different flows. Any of the ensemble members, including the solid line, are plausible trajectories for the evolution of the real atmosphere, and there is no way of knowing in advance which will represent the real atmosphere most closely. From Wilks (1995).

to measure the state of the atmosphere with sufficiently dense or accurate observations. A large number of states (or points in phase space), each consistent with the analysis data, form an ensemble that represents the initial state of the atmosphere in probabilistic form. It is assumed that the trajectories taken by the various states do not cross (they are independent), and that states are neither created nor destroyed, allowing a continuum hypothesis to be used. Then it is possible to formulate a continuity equation for probability (analogous to the continuity equation for mass), known as the *Liouville equation*, which allows the statistical properties of the ensemble of realizations (the forecast) to be known without actually calculating the larger numbers of individual trajectories associated with the ensemble members. In practice, simple correction terms are added to the prognostic equations, converting them to stochastic dynamical equations (SDEs) that predict the expected state of the atmosphere. But in order to evaluate the correction terms, additional prognostic equations are required for the higher moment terms (e.g. the variance and covariances for the variables), and so on. To close the system, an approximation is required – generally, the third and higher moment terms are neglected. This method produces forecasts that have mean square errors significantly smaller than those produced using the deterministic method. At the same time, information is provided about the uncertainty associated with the forecast, along with an estimate for how long the forecast is applicable. Nevertheless, there are significant drawbacks. The integration of the SDEs is computationally expensive due to the prediction of higher moment terms. Also, it is difficult to treat the errors in model formulation. Finally, to derive the prognostic SDEs, some closure assumptions must be used in the higher moment terms. A more general treatment of the *Liouville equation* has been made (Ehrendorfer 1994a,b) which avoids several of these difficulties, but its applicability to complex multidimensional systems such as the atmosphere is currently impractical.

In light of the disadvantages of the stochastic dynamical method or other methods utilizing the *Liouville equation* to calculate the forecast uncertainties directly, more pragmatic approaches have been developed: ensemble methods. In these methods, an ensemble

consisting of a finite number of realizations is used to sample the phase space of all available solutions. In practice, N samples of the initial state, called perturbations, are integrated forward by a model, resulting in N forecasts. The resulting subspace of solutions are then used to construct a synthetic PDF that should approximate the true PDF. By means of a large number of simulations, the expected value (the ensemble mean) and higher moment statistics of the forecast can be estimated, leading to estimates of the uncertainty of the forecast. The next section provides an overview of the various ensemble approaches, with some degree of generality.

2.2 Overview of ensemble methods

Aberson (2001) provides a useful definition of a forecast ensemble: "A forecast ensemble is, broadly, any set of forecasts that verify at the same time." There are at least five general ensemble methods in use today: consensus, lagged average forecasting, Monte Carlo methods, breeding of growing modes, and singular vector decomposition. Each of these methods seeks to treat the uncertainties associated with the initial condition and/or model formulation by taking a collection of forecasts that start from slightly different initial conditions.

The simplest ensemble approach in an operational framework is the consensus approach (also known as the "poor man's ensemble" for reasons that will become evident). In this method, the output of several independent NWP models is simply averaged together to get a consensus forecast that should on average be better than any of the individual models' forecasts. Statistics such as the variance are computed based on the individual model differences from the consensus forecast. If the individual models all predict a similar evolution, then the resulting variance will be low, and the consensus forecast can be interpreted as having better than average reliability. If their solutions rapidly diverge, the variance will be high, and one might expect that the ensemble mean will have poor reliability. Thus, an *a priori* determination of forecast skill may be possible at the time the

forecast is made. Of course, the consensus approach relies on having several skillful, yet independent NWP models. If the models are not independent, the improvement of the consensus forecast over any given model will be reduced. Models are independent to the extent that they use different initialization schemes, different formulations of the physical laws governing the atmosphere, and different numerical approaches to integrate the prognostic equations. The great advantage of the consensus approach is that no additional model runs are required (hence the "poor man" moniker), just a postprocessing of the models that are already running operationally. The uncertainties in the initial condition are treated by the diverse data initialization schemes associated with the NWP models. Each scheme may use different sources of data than the others (e.g., one scheme may ingest satellite data, while another might use aircraft data). Further differences will arise from the varied methods of constructing an analysis from the data, yet each model should be consistent with the state of the atmosphere. Uncertainties in model formulation are handled by the variety of model formulations amongst the NWP models (e.g., the models likely use different convection and radiation parameterizations, planetary boundary layer formulations, and numerical methods). The consensus method has two disadvantages: sometimes one or more of the operational suite of models will be missing, and frequent model upgrades make long-term interpretations of the consensus skill difficult.

Hoffman and Kalnay (1983) suggest another ensemble approach that uses forecasts from a single NWP model: the lagged average forecast (LAF) method. Each member of the ensemble is a forecast which starts at a different time (lagged from the most recent run), but verifying at the same time as the most recent run. The output of these forecasts are processed as in the consensus approach. One obvious advantage of the LAF method is that no additional model runs are needed, since previous operational runs are used. Also, information contained in previous analyses and forecasts are utilized, taking the uncertainty of the initial state into account. As a side note, operational forecasters have used LAF in a qualitative manner for many years. Known as d($prog$)/dt, this method is a matter of

forecaster interpretation. Run-to-run consistency engenders high confidence, while low confidence usually ensues when the model runs flip back and forth between distinctly different solutions. In this case, the forecaster may still gain some information as to the range of possible solutions.

The next and perhaps most basic ensemble method is the Monte Carlo forecasting (MCF) method. Ensemble members are created by randomly selecting a number of initial states from a collection of possible initial states. Ideally, this collection should have a density in phase space proportional to the PDF of the true state (which can be estimated through objective analysis). Deterministic forecasts are made for each of the ensemble member initial states, resulting in an equal number of ensemble member forecasts. If the perturbations to the initial state reflect the error distribution of the analysis, then the collection of forecast states should have a PDF similar to the actual atmosphere. The MCF ensemble average is essentially an approximation of the stochastic dynamical prediction method. Leith (1974) showed that the MCF method gives a more accurate forecast on average than any of the individual ensemble members. For the ensemble mean, improvement generally increases as the number of members increases, with the most improvement occurring as members are added to the smallest ensembles (i.e., the improvement gained in going from an ensemble size of 2 to 3 is greater than the improvement in going from 20 to 21 members). The main disadvantage of the method is that a large number of integrations are required to properly ascertain the higher-order moments. Another disadvantage of the MCF ensemble is that random perturbations project onto modes that tend to grow more slowly than actual modes (such as those arising from baroclinic instability) in the atmosphere. As a result, forecasts generated from random perturbations tend to be clustered too closely to the ensemble mean forecast (i.e., the ensemble spread is unrealistically small given the actual uncertainty). Thus, random perturbations are not well-suited to estimating the subspace of dynamical pathways available to the system – simply adding more members to a MCF ensemble is inefficient.

The current state of the art ensemble methods being used by operational NWP centers involve the generation of dynamically-constrained perturbations. These methods, known as the breeding of growing modes (BGM) or singular vector decomposition (SVD), attempt to find or construct perturbations that project onto the fastest-growing modes. Ehrendorfer and Tribbia (1997) have shown that it is important to sample these fast growing modes to gain the most benefit (at least for short time ranges). The BGM and SVD methods use a combination of present and past analysis data, past and present forecasts, and random perturbations to either breed (in the case of the BGM method) or construct through linear adjoint methods (SVD method) optimum perturbations that project onto fast growing modes. Since these modes represent fast growing analysis errors, they provide the best estimate of the uncertainty in the initial state.

2.3 Previous studies of tropical cyclone motion using ensembles

2.3.1 First experiments

All studies of ensemble TC track prediction are recent. Despite the theoretical development of ensemble ideas and predictability theory in the 1960-80s, operational implementation was limited by computer power and the general state of knowledge in NWP (i.e., the greatest improvements were still to be had by improving the numerical models). By the early 1990s, these considerations were no longer major impediments, and on 7 December 1992, operational ensemble forecasting commenced at the National Meteorological Center[1] (NMC), as described by Tracton and Kalnay (1993). The global ensemble utilized perturbations obtained from the LAF and the BGM methods, as outlined in Toth and Kalnay (1993). Although these perturbations were better suited for midlatitude weather systems (featuring baroclinic developments), the existence of globally-perturbed fields soon spurred some research in ensemble prediction of TC tracks.

In the first known study, Aberson et al. (1995) ran a barotropic ensemble for 59 cases

[1] Now the National Centers for Environmental Prediction (NCEP).

that occurred worldwide between 30 August and 15 September 1994. They used 16 different initial conditions, sampled on a $2.5^\circ \times 2.5^\circ$ latitude-longitude grid from the NMC medium-range ensemble suite, consisting of 11 Medium-Range Forecast (MRF) model runs and 5 Aviation (AVN) runs. These large-scale analyses were combined with a synthetic vortex and then integrated forward using the VICBAR model (for details concerning the VICBAR model, see Aberson and DeMaria 1994) to obtain 16 different forecasts. They found the ensemble was able to improve over the MRF forecast in some cases, but not in others. Many of the ensemble members clustered closely about the control forecast, but the actual track sometimes fell completely outside the envelope of ensemble forecast tracks. Their results were somewhat disappointing, but the lack of significant improvement might be attributed to the lack of perturbed boundary conditions (Mark DeMaria 2003, personal communication). In the context of a limited-area barotropic ensemble, much of the uncertainty in the track forecast does not necessarily result from uncertain initial conditions, but from uncertainties in the evolution of the environment. Their ensemble did not make use of the perturbed forecasts from the NMC ensemble, just the initial conditions.

Aberson et al. (1998) conducted a larger study of Atlantic and East Pacific TCs during the 1996 and 1997 seasons with the Geophysical Fluid Dynamics Laboratory (GFDL) model, a triply-nested primitive equation model. They again used the bred modes from the NCEP MRF ensemble members for the initial condition, and this time also used the ensemble member forecast fields to provide time-dependent boundary conditions to the GFDL model. The operational GFDL forecast used the high resolution triply-nested grid configuration, while the ensembles and control forecast (GFCT) used a lower resolution doubly-nested configuration. The resulting ensemble mean forecast (GFMN) produced forecasts better than the control and comparable to the operational GFDL. There was no clear correlation between ensemble spread and error, although the spread did seem to set an upper bound on the error. The ensemble spread did not reliably encompass the actual track, limiting its usefulness. This is probably related to the lack of spread in the bred modes from NCEP,

since the MRF ensemble system was designed primarily for the prediction of midlatitude weather systems, not systems in the tropics.

In another early study, Krishnamurti et al. (1997) illustrated the potential of ensemble forecasts for hurricane track forecasts. They conducted a limited experiment for three 1979 storms using an 8-member ensemble initialized from analyses produced by four different NWP groups (NCEP, ECMWF, GFDL, and NASA/Goddard) at two different initial times, 12 h apart. The data are from the First Global GARP (Global Atmospheric Research Program) Experiment (FGGE). Two sets of runs were conducted using the Florida State University Global Spectral Model (FSUGSM), one 8-member ensemble used physical initialization to constrain the initial condition (divergence, using observed gauge and satellite estimated rain rates), while the other did not. They found that physical initialization reduced the spread of the ensemble forecasts and resulted in a superior mean forecast, probably because the improved definition of mass sources and sinks led to a more accurate forecast for the storm environment. Thus the reduced variance was related to a decrease in false variance due to significant analysis errors, not the inherent analysis uncertainty.

Shortly thereafter, Zhang and Krishnamurti (1997) conducted another study of ensemble TC track prediction using a 15-member FSUGSM ensemble for three Atlantic hurricanes. They pointed out that the perturbation methodologies used for midlatitude ensemble prediction are not adequate for tropical weather systems, given the different physical and dynamical processes at work. Perturbation growth in the midlatitudes arises largely due to dynamical instability as described by linear perturbation theory, while in the tropics growth may be dominated by small-scale physical processes (such as cumulus convection and latent heating) interacting with large-scale systems. They proposed two classes of perturbations: perturbations to the initial hurricane position and perturbations to the environment and storm structure. The initial position of the hurricane was perturbed 50 km to the north, south, east, and west of the analyzed position, along with no perturbation (control), resulting in five different position perturbations of the initial position. The magnitudes of the

position perturbations were chosen based on characteristics of the initial position errors of various observational techniques (aircraft reconnaissance $\sim 20\,\mathrm{km}$, radar $\sim 20-55\,\mathrm{km}$, and satellite imagery $\sim 110\,\mathrm{km}$). Perturbations to the storm structure and environment were accomplished using an empirical orthogonal function (EOF)-based procedure analogous to the BGM method, starting from the ECMWF analysis. In this procedure, random perturbations (with magnitudes comparable to the observational errors) are added to the control analysis, then the model is integrated forward for 36 h. Difference fields between the perturbed and control runs are calculated every 3 h, and a data matrix is constructed. A three-dimensional EOF analysis determines the modes (eigenvectors) whose EOF coefficients increase rapidly in time. The first few modes (which are considered the fastest-growing modes), are then used to construct a perturbation field that is added to and subtracted from the control analysis. Since the resulting 3 perturbation fields grow in an environment (the full-physics 3-D FSUGSM) where the nonlinear effects of shear and convective cumulus are included, these fields include perturbations to the large-scale environment as well as storm intensity. The resulting 15-member ensemble mean forecast showed large improvement over the control forecast in two of the three cases, although care should be taken in interpreting these results given the limited sample size. They also noted that the ensemble spread did seem to offer an *a priori* measure of forecast reliability, with small spread indicating a more reliable forecast. In addition to the track problem, their update study (Zhang and Krishnamurti 1999) also examined ensemble prediction of intensity, wind field, maximum wind, and precipitation as well as methods of displaying the resulting information.

Most of the above studies assume a "perfect" model, and perturbations are used to ascertain the inherent uncertainty of the initial condition. In contrast, Ramamurthy and Jewett (1999) examined the uncertainty of model formulation using various configurations of the National Center for Atmospheric Research (NCAR)/Penn State Nonhydrostatic Mesoscale Model (MM5) (for details on the MM5 model, see Dudhia 1993). The impact of uncertainties in the initial conditions, cumulus parameterizations and boundary-layer

schemes, and boundary conditions were examined for Hurricane Opal's intensity, track, and landfall predictions. Four different cumulus parameterizations and two initial conditions were used, leading to 24 experiments. They found that both track and intensity forecasts exhibited considerable sensitivity to the choice of initial condition. The various cumulus convection schemes interacted differently with the two initial conditions, resulting in differing rates of intensification. They also found that the track and timing of landfall forecasts were improved when data assimilation was performed on the outer grid, with nudging toward the corresponding analyses for the first 24 h of integration.

2.3.2 Multimodel consensus

It has long been noted (Sanders 1973) that a consensus forecast derived from independent forecasts will usually be more accurate than any of the component forecasts. Individual forecasts may beat consensus, but over a long-term average, consensus usually wins. A theoretical discussion of why this occurs is given by Thompson (1977), who estimated that the combination of two independent forecasts would reduce the error by about 20%. Leslie and Fraedrich (1990) showed that an optimal linear combination of a NWP forecast model and a climatology and persistance model significantly reduced the mean forecast-position errors (by 15% at 24 h and 17% at 48 h) from the next best forecasting methods at those time periods.

Mundell and Rupp (1995) describe the development and operational implementation of a larger consensus, which consists of some average of nine different forecast aids available to the forecasters at Joint Typhoon Warning Center (JTWC). The simple average (consensus blend) of all nine components showed improvement of 10% at 24 h and 15% at 48 h over the average errors of the individual components. Another average, which weights the contribution of the individual component aids by past performance, showed a 24% improvement over the average errors of the individual components at these time periods. Both of these "hybrid" aids achieved these remarkable results by staying in the center of the

forecast envelope defined by the component forecasts. This property of linear combinations guarantees that while consensus may not always be the best forecast, it will never be the worst. In addition to routinely beating the best of its component models in most individual forecasting situations, the performance-weighted consensus was more consistent than the individual components, as seen by its smaller error variances. This means that the consensus forecast was consistently more reliable than any of the individual forecast aids, making it an ideal tool for track forecasting.

Goerss (2000) provided a detailed study of the performance of a simple three-model (GFDL, NOGAPS, and UKMO) consensus average for the 1995-1996 Atlantic hurricane seasons and a similar consensus (GSM, NOGAPS, and UKMO) for the 1997 Northwest Pacific typhoon season. Improvements with respect to the best individual forecast model were 16%, 20%, and 23% at 24, 48, and 72 h respectively in the Atlantic, with similar results in the NW Pacific. He also found that the actual TC track was included in the 72-h consensus envelope 79% of the time. There was no clear correlation between the ensemble spread and the error of the ensemble mean, but the spread did appear to be related to the upper bound of error. In addition, only 7 cases out of 166 exhibited the undesirable combination of low spread and high error. Finally, the size of the 95% level of uncertainty (similar to twice the standard deviation) in the 24-h forecasts was reduced by 20%, which could have a direct impact on reducing the size of warning areas associated with TC landfalls.

Several recent studies have achieved very promising results by using a statistical regression model to process the forecasts of the individual member models, obtaining a forecast far superior to that of any member or even the simple ensemble average (consensus) forecast. One such effort, called the Multimodel Superensemble, is described by Vijaya Kumar et al. (2003). The statistical model is "trained" by regressing intensity and position forecasts from the individual models against the observed values for past storms (a training sample size of 60 storms was found to be optimum). A simple multiple linear regression

technique generates weights for each model for each 12-h forecast period out through 5 days. These bias estimates are then used in the forecast phase to construct the superensemble forecast from the individual model forecasts. Results for the Pacific TCs over a 3-year period (the training phase was repeated independently each season) show skill improvements of 50, 140, 160, and 200 km for TC position errors for 24-, 48-, 72-, and 96-h forecasts over the best individual model (the model with the smallest error). The superensemble also showed considerable improvement over the simple ensemble average (consensus) as well, which had errors comparable to the individual models. The lack of improvement of the consensus suggests that the individual models are highly correlated (Goerss 2000). Thus the success of the superensemble depends on the intermodel correlation, which is likely to be highly dependent on model formulation. But the models rarely stay the same for long because the various NWP centers frequently implement changes to the models. Thus, implementation of the superensemble to operational TC forecasting is problematic.

Weber (2003) describes another statistical ensemble prediction system (named STEPS) for TC tracks in the Atlantic basin. This system assumes that position errors are systematic, depending on storm structure, location, and motion. Annual performance data for the various NHC forecasting aids (the AVN, VICBAR, and LBAR models were excluded based on sensitivity tests) are regressed against storm data contained in the NHC advisories, including storm position, radius of maximum wind speed, radius of the outermost closed isobar, pressure at that radius, central surface pressure, maximum wind speed, translation direction, an estimate of the vertical extent of the storm, and estimates of the radii of 35- and 50-kt winds in each storm quadrant. Only data from the previous season is used for development, resulting in a set of statistical weights in the form of a linear combination of expected position errors and their standard deviations, computed as functions of all possible values of storm parameters. In this way, the models which perform the best for a particular storm structure or location are given the highest weights, while the models which are least skillful for that particular situation are given low weights. In addition, the weights are ap-

	0	12	24	36	48	60	72
Inherent	39	80	97	122	149	185	218
Practical	86	139	166	213	255	327	386

Table 2.1: Inherent and practical mean position errors (km) at 12-h intervals out to 72 h using a barotropic tropical cyclone model, for storm cases from the Atlantic basin from 1993-1996. From Leslie et al. (1998).

plied to the predicted positions of all ensemble members, resulting in a track forecast that shows the statistical confidence regions of strike probability. The mean position errors from 1997-2000 for an operational version of STEPS (using information only available at the time of the forecast) were 120, 215, and 296 km at 24, 48, and 72 h, respectively. The corresponding values for the operational ensemble were 128, 238, and 336 km, only marginally larger. But the standard deviations of the STEPS position errors were of comparable magnitude to those of the best numerical models out of the operational suite, being about two-thirds the size of the actual position errors. In addition, out of 332 72-h forecasts, only 2 STEPS forecasts (6 in the operational version) had position errors larger than 1000 km (compared to GFDL, 17 cases; UKMO, 7 cases; NHC official forecast, 28 cases). The mean diameters of the 66.7% strike probability regions were 304, 682, and 1033 km at the 24-, 48-, and 72-h forecast times. The lack of a particular model for a given forecast (as happens frequently in an operational setting) does not significantly degrade the quality of the forecast. Weber concludes that STEPS is able to consistently provide high-quality forecasts with accuracy comparable to those of the best numerical models, yet without the seasonal variations in model performance. In addition to these benefits, a real time 72-h STEPS forecast only takes about 3 s of computer time on a modest PC, including the generation of the strike probability maps.

2.3.3 Barotropic ensembles

Besides the studies already mentioned under "early experiments", several other studies have utilized barotropic ensembles to study TC track prediction. Since this thesis will use a barotropic ensemble, these studies are of particular interest. The first study (Leslie et al. 1998) estimated the inherent and practical limits of predictability. As a reminder to the reader, the inherent predictability limit is the smallest mean position error that would be expected if the model and the initial condition are both 'perfect'.[2] The practical predictability limit is simply the performance achieved by current operational NWP systems with initial conditions that have errors characteristic of operational analysis schemes. They used both barotropic and baroclinic models. The barotropic model ran at 15 km resolution on a 250 × 250 km grid using 850-300 hPa deep layer-mean winds derived from the Australian Bureau of Meteorology's tropical analysis scheme. The model uses a very efficient multigrid solver on the stream function-vorticity equation. A 72-h forecast only takes 20 s, making large ensemble runs possible. Their estimates of the predictability limits using a barotropic model for the Atlantic basin are shown in Table 2.1. The barotropic model errors are about 40-50% greater than what might be expected using a perfect model with perfect initial conditions. The practical limit achieved by the baroclinic model in their study had errors that were 35-40% greater. When the baroclinic model was run with additional satellite-derived data (at higher spatial and temporal resolution), an advanced four-dimensional data assimilation scheme, and a very high model resolution (15 km grid spacing), the practical limit errors dropped to just 30-35% greater than those of the inherent limit. This suggests that there is still much room for improvement, even in today's

[2] Even if the initial state and model are perfectly resolvable down to the limits of the macroscopic scale (i.e. 1 cm), errors will still arise (eventually) due to the microscopic fluctuations (the proverbial butterfly flapping its wings). Through instability mechanisms and nonlinear scale interactions, these errors grow to the larger scales, preventing a perfect forecast. So when we opine about 'perfect' models, we are talking about a certain practical level of perfection which lies at the limits of human understanding and technological capabilities. Another level of perfection could be imagined in which each atom of the system is modeled, but even here, the Heisenberg Uncertainty Principle would prevent the initial state from ever being measured completely perfectly.

state-of-the-art NWP models. Examination of the distribution of errors shows that the incidence of very large forecast errors was reduced mainly by the inclusion of additional data, increased model resolution, and an improved assimilation scheme. In light of the limitations of barotropic models (one-level dynamics), including these additional data are difficult or impossible.

The studies most relevant to this thesis are those of Cheung and Chan (1999a,b). Their systematic studies examined the effects of different methods for perturbing the storm environment and the vortex. A barotropic model with 50-km grid spacing was used. The final analyses from the Tropical Cyclone Motion (TCM-90) Experiment (Rogers et al. 1993) were used to derive an 850-300 hPa deep-layer mean wind field for the initial condition. The axisymmetric vortex of Chan and Williams (1987) was bogussed at the best-track position. The first part of the study focused solely on perturbations to the environment, so all properties of the bogus vortex remained unchanged. Wind information from the corresponding analyses was blended in during the integration to provide time-dependent boundary conditions and to nudge the model solution towards the observed analysis within the blending zone.

In studying the MCF method, random perturbations were added to the environmental winds with root-mean-square (rms) magnitudes comparable to the square root of the error covariance of the mid-tropospheric wind field for a given NWP model. Ninety-nine ensemble members (plus one control) were thusly generated, resulting in a 100-member MCF ensemble. Results for 9 different cases showed that the ensemble members tend to spread symmetrically about the control, with the ensemble mean falling near the control forecast. Thus they concluded that the MCF method has little forecast utility in this context. No further investigation was made for the MCF method.

Next they tested the LAF method. Perturbations were constructed by filtering the verifying vortex in the analysis at the initial time, taking the difference between this filtered analysis and a forecast verifying at that time that started previously, and then inserting a

bogus vortex at the correct position. Additional perturbations were also generated by finding the short-range difference between the predicted environments of any 2 lagged forecasts, giving a total of 6 pairs. This gave a 17-member ensemble, including the control forecast. They also tested the BGM method. Eight pairs of perturbations were bred for 24 h, starting from a randomly-perturbed analysis. The bred error was rescaled back to a magnitude of $3\,\mathrm{m\,s^{-1}}$, and the resulting perturbation was added back to the filtered analysis, with a bogus vortex inserted at the correct position. This also resulted in a 17-member ensemble (in order to compare the LAF and BGM methods, it is important to use an ensemble of similar size).

Results from the LAF and BGM experiments show little significant differences between the two methods when compared to the best track forecasts. Skill scores in both methodologies degrade substantially with time, with most median scores below zero (negative skill). This may suggest problems in applying these ensemble techniques to a pure barotropic model. Nevertheless, a high degree of spread-skill correlation was demonstrated under the perfect model approach (PMA), in which one ensemble member, such as the control forecast, is taken to be "correct". The LAF and BGM techniques were found to be most successful when the synoptic environment exhibited certain characteristics: 1) a TC making a transition from one synoptic regime to another, 2) interaction with an apparent break in the subtropical ridge, 3) a rapid strengthening or weakening in the subtropical ridge, 4) potential recurvature of a TC, and 5) cases in which there were multiple TCs.

In their second study, Cheung and Chan (1999b) examined perturbations to the vortex. Two series of experiments were conducted, numbering eight in total. The first three experiments applied the MCF, LAF, and BGM methods to the analyzed vortex only, instead of the entire domain as in the first study. No synthetic vortex was used in this series of experiments (given the designations MCFV, LAFV, and BGMV, respectively). The second series of experiments viewed the synthetic vortex as part of the model configuration, so the idea was to perturb the vortex parameters systematically, as this might sample the vortex

uncertainty better than the first set of experiments, which focused on the environmental uncertainty. Thus, perturbations are made to observed storm parameters such as the initial position of the storm (designated IPER) and/or the storm's persistence (motion) vector. The addition of a persistence vector to an axisymmetric vortex profile produces asymmetric storm structure (i.e. the so-called β gyres). These experiments are designated as BETA.

To create the β gyre perturbations, the axisymmetric part of the vortex is first perturbed by adding random Gaussian noise to the three basic properties of the axisymmetric tangential wind profile: intensity (maximum sustained wind), radius of maximum wind, and the size parameter (which determines how the wind speed decreases to the environmental value in the outer core). The vortex is then allowed to spin up in a quiescent environment using a high resolution model, allowing the β gyres to develop and become quasi-steady. The spun-up vortex is then blended into the control analysis. The persistence vector is perturbed using uncertainty in the past 6-h position (determined from the error statistics of several past seasons). The BETA(0) experiment consists of the perturbed spun-up vortices only. In BETA(1), the β gyres are perturbed. BETA(2) perturbs the persistence vector, while BETA(3) perturbs both kinds of asymmetry simultaneously. All eight ensemble methodologies used an ensemble size of 17 members.

They found that the addition of random noise to the vortex (MCFV) or to the initial position (IPER) did not offer any improvement – the ensemble mean differed little from the control forecast. Apparently, the random noise was easily smoothed out by the strong cyclonic shear of the vortex. It was not possible to identify the vorticity center during the model integration for about half of the LAFV and BGMV cases, so these were excluded from the verification. These failures seemed to occur when the vortex was easily distorted by the perturbations. Most of the successful cases showed positive skill out through 72 h, using a PMA approach. The BGMV method had higher median scores than the LAFV method for most forecast time periods, suggested that the breeding method provides better samples from the analysis errors than the lagged-average approach. The BETA(0) and

BETA(1) experiments showed better skill than BETA(2) and BETA(3) under a PMA verification. Median scores tended to be positive at all times, as in the results from the LAFV and BGMV experiments. This suggests that an approach that perturbs the β gyres can be useful in a barotropic model. When just the persistence vector is perturbed, the ensemble spreads symmetrically (as in IPER), so little benefit is gained. Perturbing both the persistence vector and the β gyres did not seem to have a great effect either, with scores similar to BETA(2). But when verified by the best track, BETA(3) exhibits increasing skill by the 72-h time period. In looking at the spread-skill correlation under the PMA approach (the model is assumed to have no deficiencies), they found that the LAFV, BGMV, BETA(0), and BETA(1) experiments all showed high correlations. The authors suggest that BETA(3) should be an ideal alternative to the dynamically-constrained methods of perturbing the vortex, since the combination of asymmetric perturbations plus the persistence vector perturbation should provide alternative (and possibly correct) flows, given that real storm structure includes these effects. They conclude, however, that just perturbing the vortex is not a sufficient way to sample the uncertainty associated with the initial condition; vortex perturbations must be combined with perturbations to the storm environment to obtain an adequate ensemble framework.

This thesis aims to construct such a framework.

Chapter 3

THE MULTIGRID BAROTROPIC MODEL

3.1 Introduction

Tropical cyclones are complex systems involving dynamics on many scales, complicated physics, and multifarious interactions with ocean, land, and the surrounding environment. To include a domain large enough to predict the storm environment, yet still resolve the fine scales of motion near the eyewall, tropical cyclone models often employ variable resolution grids (e.g., Kurihara et al. 1998).

Although tropical cyclone modeling involves a wide range of scales, a reasonably accurate track forecast can be obtained without including all of the details of the inner core of the storm. For example, models based on simplified barotropic dynamics can provide useful forecasts of tropical cyclone motion. Since they are computationally cheap to run, the National Hurricane Center's (NHC) suite of operational guidance products includes barotropic models. Inherently faster than three-dimensional full-physics models, efficient barotropic models open the door to forecasting techniques that utilize very large ensembles.

The first operational barotropic tropical cyclone track models were developed during the late 1950s and early 1960s (Tracy 1966) . Due to limited computer power, these models attempted to separate the prediction of the storm environment from the smaller-scale vortex circulation. Because of these limitations, early barotropic models tended to have larger forecast errors than the simpler statistical track models available at that time.

A more successful operational barotropic tropical cyclone model (SANBAR) was developed in the late 1960s (Sanders and Burpee 1968; Sanders et al. 1975, 1980). The

Model	Dynamics	Discretization	Nested?
SANBAR	nondivergent	finite-difference	no
VICBAR	divergent	Galerkin/B-spline	yes
LBAR	divergent	Galerkin/harmonic-sine	no
MUDBAR	nondivergent	finite-difference	yes

Table 3.1: Characteristics of selected barotropic tropical cyclone models

SANBAR model was run as part of the NHC operational suite until 1989 when it was replaced by VICBAR. The VICBAR model (DeMaria et al. 1992) introduced an improved discretization scheme, based on B-splines and implemented with nesting. The LBAR model[1] (Horsfall et al. 1997), with harmonic-sine basis and without nesting, was later developed as a simpler and more portable operational alternative to VICBAR. Some characteristics of these models are given in Table 3.1.

Recently a new multigrid barotropic model (MUDBAR) has been developed (Fulton 2001). Based on the simple modified barotropic vorticity equation, MUDBAR uses an adaptive multigrid method to refine the mesh around the moving vortex, with the goal of maximizing accuracy while minimizing computational cost. MUDBAR can also be run in a mode that uses fixed-size movable meshes. This model was developed primarily as a test bed for adaptive multigrid techniques and was previously evaluated using idealized initial conditions. However, MUDBAR's speed and accuracy make it a useful alternative to existing operational barotropic models. To compare MUDBAR to operational barotropic models, the capability to include initial conditions from real data using a procedure similar to that of the LBAR model was implemented.

The purpose of this chapter[2] is to evaluate the accuracy, skill, and efficiency of the MUDBAR model by comparing it to the current operational barotropic model LBAR.

[1] Originally the Limited-Area Sine Transform Barotropic model (LASTBAR).

[2] Most of the text in this chapter closely follows a paper (Vigh et al. 2003) that was recently published in *Monthly Weather Review*. The portions of text on the LBAR and MUDBAR models were written by Mark DeMaria and Scott Fulton, respectively.

Section 3.2 briefly reviews the LBAR and MUDBAR models and Section 3.3 describes the data used for the comparison. Results for the 2001 Atlantic hurricane season are presented in Section 3.4, and conclusions are summarized in Section 3.5.

3.2 Model descriptions

Both LBAR and MUDBAR are developed on a section of the sphere, transforming longitude λ and latitude ϕ to Cartesian coordinates x and y via the Mercator projection

$$x = (\lambda - \lambda_0) a \cos\phi_0, \qquad y = \left[\tanh^{-1}(\sin\phi) - \tanh^{-1}(\sin\phi_0)\right] a \cos\phi_0, \tag{3.1}$$

where a is the radius of the earth. The projection is true at (λ_0, ϕ_0), where $(x, y) = (0, 0)$, which is taken as the center of the model domain. The models differ in the equations and discretizations used as described below.

3.2.1 LBAR

The LBAR model was briefly described in Horsfall et al. (1997). Based on the divergent barotropic (shallow water) equations with a mean fluid depth of 800 m, the model uses a harmonic-sine series representation and a Galerkin projection. Second- and fourth-order diffusion terms are included in the prediction equations. To accommodate nonperiodic boundary conditions, the dependent variables (horizontal wind components and geopotential height) are divided into boundary and interior functions using the technique described by Chen and Kuo (1992). The boundary function is the solution to Laplace's equation with inhomogeneous boundary conditions; the interior function satisfies homogeneous boundary conditions and thus can be expanded in a double sine series. Boundary conditions for the forecasts are obtained from the Aviation model run (AVN) of the National Centers for Environmental Prediction (NCEP) Medium-Range Forecast Model (MRF)[3] . The sine series include 48 terms on a 7200-km square domain. Nonlinear terms are evaluated using the

[3] As of 2002, the NCEP's global model is called the Global Forecasting System. The MRF runs have been replaced by four daily AVN runs.

transform method, with 96 transform grid points in x and y. The model uses centered time differencing with a time step of 180 s. Forward differencing is used for the diffusion terms.

The initialization procedure for LBAR is the same as for VICBAR (DeMaria et al. 1992). Large-scale wind and height fields are determined by vertically averaging (850–200 hPa) the AVN analysis fields. Heights from the global model are calculated as deviations from the heights in the U.S. standard atmosphere. The tropical cyclone is represented by the sum of an idealized vortex and an initial storm motion vector, both chosen to closely approximate the observed storm. The large-scale and tropical cyclone wind fields are blended as described in Section 3.3 to obtain the initial wind field for the model. The corresponding height field is calculated by solving the nonlinear balance equation, using the analysis heights at the boundaries.

To increase the influence of the boundary conditions, the predictive equations for wind and height contain nudging terms that damp the difference between the LBAR-predicted fields and the corresponding fields from the AVN model run. This nudging is applied in a boundary strip of width $s_0 = 3000$ km with amplitude proportional to $(1 - s/s_0)^4$, where s is the distance to the nearest boundary. The nudging coefficient is chosen so that the difference between the LBAR and AVN model fields has an e-folding time of 2 h at the boundaries.

3.2.2 MUDBAR

The MUDBAR model is based on the modified barotropic vorticity equation

$$\frac{\partial q}{\partial t} + m^2 \frac{\partial(\psi, q)}{\partial(x, y)} + \beta m \frac{\partial \psi}{\partial x} = 0, \tag{3.2}$$

where $\beta = 2\Omega a^{-1} \cos\phi$ (with Ω the rotation rate of the earth) and $m = \cos\phi_0 / \cos\phi$ is the map factor. The streamfunction ψ and potential vorticity q are related via the elliptic problem

$$\left(m^2 \nabla^2 - \gamma^2\right) \psi = q, \tag{3.3}$$

where γ is the inverse of the effective Rossby radius of deformation. The model runs reported here use $\gamma = f_0/c_0$, where f_0 is the Coriolis parameter at the reference latitude ϕ_0 and c_0 is a specified phase speed, here chosen to be $90\,\mathrm{m\,s^{-1}}$ to match the gravity wave phase speed for the 800-m shallow-water depth in LBAR. Data enter the model through the initial condition (specify q) and boundary conditions (specify ψ on the boundary and q on inflow), using the same wind field as described above for LBAR.

The equations are discretized using conservative second-order centered finite differences in space (Arakawa Jacobian) and the fourth-order Runge–Kutta scheme in time, using an adaptive multigrid method that refines the mesh around the moving vortex. The model can be run in a fully adaptive mode with fine-grid patches being created, moved, resized, and destroyed automatically as dictated by the estimated truncation error of the evolving solution (Fulton 2001). For track forecasts, the fully adaptive mode did not offer any efficiency advantages over fixed-size grid configurations, so this study uses several nested grid patches of fixed sizes that move with the vortex. Likewise, the model can be run with fourth-order space differencing, but as this did not improve the efficiency of track forecasts (for the same accuracy), only the second-order version is used.

3.2.3 Comparison

The principal differences between LBAR and MUDBAR are the dynamics (divergent vs nondivergent) and discretization (spectral vs adaptive multigrid). Beyond this, the parameters of the MUDBAR model are adjusted to make it as similar to LBAR as possible. The model domain for LBAR is a 7200-km square, but the solution is nudged toward the specified environmental flow in the outer portion of the domain. As described above, the magnitude of the boundary nudging term decreases rapidly with the distance from the boundary. Based upon the behavior of the LBAR nudging term, a domain size of a 6000-km square was chosen for MUDBAR. Choosing this domain for MUDBAR gives a domain size approximately 54° longitude by 49° latitude (depending somewhat on the reference latitude

ϕ_0). Both models are run with a fixed domain centered on the initial vortex position, and both define the storm track using the location of the maximum vorticity.

3.3 Model initialization and boundary conditions

The 0000 UTC NCEP global model analysis and 5-day forecast fields were collected for most of the 2001 Atlantic hurricane season for this ensemble study and to compare the LBAR and MUDBAR models. These data were available for all named 2001 Atlantic tropical cyclones except Tropical Storm Allison. Figure 3.1 shows the best tracks for the storms included in the comparison. Unnamed depressions, subtropical, and extratropical cases were excluded from the verification statistics. The active 2001 season provided a variety of storms across the full range of intensities and latitudes typically experienced in the Atlantic basin, making 2001 a good year in which to conduct a robust model comparison. The forecast sample includes 88 cases with at least a 12-h verification.

It should be pointed out that the global model fields used in this study are not exactly the same as those used in the operational LBAR model. This study uses fields from the control member of the NCEP Global Forecasting System (GFS) ensemble, which in 2001 was a T126 run of the MRF model (truncated to T62 at 84 h and thereafter). The operational LBAR model uses fields from a 6-h-old AVN forecast, which had a T170 truncation in 2001. The LBAR model updates the boundaries at 6-h intervals, while the GFS ensemble control data are available only at 12-h intervals. In addition, there are some differences in the data cutoff times and initialization procedures in the GFS ensemble control and the AVN. Perhaps the most significant difference is that in 2001, the AVN used the vortex relocation scheme of Liu et al. (2000), while the GFS ensemble control did not. By properly relocating the model's analyzed vortex to the operational position estimate, the relocation scheme significantly reduces binary interactions between the synthetic and analyzed vortices, resulting in substantial reduction in forecast track errors (Liu et al. 2002). To simplify the comparison between LBAR and MUDBAR, the LBAR model was rerun using the same

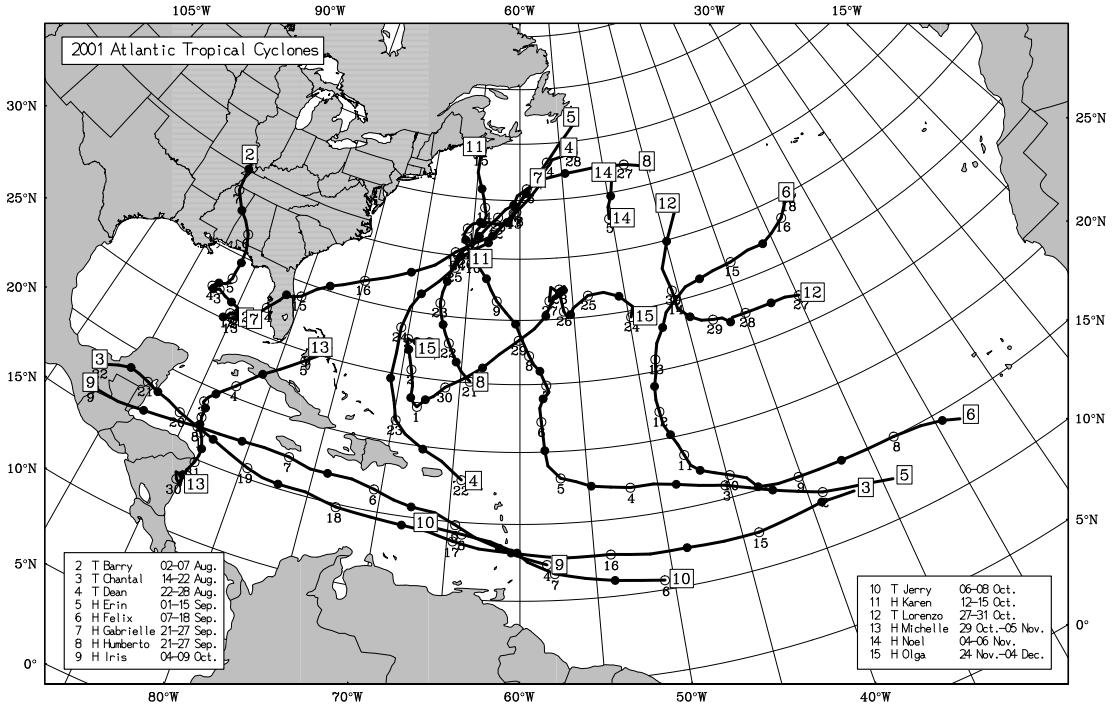


Figure 3.1: Tracks of the fourteen tropical cyclones from the 2001 Atlantic hurricane season included in the model comparison. Numbers in squares identify the storm, and filled and open circles give the storm position at 0000 UTC and 1200 UTC, respectively, on the indicated day. Extratropical and subtropical track segments are not shown.

GFS ensemble control fields that were used in MUDBAR. To distinguish the operational LBAR from the reruns, the version that uses the NCEP GFS ensemble control fields will be referred to as NBAR.

The models are initialized using the velocity field

$$\mathbf{v}_0 = (1-w)\mathbf{v}_{\text{anal}} + w(\mathbf{v}_{\text{vor}} + \mathbf{v}_{\text{cen}}), \tag{3.4}$$

where $\mathbf{v}_{\text{anal}}$ is the analyzed velocity field (from the GFS ensemble control member), $\mathbf{v}_{\text{vor}}$ is the velocity field of the specified (synthetic) vortex, and $\mathbf{v}_{\text{cen}}$ is the specified initial storm motion vector. The weighting function w smoothly blends the synthetic vortex into the velocity field; $w(r) = \exp[-(r/r_b)^2]$ is used, where r is the distance from the vortex center and r_b is the blending radius (here chosen to be 1000 km). For LBAR this velocity is used directly, with the corresponding height field obtained by solving the nonlinear balance equation. For MUDBAR the initial value of q is $\zeta_0 - \gamma^2\psi_0$, where $\zeta_0 = \mathbf{k}\cdot\nabla\times\mathbf{v}_0$ is the initial relative vorticity and ψ_0 is obtained by solving $m^2\nabla^2\psi_0 = \zeta_0$.

For $\mathbf{v}_{\text{vor}}$ we use the symmetric vortex with tangential wind given by

$$V(r) = \frac{rV_m}{r_m}\exp\left\{\frac{1}{b}\left[1-\left(\frac{r}{r_m}\right)^b\right]\right\}, \tag{3.5}$$

with maximum wind V_m at radius r_m, and where b is a size parameter. This profile has been used by DeMaria (1987), Chan and Williams (1987), Fiorino and Elsberry (1989), and DeMaria et al. (1992). The three parameters of this vortex are chosen to approximately match the intensity and size of the observed storm. The initial storm motion vector $\mathbf{v}_{\text{cen}}$ is taken from the operational NHC estimate.

For boundary data, the GFS ensemble control forecasts are spatially interpolated and vertically averaged as described above to obtain time-dependent specified geopotential and velocity fields at 12-h intervals out to 120 h. Both models use time-dependent boundary values, interpolating in time as needed. NBAR uses the GFS ensemble control fields for lateral boundary conditions on the velocity and height, and for the nudging term described previously. MUDBAR constructs boundary values of streamfunction by integrating $\mathbf{v} =$

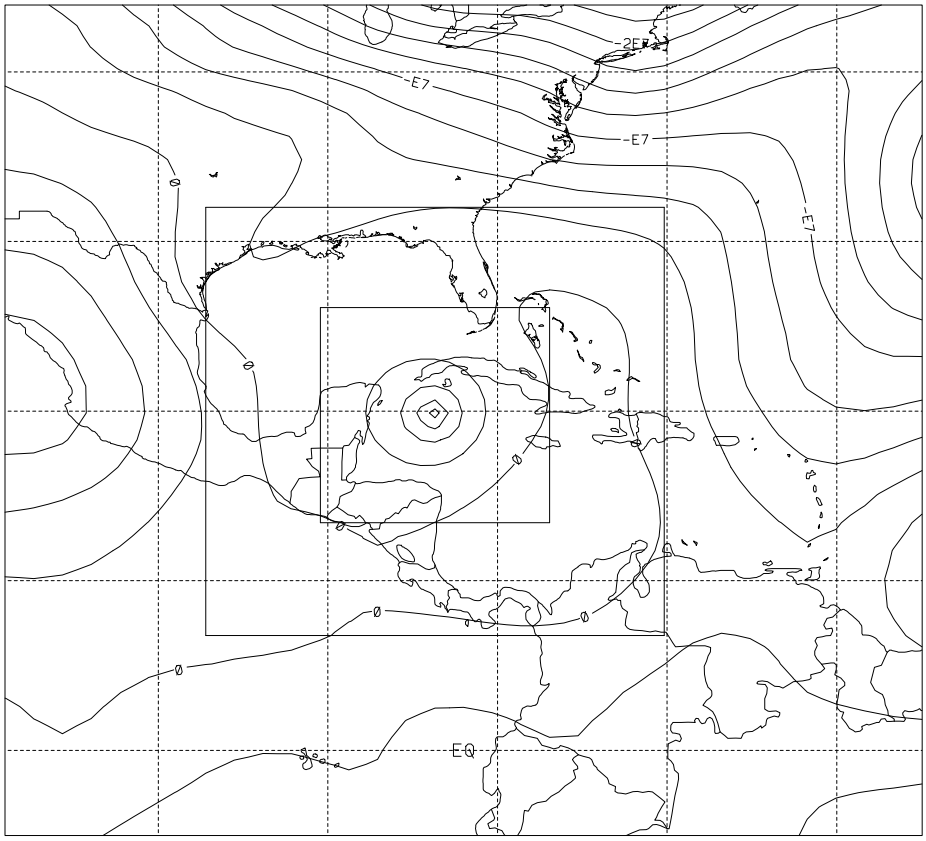


Figure 3.2: Example of the model domain and initial conditions for the MUDBAR model. Contours show the initial streamfunction (for the case of Hurricane Michelle at 0000 UTC on 4 November 2001) and squares show the boundaries of the two fine-grid patches.

$m\mathbf{k} \times \nabla\psi$ around the boundary, and uses the corresponding vorticity where there is inflow.

To illustrate the domain, grid sizes, and initial data, Fig. 3.2 shows a typical initial streamfunction field for MUDBAR. This case is for Hurricane Michelle at 0000 UTC on 4 November 2001, then a category-four storm (on the Saffir–Simpson hurricane scale) centered just south of the western end of Cuba. The details of the selection of the MUDBAR grids are described in the next section.

3.4 Results

In this section, a version of MUDBAR will be developed that provides a reasonable balance between computational cost and forecast accuracy. The forecast accuracy will be determined by defining the track error as the great circle distance between the storm position in the model and the best track position from the NHC best track data (Jarvinen et al. 1984). Once the final version of MUDBAR is chosen, the forecast results will be compared with those from the operational LBAR model, and the LBAR model run using the same fields as MUDBAR. The MUDBAR results will also be compared with some of the other operational track models used by NHC. The various models compared in this section are summarized in Table 3.2.

A statistical t test will be conducted following the method of Franklin and DeMaria (1992) to determine whether model differences are significant. Sample sizes are adjusted to account for serial correlation. The level of statistical significance is taken to be 95%.

3.4.1 Optimization of MUDBAR

The MUDBAR model can be run with any number of nested grid patches of virtually any size. The best configuration of MUDBAR for the barotropic track forecast problem is the one that gives the most accurate forecasts for the least computational cost. To determine the best version of MUDBAR, the 2001 forecast cases were run with a single grid, two grids, and three grids. Each of these grid systems were run with a range of horizontal resolutions. The computational cost was measured as the central processing unit (CPU) time required for a single forecast case on a 1-GHz personal computer (PC). The accuracy was measured by the mean track error (at times 12, 24, 36, and 48 h) over all cases for the 2001 season.

Figure 3.3 shows the accuracy obtained using various numbers and sizes of grids as a function of the cost required. Each single-grid configuration is indicated by an $\times$, each two-grid configuration by a small dot, and each three-grid configuration by a triangle. These results show that mesh refinement is especially advantageous in the short term (e.g., at

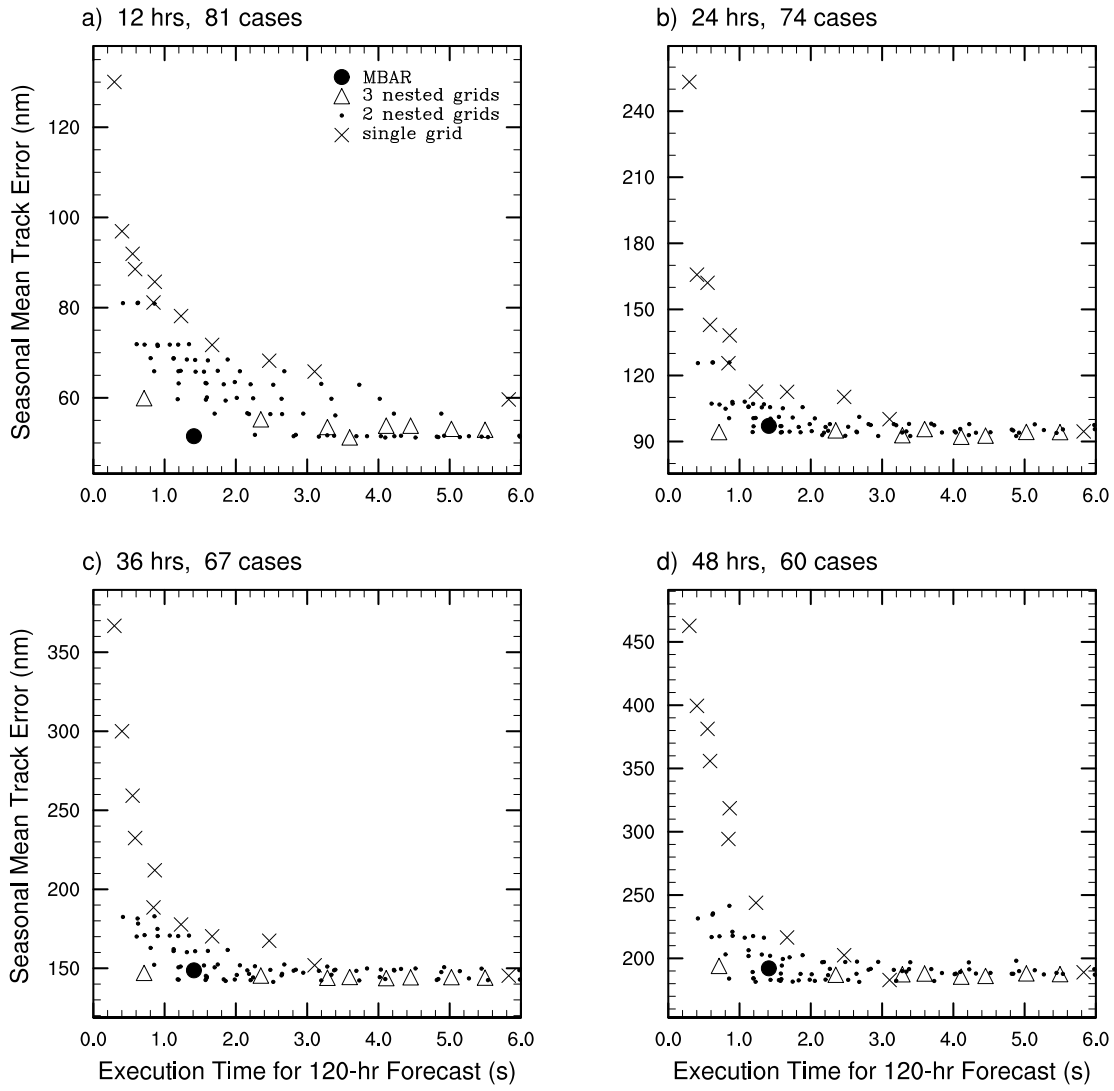


Figure 3.3: Efficiency vs. accuracy for the MUDBAR model with different grid configurations. Single-grid, two-grid, and three-grid configurations are marked by $\times$, small dots, and triangles, respectively. The large dot represents the optimal grid configuration chosen for this paper (MBAR).

Tech	Model	Model type	Domain	Nested?	Boundary data
MBAR	MUDBAR (optimal grid configuration)	nondivergent barotropic	limited area	yes	GFS ensemble (control run)
NBAR	LBAR reruns	divergent barotropic	limited area	no	GFS ensemble (control run)
LBAR	LBAR	divergent barotropic	limited area	no	GFS (Aviation run)
GFDL	GFDL	baroclinic w/physics	limited area	yes	GFS (Aviation run)
NGPS	NOGAPS	baroclinic w/physics	global	no	—
AVN0	GFS (Aviation run)	baroclinic w/physics	global	no	—
A98E	—	statistical-dynamical	—	—	dynamical predictors from GFS (Aviation run)

Table 3.2: Summary of specific model configurations used in error and skill comparisons. All configurations are operational except the first two.

12 h), when accurately capturing the initial position and motion is critical; at later times the advantage is reduced, as other sources of error (uncertainties in the environmental flow, lack of baroclinic effects and physics, etc.) begin to dominate. From this figure, one can also infer a threshold for the minimum resolution necessary to accurately resolve the scales important for the storm track. The optimal grid configuration selected for subsequent runs (indicated by the large solid dot) consists of three grids of size 32×32 grid intervals each,[4] with mesh sizes approximately 194, 97, and 48 km. In this and subsequent results we refer to this configuration of the model as MBAR, which is considered the best choice based upon computation cost and forecast accuracy.

[4] The fact that 32 is divisible only by 2 ($32 = 2^5$) may be a reason why this particular grid configuration performs so well with the multigrid methods.

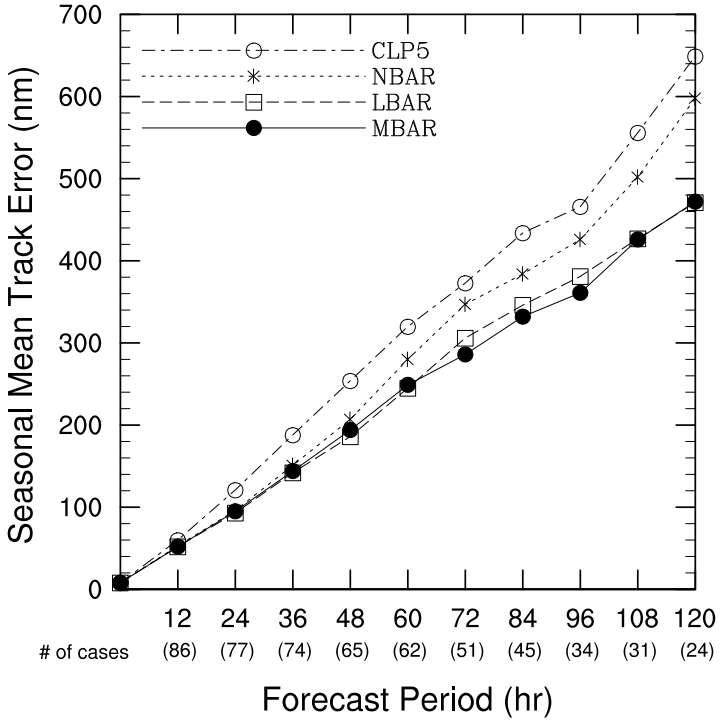


Figure 3.4: Model accuracy for CLIPER (CLP5), the LBAR reruns using the NCEP GFS ensemble control fields (NBAR), the operational LBAR, and the optimal configuration of MUDBAR (MBAR).

3.4.2 Comparison with LBAR and NBAR

Figure 3.4 shows the average track forecast error from the 2001 sample as a function of forecast time for MBAR, NBAR, and LBAR. Also included is the 5-day climatology and persistence (CLIPER) model of Aberson (1998), since it is often used as a benchmark for the evaluation of forecast skill. If a model has average track forecast errors smaller than CLIPER, it is considered to be skillful. MBAR and LBAR showed statistically significant improvement over CLIPER at all time periods, while NBAR showed significant improvement only out to 48 h.

Figure 3.4 shows that the average MBAR errors are comparable to those from NBAR out to about 48 h and are somewhat smaller after that time. The differences between MBAR and NBAR were only statistically significant at 72 h. This result indicates that MBAR is able to reproduce or slightly improve upon the operational barotropic forecast model for the case when both use the same initial and boundary condition information. The differences in the errors between NBAR and MBAR can be attributed to the differences between the two modeling systems (divergent barotropic versus nondivergent barotropic, treatment of boundary conditions, numerical methods, etc). Fig. 3.4 does not show the computational cost of the runs. Each 5-day NBAR run took an average of 99.2 s, while each MBAR run took only 1.4 s. Thus, the LBAR modeling system could be replaced by a system with about 1/70 the computational cost.

Figure 3.4 also shows the results for LBAR that were run in real time. The LBAR errors in Fig. 3.4 are smaller than those for NBAR after 48 h, and comparable to those from MBAR. The differences between LBAR and MBAR were not statistically significant at any time period, but LBAR showed statistically significant improvement over NBAR at 36 h and beyond. Since these forecasts use the exact same modeling system, the differences between LBAR and NBAR must be due to the differences in the initial and boundary conditions. As described previously, LBAR uses a 6-h-old AVN forecast at 6-h intervals for the initial and boundary conditions. Apparently, the use of a 6-h-old model run does not degrade the forecasts. This result is consistent with those shown by Horsfall et al. (1997). The AVN is also run at higher resolution than the GFS ensemble control and makes use of the vortex relocation scheme, which is not included in the fields used to initialize NBAR and MBAR. All of these factors probably contributed to the increased track errors of NBAR compared to LBAR.

3.4.3 Model skill

All three barotropic models in Fig. 3.4 had errors smaller than CLIPER. This result indicates that, despite the use of the very simple barotropic framework, the forecasts were skillful. As a further evaluation of the barotropic modeling system, Fig. 3.5 compares the skill of LBAR and MBAR to some of the other NHC operational models. In this comparison, skill is evaluated by determining the error of each model relative to that of the CLIPER model for a homogeneous sample of cases. If the relative error of a particular model is negative (the errors are less than CLIPER), then the model is considered to be skillful. The sample sizes in Fig. 3.5 are smaller than those in Fig. 3.4 because not all of the operational models ran at every forecast period; a homogeneous comparison excludes such cases, reducing the sample size.

Figure 3.5 includes a representative of each basic type of model used at NHC (see DeMaria and Gross 2003 for a description of all NHC models through 2002). The A98E model is a statistical-dynamical prediction model that uses geopotential heights from the AVN forecast to modify a CLIPER track forecast. The Geophysical Fluid Dynamics Laboratory (GFDL) model is a limited-area, nested baroclinic model that obtains boundary conditions from the AVN. AVN0 is the track forecast obtained by tracking the representation of the storm in the global AVN forecast fields, and NGPS is a similar forecast determined from the Navy Operational Global Atmospheric Prediction System (NOGAPS). This figure shows that all of the models were skillful out to 120 h, except A98E after 84 h. There was little difference between the skill of the LBAR, MBAR, NGPS, and GFDL models, but the AVN0 model appeared to have the most skill after about 24 h. The AVN0 showed statistically significant improvements over all other models at all forecast times except at 12 and 108 h. At the 72-h forecast time, all models except A98E showed statistically significant improvement over CLIPER. At 120 h, the only models showing improvement over CLIPER were the AVN0, GFDL, LBAR, and MBAR models. This result shows that although the barotropic model is very simple, it can still sometimes produce skillful track forecasts that are

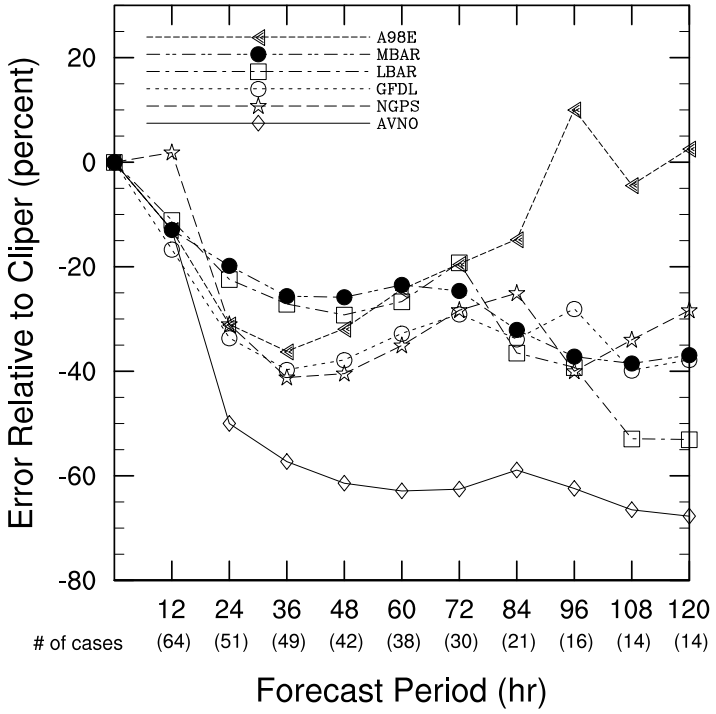


Figure 3.5: Skill relative to CLIPER for a statistical-dynamical model (A98E), two barotropic models (MBAR and LBAR), a full-physics 3-D model (GFDL), and two global models (NGPS and AVN0).

competitive with some of the more general models such as the GFDL model and NOGAPS.

3.5 Conclusions

We conclude that the adaptive multigrid tropical cyclone track model MUDBAR can achieve accuracy similar to the operational shallow-water model LBAR with far less computational work (approximately a factor of 70). Both models show skill (relative to CLIPER) out to five days. The fact that MUDBAR can produce an accurate and skillful forecast in almost negligible computer time (about 1.4 s for a 5-day forecast on a modest PC)

makes it a promising tool for use in an ensemble forecast scheme employing a large number of members perturbed in a multi-dimensional parameter phase space. Such a scheme has been developed and tested as the subject of this thesis (Vigh 2002; Vigh 2004).

The great improvement in computational cost of MBAR relative to LBAR illustrates the utility of the multigrid approach in a simple framework. In this barotropic context there is little need for fully adaptive grids (at least for tropical cyclone track forecasting), so the version of MBAR discussed here simply used fixed-size grids. It is left as a topic for future research to determine how much improvement in computational cost might be obtained by applying fully adaptive multigrid methods to a more general primitive equation hurricane modeling system.

Chapter 4

DESIGN OF THE KILO-ENSEMBLE

4.1 Introduction

The chapter presents the development and implementation of a 1980-member ensemble (aka, the 'kilo-ensemble') for tropical cyclone track forecasting. The ensemble design philosophy will be discussed, as will the selection of perturbation classes and associated magnitudes by means of exploratory sensitivity tests. The resulting ensemble system is implemented and tested in a semi-operational environment similar to other operational TC forecasting aids.

4.2 Design philosophy

This study aims to develop an ensemble-based forecasting system that possesses the following characteristics:

1. The ensemble system should be framed within an operational setting to allow comparison with other operational forecast aids.
2. To facilitate the calculation of spatial probabilities, the system should generate a large number of realizations.
3. It should have a way of correcting systematic biases.
4. All steps (data acquisition, ensemble runs, postprocessing, product dissemination) should be automated with no input required of the user.
5. The entire process should fit within a time-span that allows the ensemble to provide useful operational guidance, using a modest PC.

Most previous ensemble studies of TC track forecasting focused on testing the efficacy of ensemble methods in a research setting, often using analyses derived from atmospheric

field studies and best-track storm data. I feel that the time has come for ensemble methods to be tested in an operational framework so that the ensemble forecasts can be readily compared with other forecast aids and models. With this goal in mind, the ensemble should only use analysis fields that are available to other forecast models or aids at forecast time. Also, the operational estimates of storm position and structure should be used, not the best-track values (which are only available at the end of the season). All input and output data files should follow (or be easily convertible to) the Automated Tropical Cyclone Forecast (ATCF) format used by all U. S. operational tropical cyclone forecast centers. By adhering to these considerations, it should be trivial to adapt the ensemble system for use by an operational TC forecasting center.

No previous ensemble systems made explicit calculations of spatial strike probabilities. While it is possible to *estimate* strike probabilities from a small ensemble (10-50 members), not many studies have done so. This avenue has been followed on an experimental basis by some regional numerical modeling centers, such as NCEP's GFS Ensemble.[1] This work aims to develop an ensemble with at least 1000 members (hence the 'kilo-ensemble' moniker), making possible explicit calculations of spatial strike probabilities. The tremendous efficiency of the MUDBAR model (Chapter 3) makes an ensemble of this size feasible.

Since the ensemble system uses a nondivergent barotropic model, considerable systematic biases are expected. If possible, the ensemble should be formulated in such a way that these biases cancel (in the case of the ensemble mean forecast) or can be removed from the final forecast products (for the spatial probabilities).

The ensemble system should be a complete autonomous forecasting system, capable of producing forecasts with no user input. To achieve this, it should use input from the ATCF-formatted A-decks, which are files that contain the operational estimates of storm position, structure, and motion (these files are produced by a TC forecaster on an ATCF workstation). The system should also ingest all fields needed for the initial condition and

[1] The experimental NCEP strike probability web page is at:
http://www.emc.ncep.noaa.gov/gmb/tpm/emctpc/ens/index.html.

boundary conditions, then run the ensemble. The resulting output should be processed to give output in the ATCF format as well as in a graphical form. The system should then disseminate these products to the forecast users, perhaps via the Internet.

The use of the efficient MUDBAR model (each 120-h forecast only takes 1.4 s) allows a 1980-member ensemble to be run in about 1 h on a 1-GHz PC. Adding in the time for the data ingest and postprocessing, a forecast is available approximately 1.3 h after the analysis and forecast fields become available. These fields are derived from the NCEP Global Forecasting System (GFS) ensemble and are available roughly 9 h after the synoptic time. This means that the kilo-ensemble output becomes available about 10 h after synoptic time, putting it in the 'late-cycle' suite of forecast products. The delay may give the ensemble an artificial advantage over other operational aids, since more information may be available to the ensemble at run-time than was available to the other late aids at the forecast valid time. To do a completely fair comparison, the ensemble should be started from the 6- or 12-h GFS forecast fields (instead of the operational analysis) using the operational information that is available at the time the ensemble is run. This approach was not used in this study, however, due to the complexity involved and the potential for spurious binary interactions due to differences in the vortex location in the 6- or 12-h GFS forecast fields and the operationally-estimated storm location. A truly-operational implementation of the kilo-ensemble should deal with these issues.

4.3 Selection of perturbation generation methodology

The two obvious methods for generating perturbations are dynamically-constrained methods or parameter-based approaches. Two dynamically-constrained methods, the Breeding of Growing Modes (BGM) and Singular Vector Decomposition (SVD), represent the current state-of-the-art ensemble methods used at various regional numerical modeling centers for modeling the global atmosphere (Toth and Kalnay 1997; Molteni et al. 1996). It should be noted that there are other ensemble generation schemes that may prove more efficient

and accurate than the BGM or SVD methods, such as the Ensemble Transform Kalman Filter (ETKF) ensemble forecast scheme (Wang and Bishop 2003), but none of these has been implemented in a global ensemble by an operational forecast center. The dynamically-constrained methods have been applied to the specific problem of TC forecasting by several studies with varying degrees of success (see Chapter 2).

The simpler parameter-based approach varies storm parameters such as position, motion, size, and steering layer depth. Cheung and Chan (1999b) found that dynamically-constrained perturbations generally performed better than simple changes to the vortex parameters. However Zhang and Krishnamurti (1997) point out that a given model may incorrectly suppress important modes due to selection rules that result from its dynamical and physical framework. Thus a dynamically-constrained perturbation approach may not be ideal for a barotropic model. And from a more pragmatic consideration, it is difficult to envision how (or why) one might generate a 1000-member ensemble using dynamically-constrained perturbations, and whether this would shed any more insight than an ensemble with 50 or 100 members. Dynamically-constrained perturbation modes are somewhat inaccessible to a simple analysis: given their complicated spatial characteristics, it is difficult to 'see' where they come from and how they influence the forecast (although such an effort may shed additional light on the nature of these modes). Finally, dynamically-constrained methods have been shunned because of their complexity and potential to add substantial computational cost for an ensemble of this size.

In light of the above goals and design considerations, a simple parameter-based ensemble perturbation method will be used. While less general than dynamically-constrained approaches, the parameter-based perturbations can be systematically formulated in terms of easily measurable (tangible) quantities, such as vortex parameters and the storm motion vector; this simplifies interpretation of the ensemble forecasts. An added benefit derives from the fact that this entire study can be viewed from the perspective of a sensitivity experiment of TC motion in the barotropic framework. Indeed, the forecast output can

be formulated in terms of the parameters used for the perturbations, perhaps giving an operational TC forecaster insight for a particular forecast situation (i.e., a parameter-based ensemble could be used to clarify forecast reasoning by helping a forecaster determine which factors are important for the forecast). For this study, a parameter-based approach is attractive because it requires no additional computation to generate the perturbations. One of the chief questions this study seeks to answer is whether such a parameter-based approach can adequately sample the subspace of dynamical pathways available to a tropical system.

4.4 Selection of perturbation classes

To design an effective ensemble, it is important to choose perturbations that represent uncertainties in the initial analysis or model formulation. Five classes of perturbations are selected for use in the kilo-ensemble: in the background environmental flow, in the depth of the layer mean wind average, in the equivalent phase speed c_{eqv}, in the vortex size/strength, and in the storm motion vector. The choice of each of these perturbation classes is described in depth below.

For TC track prediction, the most important source of uncertainty is probably associated with the evolution of the environment, both near and (at longer time frames) far from the storm. Since a limited-area barotropic model cannot accurately predict the evolution of the midlatitude and tropical environment several days out into the future, the kilo-ensemble uses the analysis and forecast fields from ten ensemble members and control forecast of NCEP's Global Forecasting System (GFS). The GFS ensemble system uses five independent breeding cycles (following the BGM method) during the analysis cycle to estimate the subspace of fast-growing perturbations, which are superpositions of the atmosphere's time-dependent Lyapunov vectors. The fast-growing analysis errors are represented by the bred vectors (Toth and Kalnay 1997). These vectors are then added and subtracted from the control analysis to obtain a total of 10 perturbed initial analyses. Then the 11 analyses are integrated through the forecast cycle to provide 11 unique forecast evolutions

of the storm environment. This information enters the kilo-ensemble by means of the initial condition and time-dependent boundary conditions. Since no other attempt is made to treat the uncertainty relating to the environment, the kilo-ensemble is essentially 'piggy-backing' off of the GFS ensemble in this regard.

The GFS ensemble system is optimized to treat midlatitude weather prediction over the entire globe, so there is no guarantee that the bred modes are suitable for a tropical environment, which may be affected by other TCs, tropical convective systems, and waves. Recently, the feasibility of a new tropical-SVD method has been tested using the ECMWF's ensemble prediction system (Puri et al. 2001), with encouraging results. One of the major conclusions of their study was that the tropical-SVD vectors should be developed from a target area near the TC. If the singular vectors are developed from the entire tropical strip, the perturbed modes are not necessarily in the TC vicinity, causing unsatisfactorily small spread in the ensemble track forecasts. This result suggests that the GFS ensembles may not work well for TC track forecasting, since their perturbations are global in nature. Nevertheless, the use of the GFS ensembles can be justified for this study to the extent that midlatitude systems affect the tracks taken by tropical systems. This is particularly true for the case of recurvature, in which the position of a trough can have a dramatic effect on the track. If the trough remains too far to the north or west of the storm, the storm may continue westward. If the trough digs further south and amplifies, the westerlies often pick up the storm and shunt it off to the north and east. Finally, the GFS ensemble fields are readily available in an operational context, while tropical-SVDs have not yet been implemented; given the lack of any suitable alternative, the GFS ensemble fields are used for this study.

Another source of uncertainty arises concerning the depth of the storm 'steering layer'. The real storm environment may have a complicated shear profile, but a barotropic model operates on only one layer; so the environmental wind information must be reduced to just one layer. The most practical way to do this is to take a pressure-weighted vertical average

over a layer of some depth. Some storms may be strongly coupled to the environment over a deep layer (especially if deep convection is present), so averaging over a deep layer would be appropriate. In situations of strong vertical shear, the storm may decouple from the upper-level winds and move with the lower-level winds, so a shallow layer average might be more appropriate. It is difficult to determine the optimum averaging depth, if one even exists. Velden (1993) found a relationship between optimum steering level depth and storm intensity, which is generally assumed to result from the vertical coupling of the storm's deep convection and the environment. Such deep convection is often associated with intense storms, but it should also be noted that relatively weak storms can also possess strong convection. Since storms can rapidly intensify or weaken, the choice of an optimal averaging depth for a 5-day forecast is even more murky. Thus the kilo-ensemble will use several layer means of different depth to treat the uncertainty associated with the storm steering layer depth.

The nature of nondivergent barotropic dynamics provides yet another source of uncertainty. It has been recognized (Wiin-Nelson 1959) that ultralong Rossby waves experience excessive retrogression in nondivergent barotropic models. The inclusion of the Helmholtz adjustment term (which depends on $\gamma = f/c_{eqv}$) in the governing equation can reduce this spurious retrogression. With f simply a function of latitude, this term depends on the equivalent phase speed $c_{eqv} = \sqrt{gh_{eqv}}$. Decompositions of the tropical atmosphere project onto a variety of vertical modes, ranging from the external and first internal modes (for subtropical highs) to higher internal modes (for flow forced by deep convection). To handle this source of uncertainty, several values of c_{eqv} will be used.

The size and strength of a TC vortex may change greatly during the storm's life cycle. Since a synthetic vortex is used to initialize each ensemble member in MUDBAR, following the method in Chapter 3, vortex parameters must be chosen. It is often unclear which parameters are ideal for a given storm, especially one undergoing rapid intensification or decay, so several vortex sizes/strengths will be simulated by varying the vortex parameter

for maximum wind speed, v_m.

The final perturbation class accounts for uncertainties in the initial storm motion (or persistence) vector. Previous researchers have demonstrated that numerical models are sensitive to the position of the storm in the initial state (DeMaria 1985a; Leslie and Holland 1995), yet there is often considerable uncertainty in the initial storm position, especially when aircraft reconnaissance data are not available for the center fix. Given this uncertainty, it would be logical to perturb the storm's initial position, as has been done in several studies (Zhang and Krishnamurti 1997; Cheung and Chang 1999b). However, Cheung and Chan showed that perturbations to the storm's initial position alone were not very helpful in a barotropic ensemble, but perturbing the vortex asymmetric structure (β-gyres) can give better results. DeMaria et al. (1990) showed that perturbations to the initial position have little effect on the average forecast track error after 12 h, largely because the storm movement is sensitive to much larger scales (1000s of km) than the initial position displacement (100 km) – there is very little variation in the synoptic steering flow on a 100 km scale. Their study showed that perturbations to the initial storm motion vector have a much greater impact on the track.

There is another reason not to perturb the storm's initial position: the possibility of adverse binary interactions between the analyzed and synthetic vortices. This study bogusses a synthetic vortex to the operational position estimate of the actual storm (as described in Chapter 3), but the environmental background flow, derived from the GFS ensemble system also contains an analyzed vortex. Even though the kilo-ensemble uses a weighting function to blend the environmental flow field with the synthetic vortex, a synthetic vortex whose position has been perturbed from the actual operational position could still interact with the analyzed vortex in the environmental flow from the GFS if it is not fully removed. The job of the blending function is to filter out the analyzed vortex and replace it with a more appropriate synthetic vortex. Since the blending function does not attempt to account for the size of the analyzed vortex that will be replaced, it is possible that

large analyzed vortices could leave some residual flow in the initial state used to initialize MUDBAR. If this is the case, then perturbations to the synthetic vortex position could still cause undesirable interactions and affect the track forecast.

Indeed, further mischief may occur due to binary interactions. Even though the analyzed vortex is relocated to the operational position estimate in the GFS control analysis (following the vortex relocation scheme of Liu et al. 2000), the other GFS ensemble members are under no such constraint, except that the breeding of their perturbed initial states begins from the control analysis. In the course of the breeding cycle, the perturbed storm may end up some distance from the operational storm location. Then, even if the synthetic vortex is bogussed to the 'correct' location with no position perturbation, a binary interaction could still be possible, since the analyzed vortex would now be in the incorrect location. Again, this effect will be more pronounced for large vortices in the GFS analysis. No vortex relocation is done for the GFS ensemble members (apart from the control analysis), perhaps because a position perturbation is okay in that framework. In the kilo-ensemble framework however, having the analyzed and synthetic vortices in different locations may cause problems.

In light of these issues, the kilo-ensemble does not use perturbations to the storm initial position, but instead perturbs the storm motion vector. By perturbing the motion vector, the uncertainties in the movement of the storm are simulated. Adding a persistence vector to a symmetric vortex adds asymmetrical structure to the vortex similar to the outer-vortex β-gyres studied by Peng and Williams (1990). These gyres result in a ventilation flow which advects the cyclone, forcing it to go in the 'correct' direction (the prescribed storm motion vector). Without the addition of a persistence vector, it could take up to a couple days for the vortex to adjust to the environmental flow. Cheung and Chan (1999) showed that the presence of β-gyres was helpful in an barotropic ensemble framework, so this effect is included in a roundabout manner, through the inclusion of the prescribed motion vector.

In addition to the storm position, other classes of storm parameters could have been

perturbed in this study, but were not for various reasons. Examples include vortex parameters that control the shape of the vortex's radial wind profile, additional modes of storm asymmetry such as vortex eccentricity, and the storm size parameter. As an example, Leslie and Holland (1995) found that forecasts of TC track were quite sensitive to the vortex profile used for the synthetic vortex. This sensitivity resulted in direct ways, such as the increased beta drift resulting from stronger outer vortex circulation and attendant outer-vortex β-gyres, as well as in indirect ways, such as a profound modification of the surrounding environment by the vortex. Since operational estimates of the radial wind profile or vortex asymmetries are not consistently available in a timely and systematic way,[2] these perturbation classes were not investigated in this study. Other important sources of track uncertainty arise from factors such as storm intensity, but the nondivergent barotropic dynamical framework used in this study does not lend itself to investigation of these factors.

4.5 Selection of parameter magnitudes for each perturbation class

The magnitudes for the ensemble parameters were chosen through a series of sensitivity experiments, as were other model parameters, such as domain size, nesting configuration, mesh size, and the maximum wind permitted, v_{max} (which is used in setting the model's CFL condition). For each perturbation class (except the environmental perturbations, which are not parameter-based), a range of parameter values was used to produce forecasts for 2001 and/or 2002 Atlantic storms using a control version of MUDBAR similar to the configuration of MBAR and LBAR used in the comparison study described in Chapter 3, but varying the parameter of interest. The average track errors and x- and y-biases for the entire season were then compared for each parameter value (the definitions of the mean track error and mean track biases are deferred to the discussion on verification in Chapter 5; see Section 5.3.3 for the definition of track error, and Section 5.4.3 for the defi-

[2] Such wind profiles are now routinely generated in near real-time by scientists from the Hurricane Research Division using aircraft and dropwindsonde data (Powell et al. 1998), but these data sources are only available when storms are close and threatening enough to mandate aerial reconnaissance. Various remote sensing technologies may allow such data to be ubiquitous in the future.

nition of bias and skill). Those parameter values which had relatively small average track errors and reasonable biases with respect to the other values were selected for inclusion in the ensemble. This process was framed within the context of choosing a range of values thought to be representative of the characteristic errors associated with each parameter. Another goal of this process was to construct a 'zero' bias ensemble by choosing the parameter ranges such that the x- and y-biases cancel for a mean forecast based on all members of that parameter class. This ideal is based on the assumption that the biases depend linearly on the parameter of interest. This simplistic argument fails to the degree that the biases do not depend linearly on the given parameter. It would probably be better to construct the various parameter ranges based on the actual error characteristics associated with a given parameter (for instance, use the actual error characteristics of storm motion vector to come up with an appropriate range for the magnitudes and directions of the persistence vector used in the ensemble perturbations). However, this error information is not really known since there are inherent uncertainties even in the best-track data. The use of sensitivity studies to define the parameter ranges bypasses this difficulty and seeks to create a responsive ensemble that is based on some sense of 'reality' with respect to the factors most important for track forecasting.

4.5.1 Perturbations to the environment

For perturbations to the environment, no parameters are used *per se*; however, a test was done to determine the sensitivity of the MUDBAR ensembles to the GFS ensemble member analysis and forecast fields. To conduct this test, forecasts were generated for storm cases from the 2002 Atlantic hurricane season. To determine the error characteristics introduced by MUDBAR, track forecasts were taken directly from the GFS ensemble members (the ensemble tracks are available operationally in the ATCF 'A'-deck files) and compared to the track forecasts of the individual MBAR-type control run based on each respective GFS ensemble member's analysis and forecast wind fields.

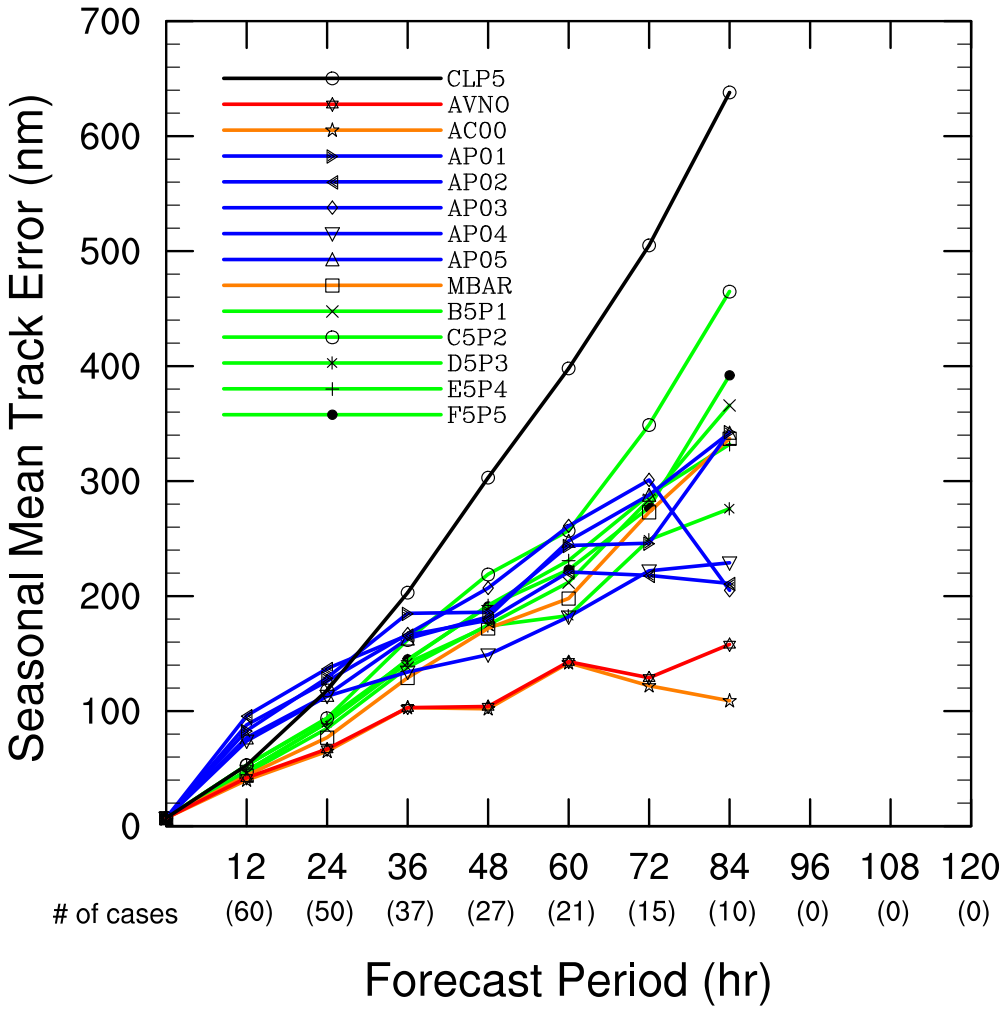


Figure 4.1: Mean track errors for the 2002 Atlantic hurricane season for the positively-perturbed GFS ensemble members (AP01, AP02, AP03, AP04, AP05), control (AC00), and official Aviation model track forecast (AVN0), the MBAR-type configurations of MUDBAR using the initial conditions and time-dependent boundary conditions from the respective GFS ensemble member (B5P1, C5P2, D5P3, E5P4, F5P5), and MBAR which is the run of MUDBAR using the GFS control.

Figure 4.1 shows the mean track errors for positively-perturbed GFS ensemble members[3] (designated by identifiers AP01, AP02, AP03, AP04, and AP05), the control forecast (AC00), and the MBAR-type configurations of MUDBAR run using the initial conditions and time-dependent boundary conditions from the respective GFS ensemble member with the standard (850-200 hPa) deep layer-mean wind (designated as B5P1, C5P2, D5P3, E5P4, F5P5, and the control MBAR). For comparison, the official Aviation model track forecast (AVNO) is also shown, as is the 5-day Climatology and Persistence (CLP5) forecast. A model's forecast is considered to be skillful if its errors are smaller than those of the CLIPER forecast. As expected, the AVNO and AC00 had the best track forecasts for the entire period. Surprisingly, the MBAR-type forecasts have lower errors than their associated GFS ensemble track forecasts out through 36 h. This may be due to the fact that no vortex relocation is used in the GFS ensemble members (although vortex relocation is conducted for the GFS control analysis fields, on which the perturbed fields are bred from), so the vortex is not constrained to start in the 'correct' location. The MBAR-type forecasts bogus the vortex to the operationally-estimated location however; this gives them an advantage which lasts about a day. The errors are of similar size between 36 and 72 h, but by 84 h, the GFS ensemble members have decidedly smaller errors. The small number of cases dictates caution in interpreting the results at 72 and 84 h.

Figure 4.2 is the same as Fig. 4.1, but for the seasonal mean track x-bias, computed by averaging the x component of the error (a positive x-bias corresponds to a track forecast that is east of the observed storm location). The GFS ensemble track forecasts display increasingly wider ranges of x-bias at longer time periods, with 4 of the 5 members displaying a positive x-bias. The MBAR-type runs have x-biases that cluster near zero through 72-h, then trend positive thereafter, but at a slower pace relative to the GFS ensembles. MBAR, AVNO, and AC00 possess consistently small biases. For comparison, CLIPER has an x-bias

[3] The GFS ensemble track forecasts only extend out through 84 h for the 2002 season. This cutoff is likely dictated by the fact that at 84 h, the ensemble fields were degraded to a lower resolution, making accurate vortex tracking difficult).

which becomes quite large and positive after 48 h.

Figure 4.3 is the same as Fig. 4.2, but shows the seasonal mean track y-bias (positive y-bias corresponds to a forecast that is north of the observed storm location). The y-bias of all models is quite small (less than 30 n mi) through 48 h. After that, the biases are mostly positive through 72 h, then trending towards negative values at later forecast periods. Again, care should be taken in interpretation of these results at the longer forecast periods, given the small number of cases.

Figure 4.4 is just like Fig. 4.1, but for the negatively-perturbed ensemble members. Likewise, Fig. 4.5 and Fig. 4.6 show the x-bias and y-bias for the negatively-perturbed ensemble members. The results are generally similar as expected, with some differences at the longer forecast periods, which probably reflects the decreased robustness of the error and bias statistics due to small sample size.

Although one might expect the GFS ensemble perturbations to have similar mean track error characteristics over the course of an entire season, substantial differences were apparent. Since a breeding method is used to generate the GFS perturbations, it is difficult to explain why some members are more skillful than others, given that each bred mode is assumed to be fairly independent after several days (i.e., a bred mode loses 'memory' of its previous incarnation after several days). It might be possible for some slow-growing modes to be retained in the analysis and perturbation system, which could then cause adverse or beneficial effects for the forecasts of a given mode. In extreme cases (84-h track errors greater than 500 n mi are sometimes observed for several days in a row for a particular GFS ensemble member), these effects might skew the statistics for an entire season. If the GFS ensemble track forecasts show significant autocorrelation, especially in the case of larger errors, then this could be evidence that the modes are not independent from breeding cycle to breeding cycle. Otherwise, the differences may simply be due to sampling error.

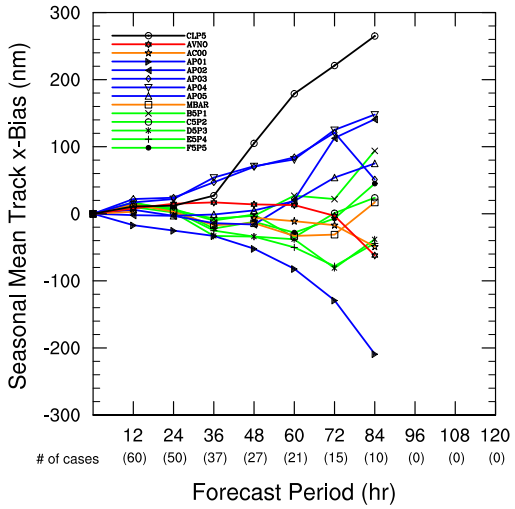


Figure 4.2: As in Fig. 4.1 but for the mean track x-bias.

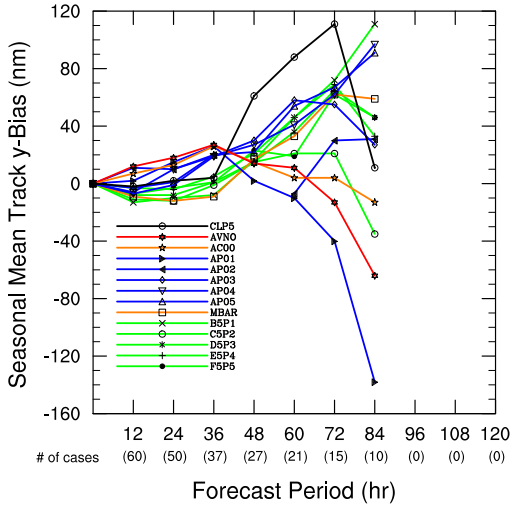


Figure 4.3: As in Fig. 4.1 but for the mean track y-bias.

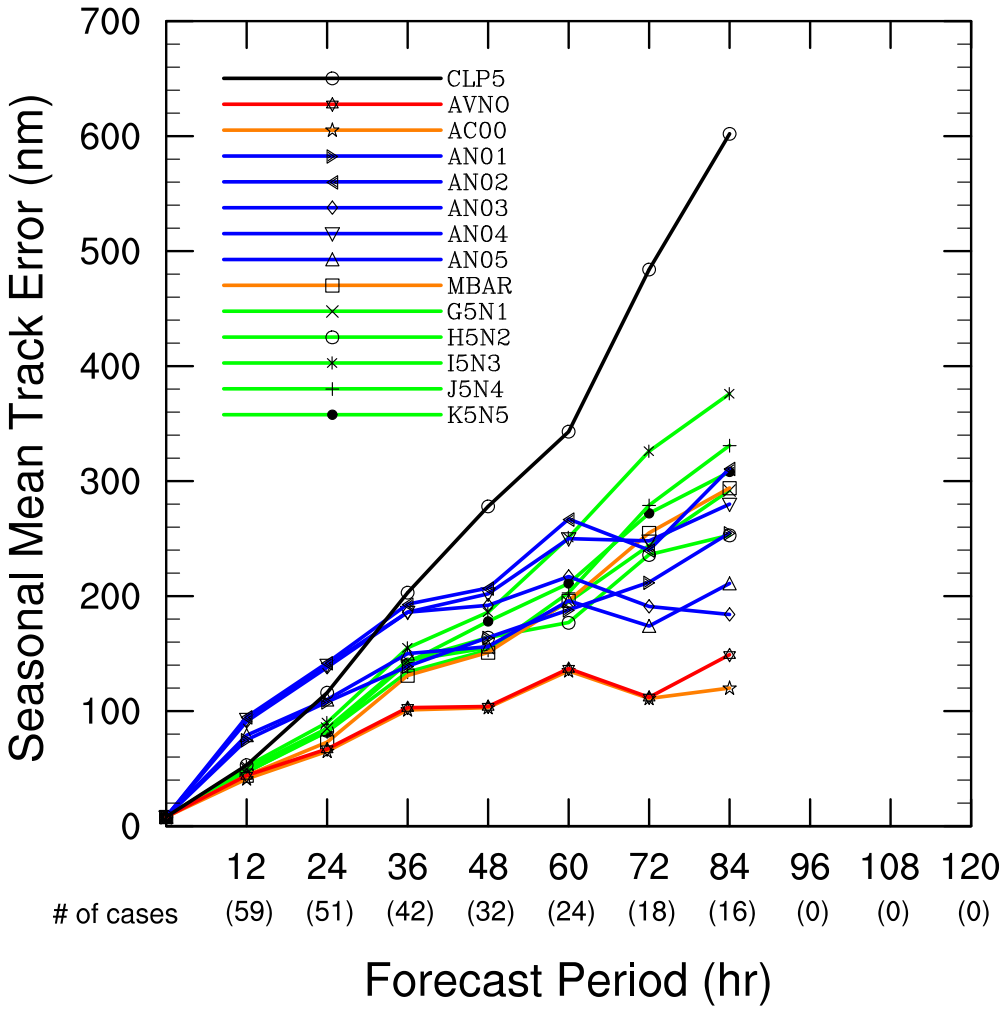


Figure 4.4: Mean track errors for the 2002 Atlantic hurricane season for the negatively-perturbed GFS ensemble members (AN01, AN02, AN03, AN04, AN05), control (AC00), and official Aviation model track forecast (AVN0), the MBAR-type configurations of MUDBAR using the initial conditions and time-dependent boundary conditions from the respective GFS ensemble member (G5N1, H5N2, I5N3, J5N4, K5N5), and MBAR which is the run of MUDBAR using the GFS control.

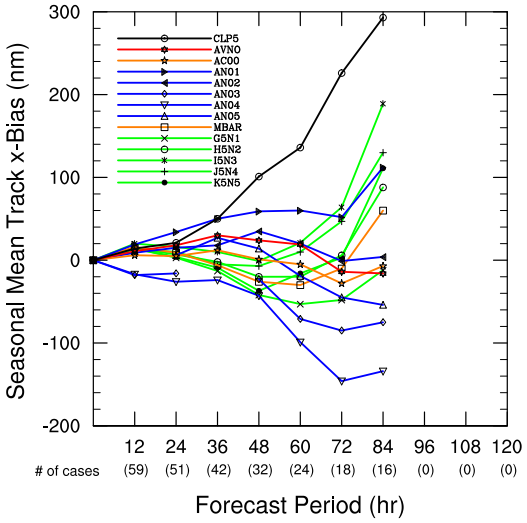


Figure 4.5: As in Fig. 4.4 but for the mean track x-bias.

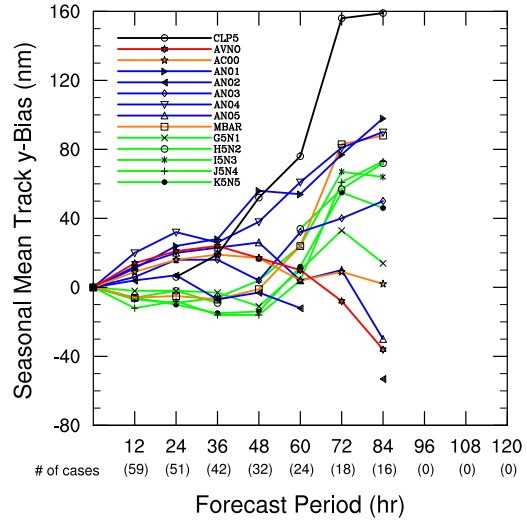


Figure 4.6: As in Fig. 4.4 but for the mean track y-bias.

4.5.2 Perturbations to the depth of the 'steering layer'

A limited sensitivity experiment was conducted for the choice of appropriate averaging depths for the deep layer-mean wind. Given the large amount of data and processing required to generate the deep layer-mean wind fields, just four were chosen. Standard pressure-weighted averages are used, such as the standard deep layer mean (DLM5: 850-200 hPa), a medium depth layer mean (DLM6: 850-350 hPa), and a shallow depth layer mean (DLM7: 850-500 hPa). In addition, a fourth experimental very deep layer-mean wind (DLM4: 1000-100 hPa) was also added, resulting in four deep layer means in this perturbation class. The idea behind the very deep layer mean is to use all the wind information from all levels when creating a mean wind for the barotropic ensemble. Although it may be dubious to assume that the winds near the surface (1000 hPa) and in the lower stratosphere (100 hPa) will be relevant to the track problem (at least for a barotropic model), it might provide useful information. This very deep layer (1000-100 hPa) is the traditional deep layer mean used by NHC's statistical models.

Figure 4.7 shows the seasonal mean track errors for 2001 Atlantic storms for a MBAR-type configuration of MUDBAR using various deep layer-mean winds. The MBAR run that uses the standard deep layer (DLM5) has a mean seasonal track error which grows steadily towards the end of the forecast period, reaching 289 n mi by 72 h and 496 n mi by 120 h. The forecasts based on the very deep layer mean (DLM4) are not quite as accurate, but comparable. The errors of the runs using the medium (DLM6) and shallow layer (DLM7) means are considerably worse, reaching 611 and 693 n mi at 120 h. Interestingly, the rate of error growth levels off at 108-h for all but the standard deep layer mean. This is likely due to a sampling issue since the number of storms in the verification decreases towards the latter forecast periods (many storms transition to extratropical cyclones or dissipate entirely before the end of any given forecast period). For instance, one case of a poorly-forecast storm that had a verifying position at 108 h, but not 120 h, could skew the statistics quite readily if the number of cases with a 108-h verification is already small. If the storm dissipated

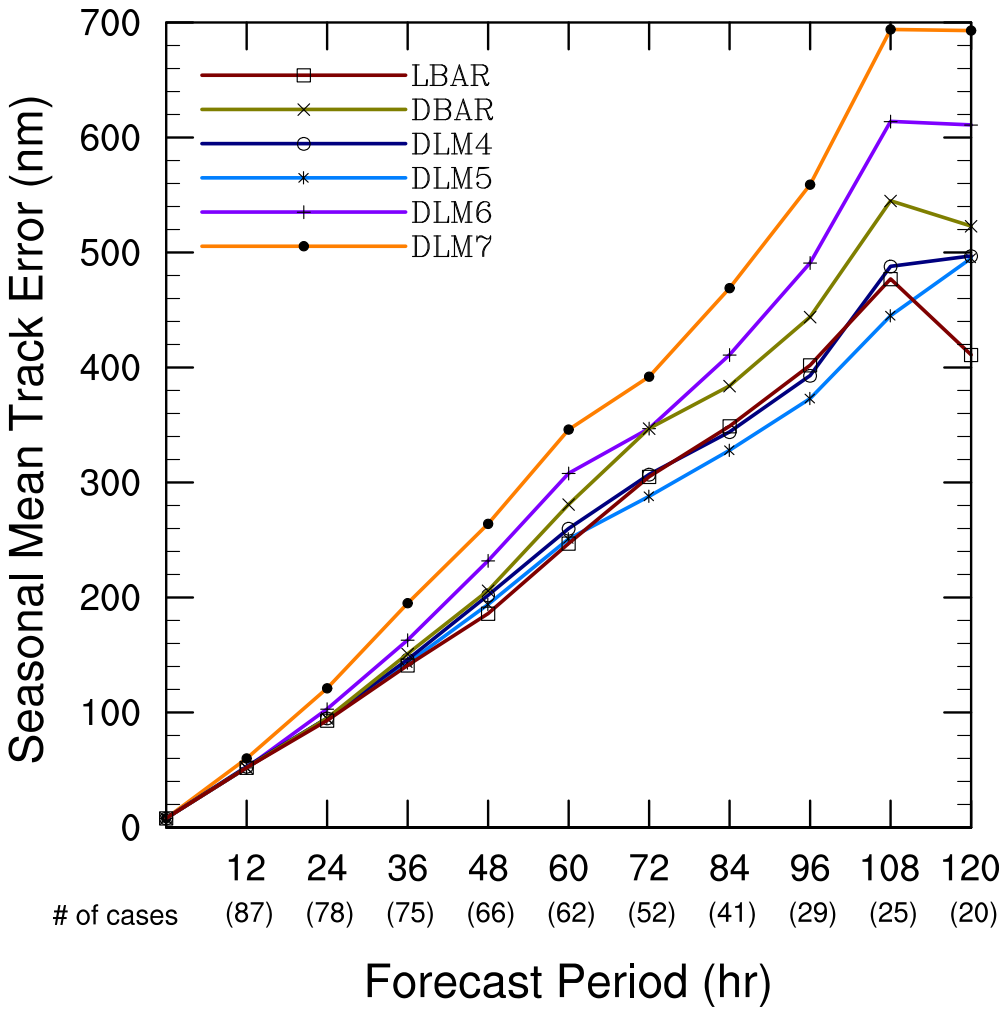


Figure 4.7: Mean tracks errors for the 2001 Atlantic hurricane season for various deep layer-mean winds in a MBAR-type configuration of MUDBAR. LBAR indicates the operational shallow-water equation model, DLM4, DLM5, DLM6, and DLM7 indicate the MBAR-type configurations using a very deep layer mean (1000-100 hPa), standard deep layer mean (850-200 hPa), medium-depth layer mean (850-350 hPa), and a shallow-depth layer mean (850-500 hPa), respectively.

before the 120-h forecast verification time, the error statistics for the 120-h period would not be affected, causing it to appear to have a lower mean error than the 108-h forecast period. This is one example of how small sample sizes can make difficult the interpretation of error statistics based on only one season.[4]

Figure 4.8 is as Fig. 4.7, but for the seasonal mean track x-bias. All the MBAR runs have a westward track bias, while LBAR and DBAR have an eastward bias. Since both models use the same storm data and environmental data, this difference must be related to differences in their dynamics (nondivergent vs. shallow-water), their numerics (multigrid methods vs. cubic B-splines), or in how the models treat the boundary conditions.

Figure 4.9 is as Fig. 4.7, but for the seasonal mean track y-bias. At intermediate time periods, all runs exhibit a slight to moderate northward bias. The standard deep layer mean has the least bias through 72 h, followed by the very deep layer mean, LBAR, and the medium layer mean. The shallow depth-layer runs and DBAR have a significant northward bias by 72 h. Interestingly, all runs trend toward a more southward bias by 120 h. The stratification remains essentially the same, with the standard deep layer mean and very deep layer mean developing the largest southward bias; DBAR has a somewhat reduced northward bias.

4.5.3 Perturbations to the equivalent phase speed parameter

To determine the sensitivity to c_{eqv}, the model was run with a variety of phase speeds ranging from 30 to 210 $\mathrm{m\,s^{-1}}$. Figure 4.10 shows the mean track errors for the 2001 Atlantic hurricane season using a configuration similar to the optimized version of MUDBAR (MBAR), but with values of c_{eqv} ranging from 30-210 $\mathrm{m\,s^{-1}}$. For comparison, the forecasts from DBAR (a shallow-water equation model that uses an equivalent depth of 800 m, which has an equivalent phase speed of 90 $\mathrm{m\,s^{-1}}$ are included. In general, as the equivalent phase speed approaches infinity (γ goes to zero), the track errors asymptotically

[4] Obviously, it would be better to use statistics based on multiple seasons to get a larger sample size, but data was only available for one season at the time this analysis was done.

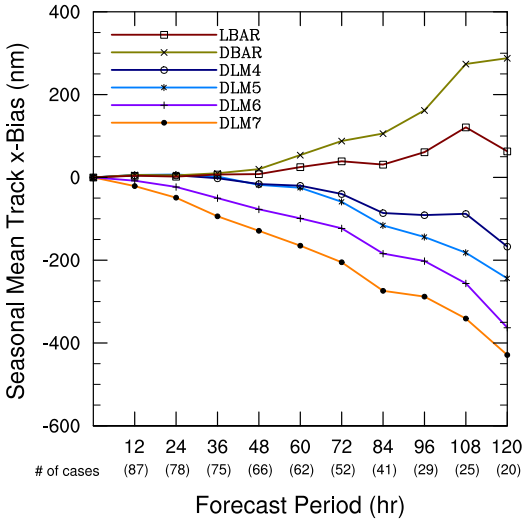


Figure 4.8: As in Fig. 4.7 but for the mean track x-bias.

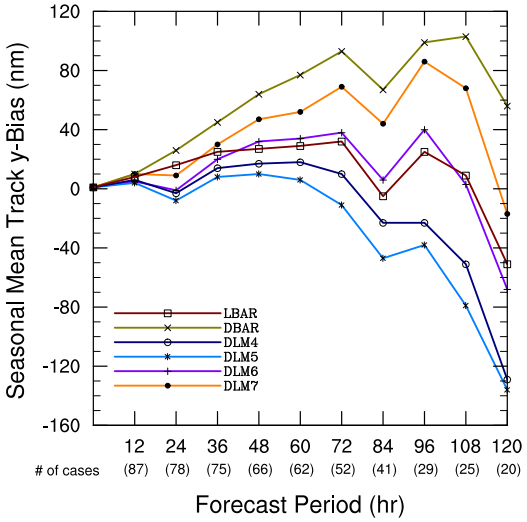


Figure 4.9: As in Fig. 4.7 but for the mean track y-bias.

approach a lower limit. The track errors for DBAR grow close to MBAR's high-c_{eqv} limit through 84 h, but then achieve lower errors thereafter. This may be due to the fact that DBAR's shallow-water equation formulation allows a better simulation of the large-scale subtropical highs, which are often important steering features for tropical cyclones.

Figure 4.11 shows the average seasonal track x-bias for the 2001 Atlantic storms for the various values of c_{eqv}. The forecasts that use a slow c_{eqv} have a large westward bias that grows in time, while the runs with faster c_{eqv} have smaller but still substantial westward biases. As in Fig. 4.10, asymptotic behavior results as c_{eqv} trends towards infinity. For comparison, DBAR has very little x-bias until 108 h.

Figure 4.12 shows the average seasonal track y-bias. The MBAR run using the slowest equivalent phase speed shows a steadily increasing southward bias in time, while the runs using c_{eqv} of 70 m s^{-1} and higher are tightly clustered, having a negligible y-bias through 84 h, then a growing southward bias afterward. DBAR has small northward bias at intermediate times, then a southward bias by 120 h.

Interpretation of seasonal statistics of track x- and-y-biases is complex due to the typical life cycle of tropical cyclones. Many Atlantic tropical cyclones undergo recurvature and extratropical transition near the end of their life. During this stage, the storms are often impacted by significant westerly flow, often in a high-shear environment. Barotropic models may not capture the rapid movement that ensues when a storm is caught up by the westerlies (probably because they use layer-mean winds). This may explain the large south and westward biases exhibited by barotropic models, especially at later forecast periods when storms are more likely to be nearing the end of their lives.

The above explanation may be too simplistic to sufficiently explain the large south and westward biases, so here is a more complex explanation that invokes the effect of equivalent phase speed. It seems obvious that the most appropriate choice of an equivalent phase speed for a deep convective system like a TC would be a slow c_{eqv} (high γ) that corresponds to the atmosphere's higher internal modes. However, TC's are often steered

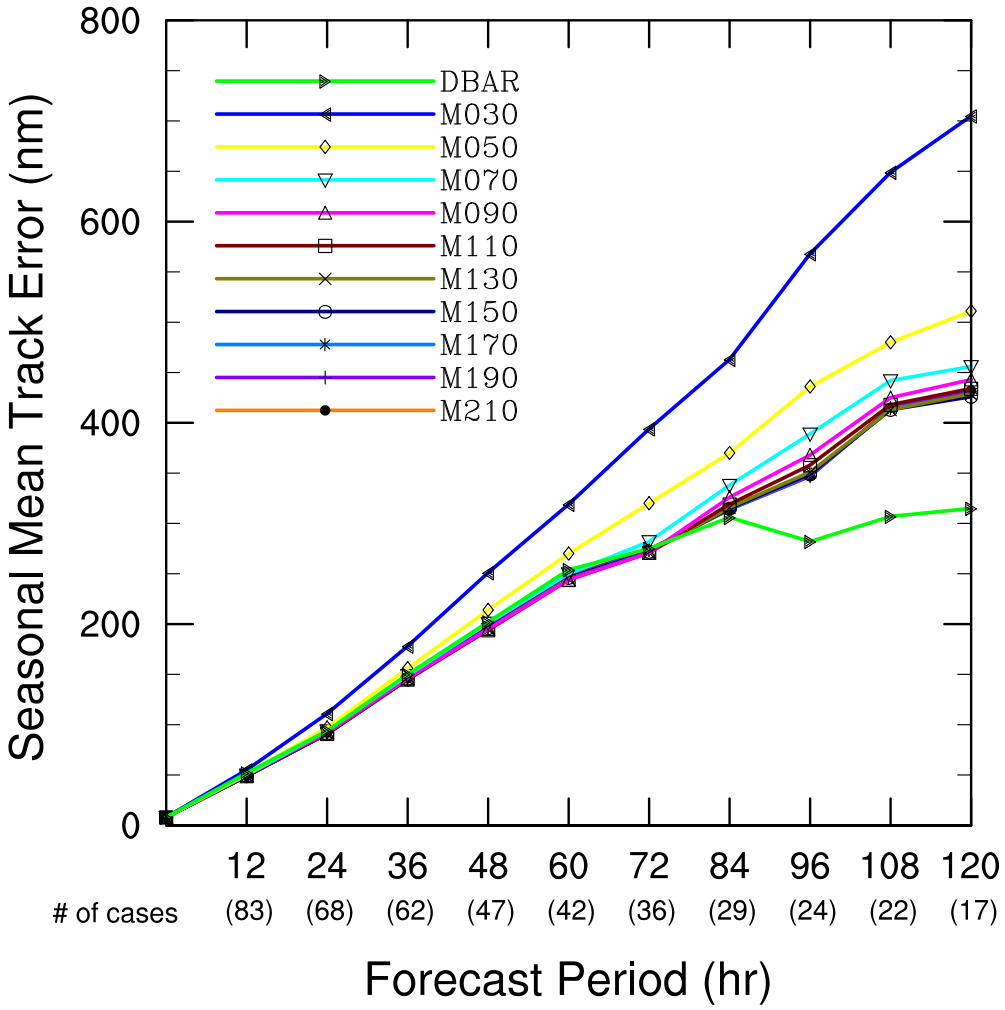


Figure 4.10: Mean tracks errors for the 2001 Atlantic hurricane season for various values of c_{eqv} in a MBAR-type configuration of MUDBAR. LBAR indicates the operational shallow-water equation model, while M030, M050, M070, ..., M210, indicate the MBAR-type configurations using equivalent phase speeds of 30, 50, and 70 $\mathrm{m\,s^{-1}}$ respectively, up to 210 $\mathrm{m\,s^{-1}}$.

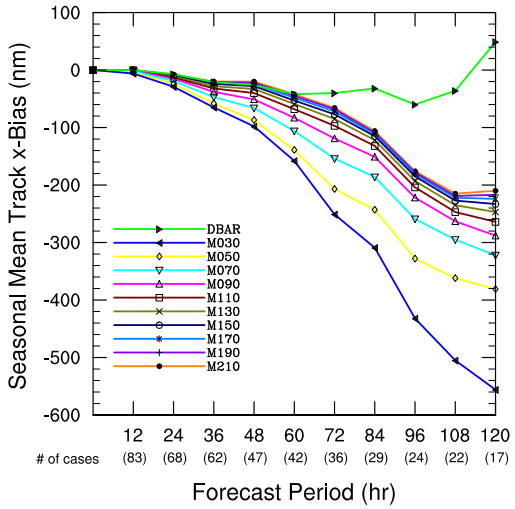


Figure 4.11: As in Fig. 4.10 but for the mean track x-bias.

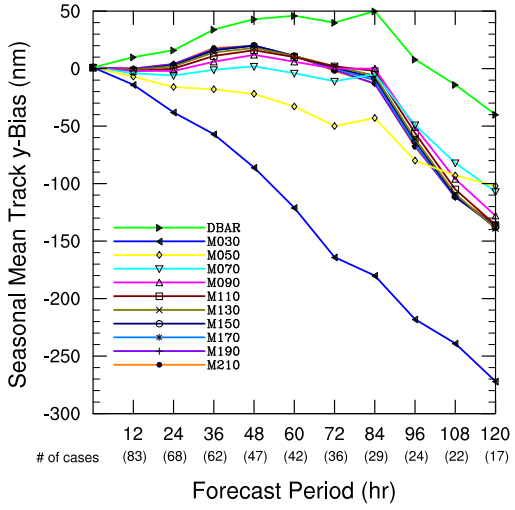


Figure 4.12: As in Fig. 4.10 but for the mean track y-bias.

by the flow resulting from large-scale subtropical ridges which can dominate over all other steering mechanisms. These subtropical highs project onto the external and first internal modes of the real atmosphere, so if the model uses a slow c_{eqv} that is appropriate for deep convection, the large-scale subtropical features retrograde too fast in the model. This allows the model storm to move further westward and southward (or, not as far to the north and east) than a storm in the real atmosphere. In addition, this bias trend can be exacerbated if the model storm does not travel poleward enough to be caught up by the westerlies. This failure to recurve may explain the sharp southward and westward biases for Atlantic storms after 84 h for the low c_{eqv} runs. The results from this sensitivity study suggest caution in the use of a barotropic model for forecasts beyond 84 h. Since this study is investigating the efficacy of ensemble methods out to 120 h, where it is believed that such methods may offer substantial benefits, it may be appropriate to try to remove these severe biases at the later forecast periods so that the ensemble's spatial probabilities are not skewed southward and westward.

In light of these results, c_{eqv} values of 50, 100, and 300 $\mathrm{m\,s^{-1}}$ were chosen to simulate the uncertainty associated with the projection of the tropical atmosphere onto features relevant for tropical cyclone track forecasting. The 30 $\mathrm{m\,s^{-1}}$ value was shunned in light of its extremely large biases. The upper value of 300 $\mathrm{m\,s^{-1}}$ was chosen since this is close to the atmosphere's external mode. Although it was not tested here, the asymptotic behavior of the previous figures suggest that it should be representative of the near-zero γ-limit, corresponding to the atmosphere's deepest vertical modes.

4.5.4 Perturbations to the vortex maximum wind speed parameter

The maximum wind speed parameter, v_m, controls the strength of the entire vortex perturbation wind field, and therefore the size of the bogus vortex. For barotropic track prediction, the size of the outer vortex is what really matters, not the vortex maximum wind speed. The inner core is usually not resolved in barotropic track models, so the maximum

wind speed is moot. The size of the outer vortex, however, largely determines the strength of the β-gyres, which in turn can have a significant impact on the ultimate storm motion (DeMaria 1985; Fiorino and Elsberry 1989; Peng and Williams 1990; Leslie and Holland 1995). The inclusion of this parameter is motivated by the variety of storms that must be dealt with in an operational framework. Not only are there weak and strong storms to forecast, but storms often undergo considerable intensity change during the forecast period. Also, storms often grow in size during their lifetimes. To account for these uncertainties, three perturbations to v_m are chosen: 15, 30, and $50\,\mathrm{m\,s^{-1}}$. These maximum wind speeds correspond to a minimal tropical storm, cat. 1 hurricane, and a cat. 4 hurricane, respectively. In reality, the barotropic model represents these three storm strengths as small, medium, and large storms, respectively. Figure 4.13 shows the radial wind profiles that correspond to these values of v_m for Hurricane Isabel at 0000 UTC on 17 September 2003, using the vortex parameters $b = 0.65$, $r_m = 101.8\,\mathrm{km}$, and $v_m = 37\,\mathrm{m\,s^{-1}}$.[5] To determine the effect of perturbing v_m on the mean track error and x- and y-biases, a sensitivity study is now conducted.

Figure 4.14 shows the 2001 Atlantic seasonal mean track errors for configurations of MBAR that use various values of v_m. The runs that use a small (weak) vortex have larger track errors than those with large (strong) vortices, which is no surprise, since a 10 or $20\,\mathrm{m\,s^{-1}}$ vortex is probably not very realistic for many of the cases. The optimal MBAR configuration that uses the operationally-estimated v_m multiplied by a reducing factor gives good results, as does the runs which used a $30\,\mathrm{m\,s^{-1}}$ bogus vortex.

Figure 4.15 shows the 2001 Atlantic seasonal mean track forecast x-biases for various configurations of MBAR that use different values of v_m. The simulations using small and medium-sized vortices have low x-biases through about 60 h, but then grow increasingly negative through the end of the forecast period, indicating that their forecast tracks are too far to the west. Runs with a larger (stronger) vortex show a similar steady decline in x-bias

[5] The actual storm had a sustained maximum wind speed of $47\,\mathrm{m\,s^{-1}}$ (95 kt) at this time, but the LBAR/MBAR system uses a v_m that is 80% of the actual maximum sustained winds.

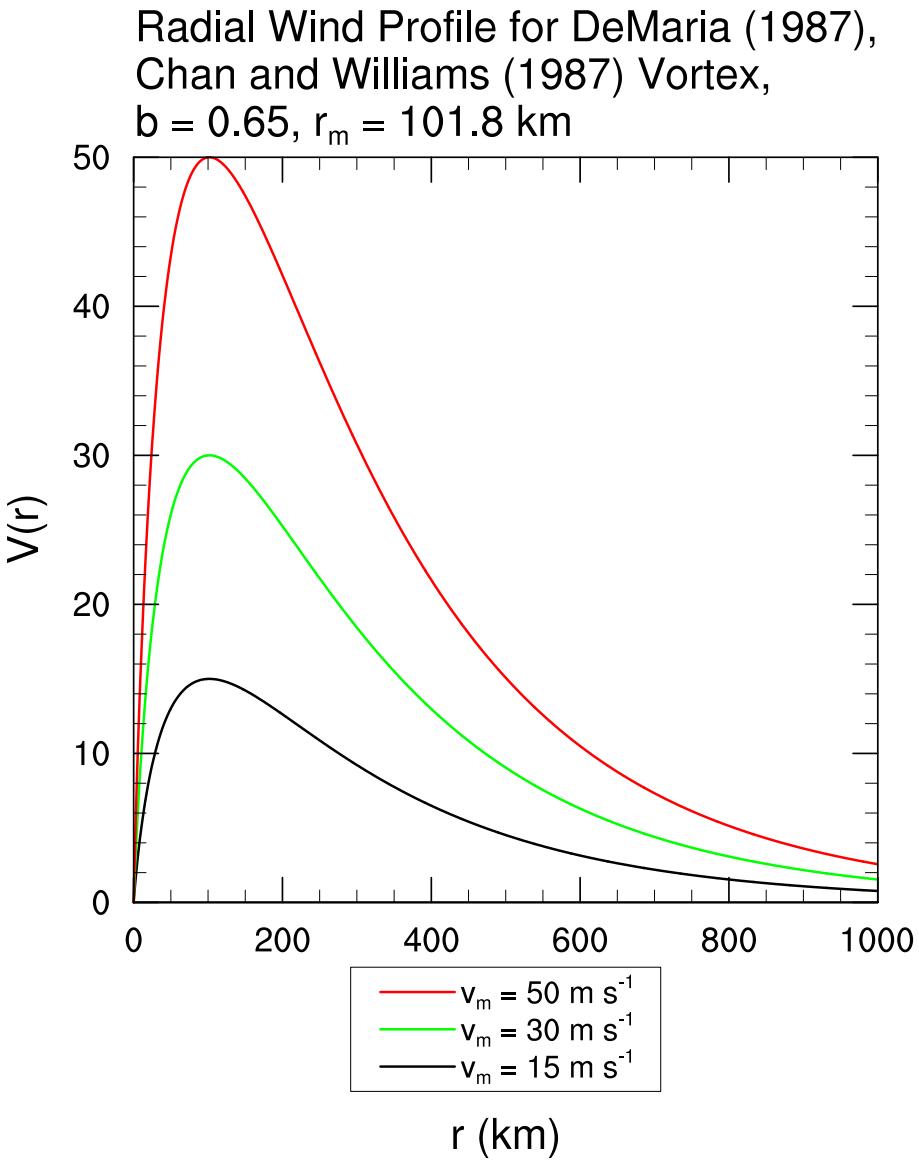


Figure 4.13: Radial vortex profile for the ensemble's three representations of Hurricane Isabel at 0000 UTC on 17 September 2003. The actual storm had a maximum sustained wind of $47\,\mathrm{m\,s^{-1}}$ (95 kt), which corresponds to a v_m of $37\,\mathrm{m\,s^{-1}}$ in MBAR, which uses a value that is 80% of the operationally-estimated maximum sustatined wind. Shown are the DeMaria (1987) and Chan and Williams (1987) profiles for the three v_m perturbations: 15, 30, and $50\,\mathrm{m\,s^{-1}}$, which correspond to actual storm maximum sustained winds of $18.8\,\mathrm{m\,s^{-1}}$ (36 kt), $37.5\,\mathrm{m\,s^{-1}}$ (73 kt), and $62.5\,\mathrm{m\,s^{-1}}$ (121 kt), respectively.

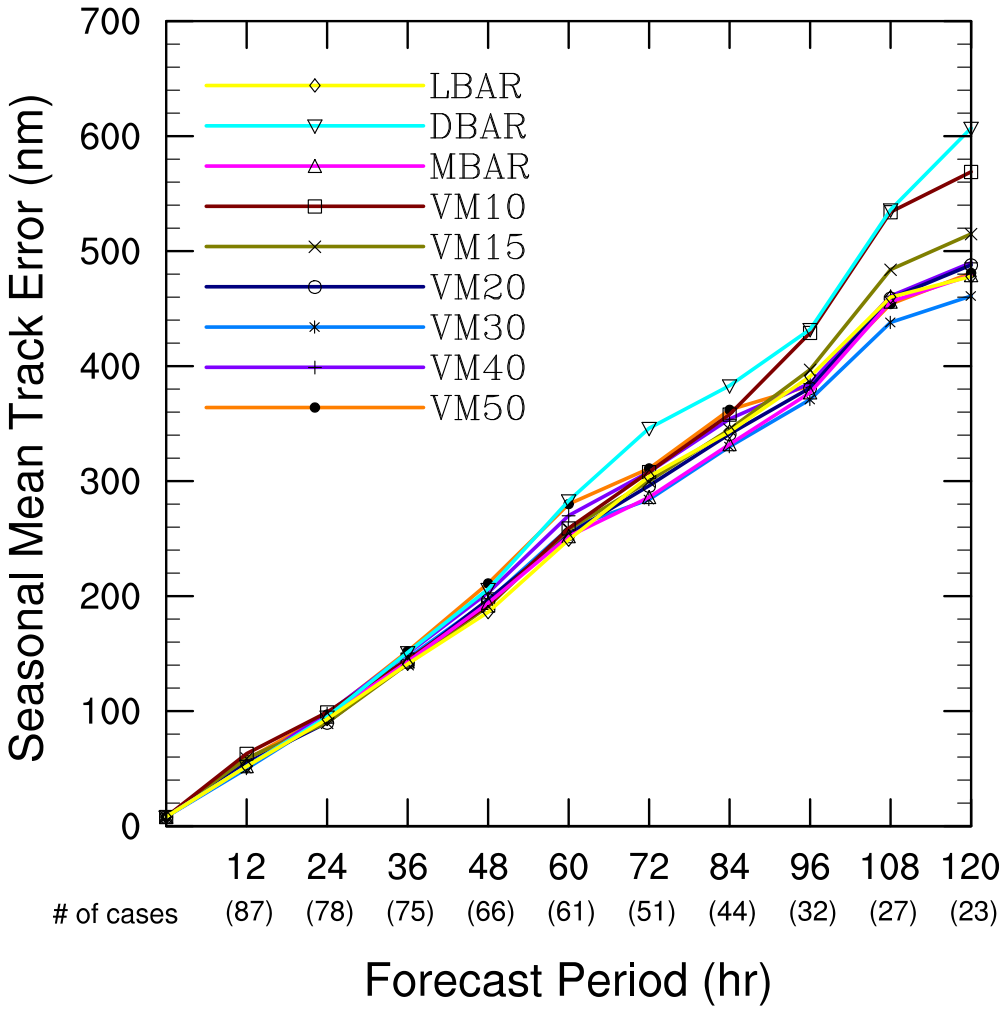


Figure 4.14: Mean tracks errors for the 2001 Atlantic hurricane season for vortices of various sizes in a MBAR-type configuration of MUDBAR. LBAR indicates the operational shallow-water equation model, while VM10, VM15, VM20, ..., VM50, indicate the MBAR-type configurations using a v_m of 10, 15, and 20 $\mathrm{m\,s^{-1}}$ respectively, up to 50 $\mathrm{m\,s^{-1}}$. MBAR indicates the optimal MBAR configuration using the operationally-estimated v_m multiplied by a reduction factor.

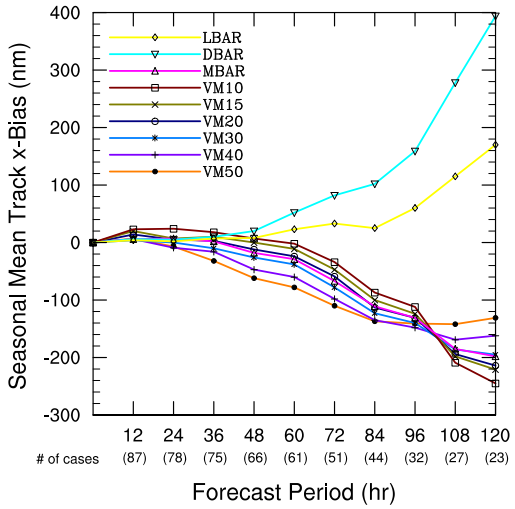


Figure 4.15: As in Fig. 4.14 but for the mean track x-bias.

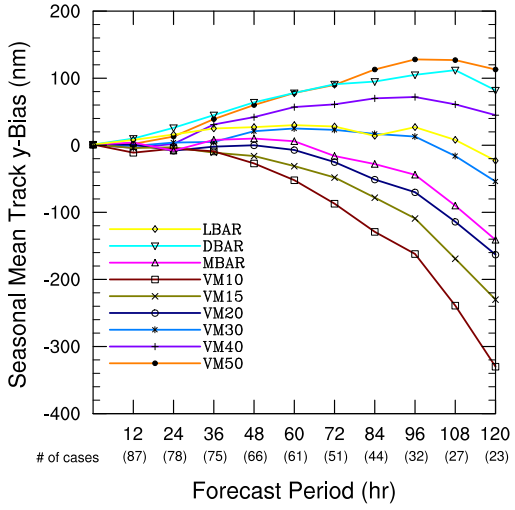


Figure 4.16: As in Fig. 4.14 but for the mean track y-bias.

at intermediate forecast periods, then level off at the end of the forecast period.

Figure 4.16 is the same as Figure 4.15, but for the seasonal mean track y-bias. A wide spectrum of behavior is evident. The largest vortex simulations ($50\,\mathrm{m\,s^{-1}}$) possess a northward bias, while the medium-sized vortex runs ($30\,\mathrm{m\,s^{-1}}$) show almost no y-bias. The small vortex runs show a large southward bias which grows significantly after 72 h. For comparison, the optimal MBAR has little y-bias through 60 h, then has an increasingly southward bias through the end of the forecast period. LBAR possesses little y-bias. This behavior can be explained by the fact that larger vortices possess stronger β-gyres, which add a more significant westward and poleward contribution to the storm motion vector. During the early and intermediate forecast periods, this causes the large (strong) vortices to move too far to the north and west. Unrealistically small (weak) vortices will not have this effect and end up too far to the south. In some cases, the small vortices will be too far west and south to undergo recurvature in cases when the real storm does recurve. This nonlinear effect may explain the significant degradation in both x- and y-biases after 96 h.

4.5.5 Perturbations to the storm motion parameter

Finally, experiments were conducted to determine the sensitivity of the storm track to the motion vector. The goal was to pick a value representative of the uncertainty of the initial motion estimate and then perturb the vortex motion vector in the four cardinal directions by differing amounts. For this sensitivity experiment, perturbations of 1 and $2\,\mathrm{m\,s^{-1}}$ to the north, east, south, and west are added to the operationally-estimated storm motion vector, and the resulting seasonal mean track errors and x- and y-biases are examined.

Figure 4.17 shows the seasonal mean track errors for the 2001 Atlantic storms for fixed cardinal perturbations of both 1 and $2\,\mathrm{m\,s^{-1}}$ to the operationally-estimated motion vector. The track forecasts do not exhibit much sensitivity, at least in the seasonal track error statistics. The perturbations to the west and south of the motion vector have slightly larger errors at later forecast periods, with the $2\,\mathrm{m\,s^{-1}}$ perturbations possessing the largest

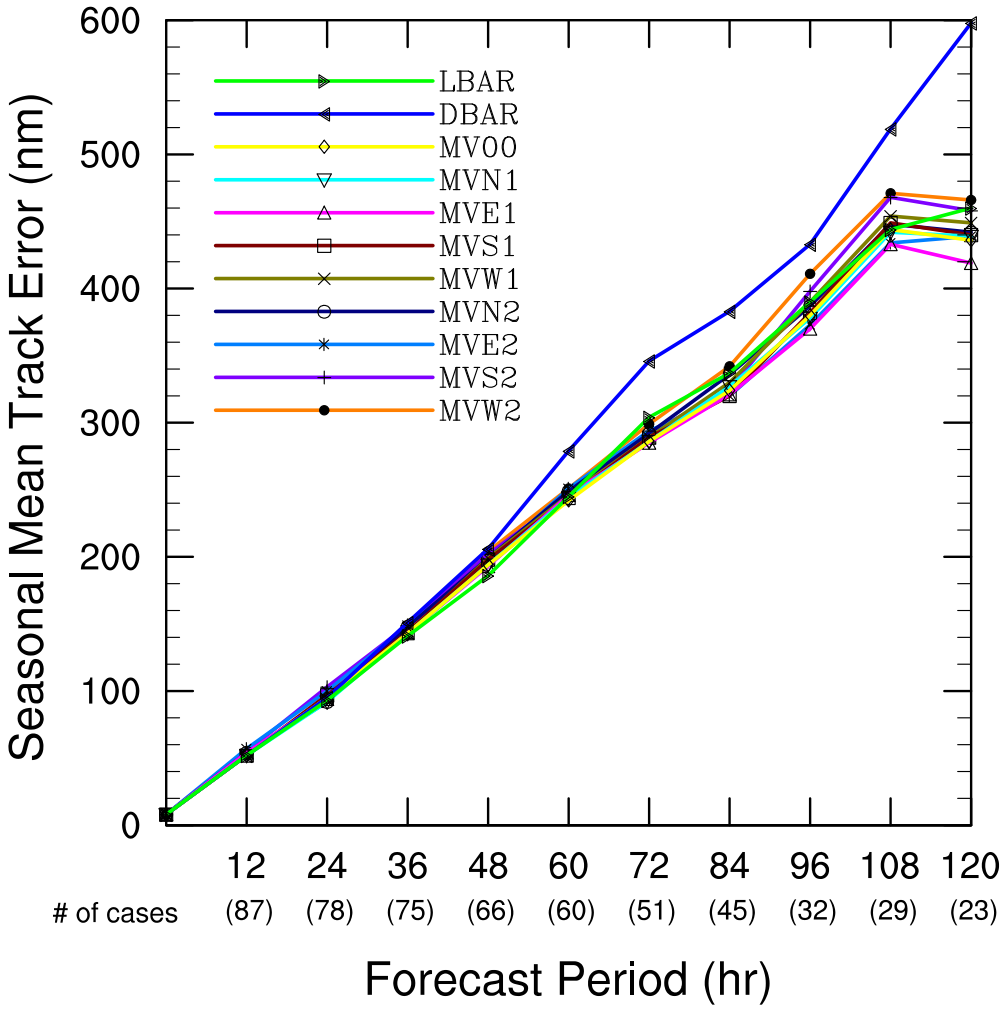


Figure 4.17: Mean tracks errors for the 2001 Atlantic hurricane season for perturbations to the storm motion vector in a MBAR-type configuration of MUDBAR. LBAR indicates the operational shallow-water equation model, while MVE1, MVS1, MVW1, and MVN1 indicate the MBAR-type configurations with a motion vector perturbation of $1\,\mathrm{m\,s^{-1}}$ to the east, south, west, and north of the operationally-estimated motion vector, respectively. MVE2, MVS2, MVW2, and MVN2 indicate similar directional motion vector perturbations, but with a magnitude of $2\,\mathrm{m\,s^{-1}}$.

track errors. Nudging a storm too far to the west or south leads to a failure to recurve in some cases, which might explain this difference.

Figure 4.18 is as Fig. 4.17, but for the seasonal mean track x-bias. As one might expect, the $2\,\mathrm{m\,s^{-1}}$ westward motion perturbation show the largest westward bias at all time periods. The $2\,\mathrm{m\,s^{-1}}$ southward motion perturbation also show a similarly large westward bias of up to 200 n mi at later forecast times. The $1\,\mathrm{m\,s^{-1}}$ southward and westward motion perturbations follow a similar development, but with somewhat smaller westward biases. Both the 1 and $2\,\mathrm{m\,s^{-1}}$ eastward motion perturbations produce forecasts with an eastward bias through intermediate time periods (again, this is no great surprise, since storms that start off moving more slowly westward will probably not travel as far westward), but a westward bias develops after 84 h for all perturbations, in keeping with the observed overall westward bias that MUDBAR seems to have at longer time periods. The northerly motion perturbations produce forecasts with similar x-bias characteristics to the eastward motion perturbations. The motion vector perturbation with the smallest overall x-bias appears to be the northward $2\,\mathrm{m\,s^{-1}}$ motion perturbation. The control forecast (no motion vector perturbation) MV00 (which is the same as MBAR in the other plots in this chapter) has an x-bias right down the middle between the northward and eastward perturbations and the southward and westward perturbations. Again, we would hope this would be the case.

Figure 4.19 is as Fig. 4.17, but for the seasonal mean track y-bias. The interpretation is also similar: the northward perturbations have northward biases at intermediate time periods, then southerly biases at the longer time periods (again, in keeping with the overall trend of MUDBAR to move storms too far to the south and west after 84 h). Eastward perturbations follow a similar trend with somewhat smaller northward biases at intermediate times, then southerly biases towards the end of the forecast period. The no perturbation run (MV00) is in the middle of the pack, with the westward and southward perturbations having the greatest southward biases. The motion vector with the smallest overall y-bias at all time periods appears to be the eastward $1\,\mathrm{m\,s^{-1}}$ perturbation.

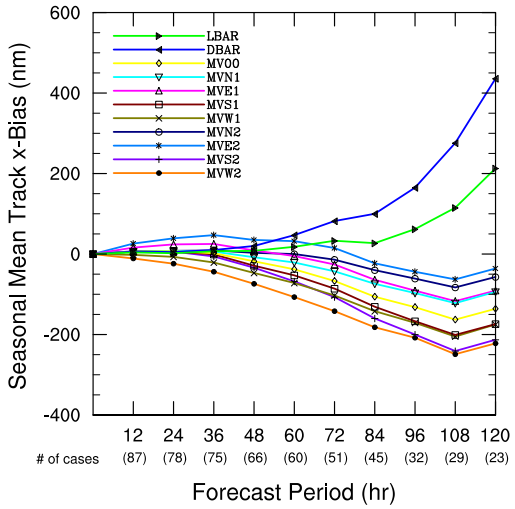


Figure 4.18: As in Fig. 4.17 but for the mean track x-bias.

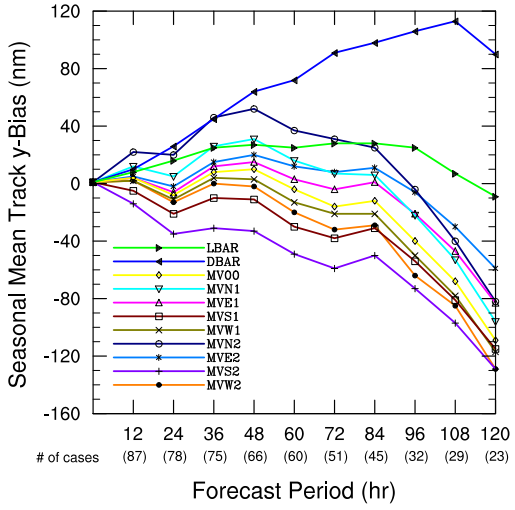


Figure 4.19: As in Fig. 4.17 but for the mean track y-bias.

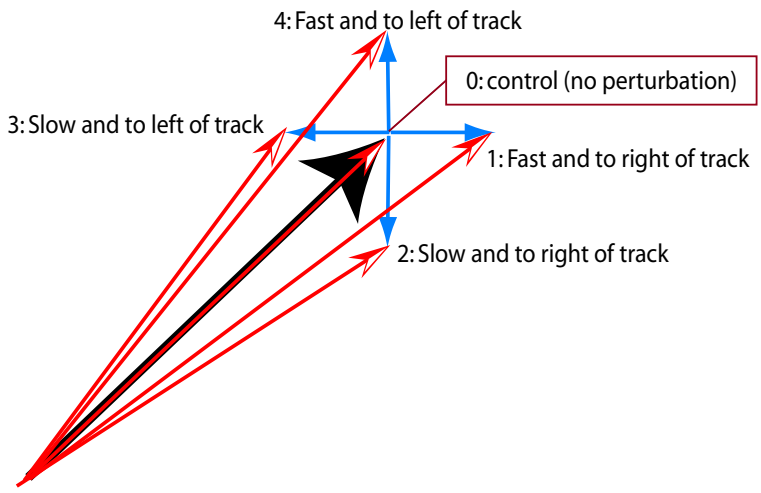


Figure 4.20: Schematic of the perturbations to the storm motion vector used by the kilo-ensemble. The specified (operationally-estimated) storm motion vector is represented by the large black arrow. The perturbation vectors (blue arrows) have a magnitude of $1\,\mathrm{m\,s^{-1}}$. The resulting perturbed storm motion vectors used in the kilo-ensemble are represented by the red arrows.

Overall, the results of this section suggest that if the motion vector was not perturbed in an ensemble framework, then the best reduction in bias might be found by adding an eastward and northward vector to the specified storm motion vector. This would help counteract the excessive southward and westward biases that seem to be inherent to MUDBAR.

Although the sensitivity experiments used motion vector perturbations fixed in the four cardinal directions (N, E, S, and W), which made for an easy interpretation of their effect on the mean track x- and y-biases, the kilo-ensemble uses a somewhat different method. Since storms often move to the northwest early in life and then later move to the northeast after recurvature, it was decided that the motion vector should be perturbed relative to the along- and cross-track directions rather than fixed cardinal directions. The drawback to

this relative approach is that interpretation of this perturbation becomes difficult since the storm could be moving in any direction. The perturbations are implemented[6] by adding the following four vectors to the storm motion vector (as well as a zero vector for no perturbation): fast and to the right, slow and to the right, slow and to the left, and fast and to the left, as shown in Fig. 4.20. The entire kilo-ensemble was run twice for the 2001 season: once with a magnitude of $1\,\mathrm{m\,s^{-1}}$, and again with a magnitude of $2\,\mathrm{m\,s^{-1}}$. The error and spread of various subensembles were calculated as described in Chapter 5, and spatial probability plots were constructed and evaluated qualitatively. Generally, the $2\,\mathrm{m\,s^{-1}}$ motion vector perturbations did not substantially increase the skill of the ensemble, although the spread increased somewhat. Since the ensemble seemed to possess adequate spread (at least in a qualitative sense) using the $1\,\mathrm{m\,s^{-1}}$ motion vector perturbations, this magnitude was chosen for use in the operational ensemble.

4.6 Implementation of ensemble forecasting system

A comprehensive multi-parameter kilo-ensemble forecasting system is constructed from the above considerations by cross multiplying the various perturbations by class such that MUDBAR is run for every possible combination between a given perturbation and all perturbations within other classes:

---

[6] To actually arrive at these perturbed motion vectors, the components of each perturbation vector are defined relative to the storm persistence vector, in the along-track and cross-track directions. Then these vectors are transformed into x-y space and their respective x- and y-components are added to the components of the specified storm persistence vector. For example, the first motion vector perturbation, which is perturbed to be fast and to the right of the track, consists of positive along- and cross-track components with magnitude $\frac{\sqrt{2}}{2}$. The resulting perturbation vector has length of $1\,\mathrm{m\,s^{-1}}$ and makes a 45° angle with the specified storm motion vector. The perturbed storm motion vector is the sum of the storm motion vector and the perturbation vector.

11 environmental evolutions

4 deep layer-mean winds

3 equivalent phase speeds

3 vortex sizes (strengths)

$\times$ 5 motion vectors

1980 members

To implement such a system in an operational framework, it was necessary for MUDBAR to be able to ingest both the initial conditions and boundary conditions as well as the operationally-estimated storm information produced by NHC/TPC (or other operational prediction centers). A script program system was designed to accomplish these front-end tasks for the optimal MBAR configuration described in Chapter 3. To allow the running of ensembles, this front-end structure was modified to allow variations of the external perturbation classes, such as the environmental evolutions provided from the GFS global ensemble, and the various deep layer-mean winds. Then the MUDBAR code was modified somewhat to conduct the internal perturbations over c_{eqv}, vortex size, and motion vector. The resulting track output is then collected and processed into a format compatible with the ATCF format. Postprocessing and graphical product generation is accomplished through an additional layer of scripts. Figure 4.21 summarizes these procedures.

4.7 Summary

This chapter discussed the ensemble design philosophy, choice of perturbation classes, and the choice of the magnitudes used in those perturbation classes. The ensemble design was guided by a desire to create an ensemble which would be simple, efficient, and yet provide enough members to allow explicit strike probabilities to be made. To this end, a parameter-based perturbation methodology was chosen. Five perturbation classes have been used to simulate the inherent uncertainties of the evolution of the environment and steering layer depth, the vertical decomposition of the tropical atmosphere, the size of the

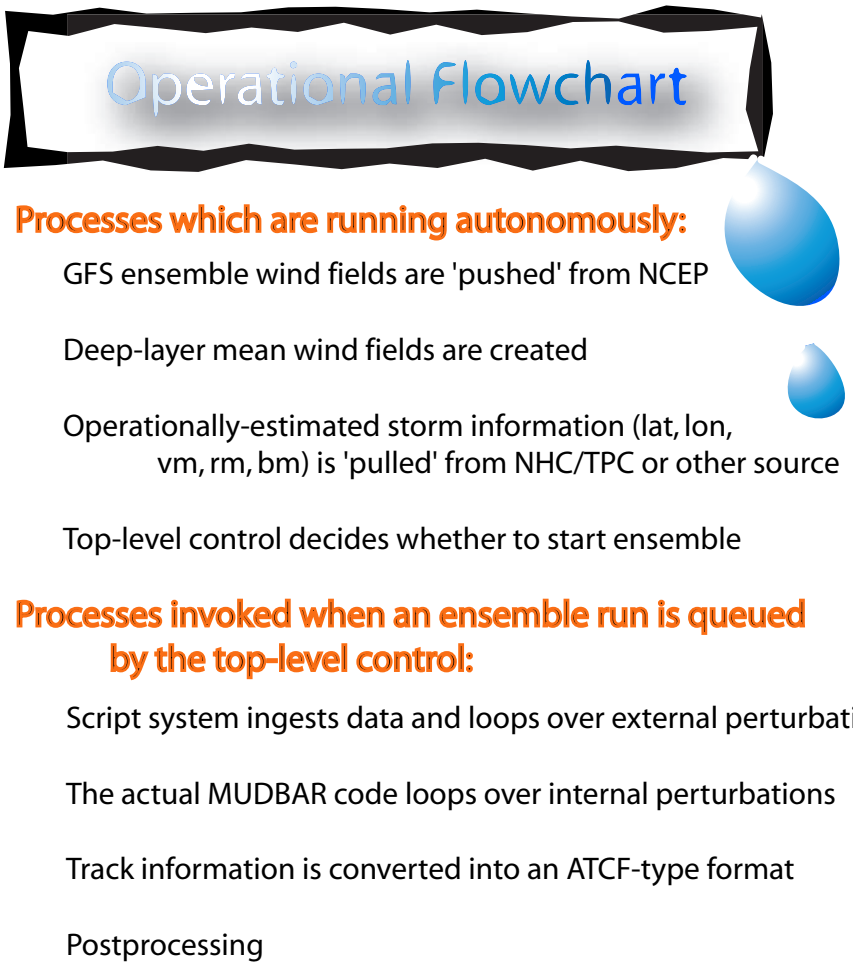


Figure 4.21: Operational flowchart of the automated kilo-ensemble forecast system.

vortex, and the storm persistence vector. Sensitivity experiments were conducted for the 2001 Atlantic hurricane season to examine the characteristics of the forecast track error, x-bias, and y-bias. From the results of these experiments, a 1980-member ensemble was built and implemented in an operational framework: the kilo-ensemble.

Chapter 5

POSTPROCESSING AND VERIFICATION

5.1 Introduction

It is one thing to design, build, and run a kilo-ensemble, but it is quite another thing to verify the ensemble forecasts and derive overall measures of the ensemble performance. This chapter seeks to accomplish this task. The organization of the kilo-ensemble track output is discussed first, followed by a description of the postprocessing procedures used to obtain the ensemble mean forecasts and spatial strike probabilities. The second half of the chapter deals with the verification of these derived products. The relative merits of various verification methods are discussed. Results follow.

5.2 Organization of the kilo-ensemble track output

The kilo-ensemble was run for the 0000 UTC cases of storms that occurred during the Atlantic hurricane seasons of 2001-2003.[1] Out of the 332 possible cases (104 cases in 2001, 92 in 2002, and 136 in 2003), the kilo-ensemble successfully ran a total of 294 times. Thus, the complete kilo-ensemble was available for 88.5% of all possible cases. Although a thorough investigation of the unavailable cases was not undertaken, it is believed that most of these failed cases resulted from missing or corrupted background wind fields from the GFS ensembles. Each ensemble run for each case has a total of 1980 members, so this

[1] For the 2001 season, the real-time collection of the GFS background wind fields did not commence until after Tropical Storm Allison and Tropical Depression Two, so these storms are not included in the total number of possible cases.

represents a total of 582,120 individual MBAR model runs. The individual track forecast positions were recorded at 12-h intervals for each ensemble member through 120 h, for a total of 10 forecast positions. All in all, there are about 5.8 million forecast positions to organize and analyze.

The ensemble output for each case is stored in an ASCII text file using a data format similar to the ATCF format used by all U. S. operational TC warning centers. The file contains CARQ lines that give the initial position of the storm at the time for which the forecast is valid (forecast hour 0, or more simply, $\tau = 0$). Each of the remaining lines corresponds to an individual ensemble member forecast. Each line consists of a tech number,[2] the 4-letter tech identifier (described below), the date/time at $\tau = 0$, followed by the forecast positions (lat/lon pairs) for $\tau = 12, 24, 36, \ldots, 120$ h.

Each ensemble member is given a unique 4-character 'tech' identifier which is used to designate the run-time characteristics of that member.[3] The first character is a letter ('A' through 'K') which designates the GFS ensemble member used for the initial and background wind fields: 'A' designates the kilo-ensemble runs that used the control member of the GFS ensemble, letters 'B' through 'F' designate runs which use the positively-perturbed GFS ensemble forecasts, and letters 'G' through 'K' designate runs which use the negatively-perturbed ensemble forecasts. The second character is a number that designates which deep-layer mean is used for the vertical averaging of the wind fields: '4' represents the very deep-layer mean (1000-100 hPa), '5' represents the standard deep layer mean (850-200 hPa), '6' represents a medium-depth layer mean (850-350 hPa), and '7' represents a shallow-depth layer mean (850-500 hPa). The third character is a letter which designates both the size of the vortex and the value of c_{eqv} that is used in the run. Small storms ($v_m = 15\,\mathrm{m\,s^{-1}}$)

[2] The tech number is a unique identifying number given to operational track guidance products. It seems that older verification programs used this number, but now it is apparently meaningless, at least in this work. All the individual kilo-ensemble member forecasts are given a tech number of 77, whereas experimental forecasts aids are usually designated by 99.

[3] This usage is a bit different than in the traditional ATCF, which uses tech identifiers simply to designate the various operational forecast aids (e.g., AVNO, LBAR, etc.)

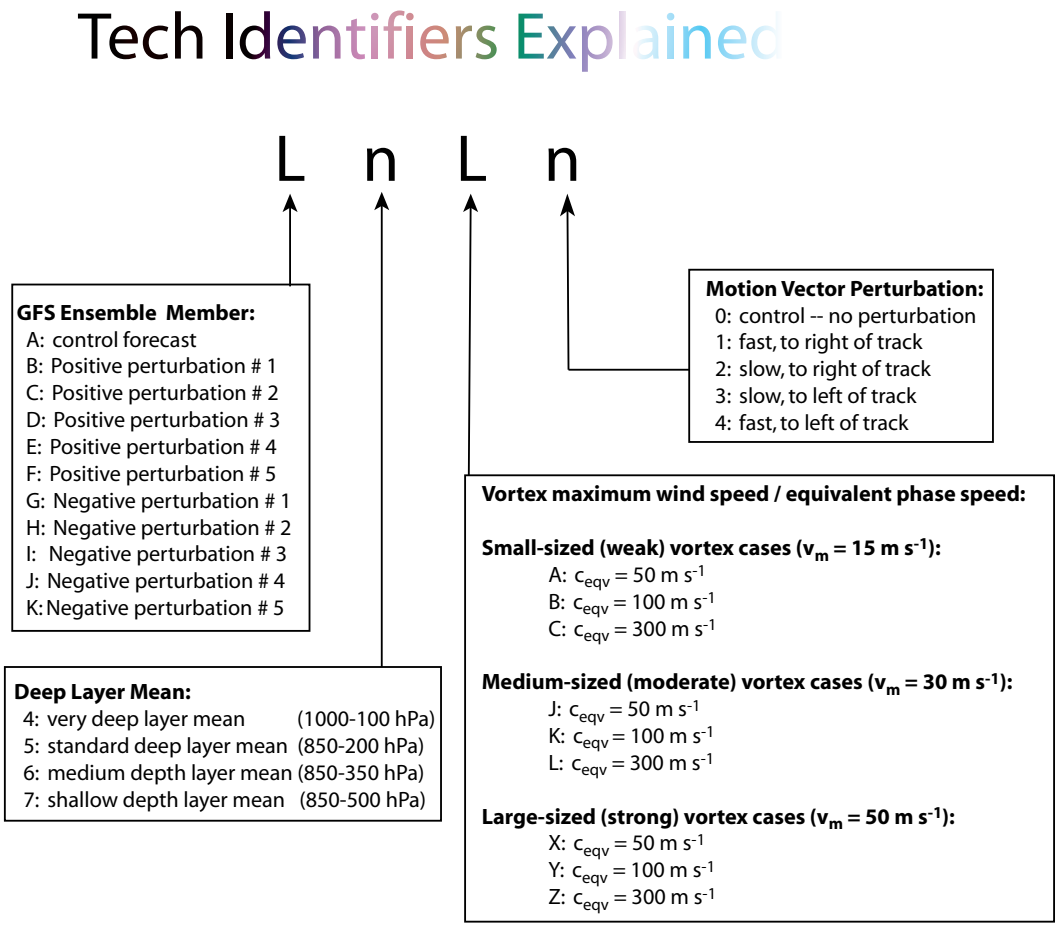
are designated by 'A', 'B', and 'C' with c_{eqv} values of 50, 100, and 300 m s^{-1}, respectively. In a similar fashion, medium-sized storms ($v_m = 30\,\mathrm{m\,s^{-1}}$) are designated by 'J', 'K', and 'L', while large storms ($v_m = 50\,\mathrm{m\,s^{-1}}$) are designated by 'X', 'Y', and 'Z'. The fourth character is a number which designates the motion vector perturbation: '0' is the control (no perturbation), '1' is the perturbation that is fast and to the right of the storm motion vector, '2' is slow and to the right, '3' is slow and to the left, and '4' is fast and to the left. This tech naming convention is summarized in Fig. 5.1.

5.3 Postprocessing

During the postprocessing, the ensemble output files are processed by a program that computes the forecast tracks of the total ensemble mean and various subensemble means. The ensemble spread is also calculated. In addition, the program searches through the ATCF 'best track' files (known as the 'B-decks') and inserts a line (with tech identifier 'STRK') that contains the official best track positions that correspond to the forecast period (starting at $\tau = 0\,\mathrm{h}$). This 'best track' is a best estimate of the actual storm track history, as determined by experts at TPC/NHC using all available information sources, both at forecast time and afterwards. For verification purposes in this study, the best track is considered to be the actual storm track. For further description of the best track history files, see Jarvinen et al. (1984).

5.3.1 Total ensemble mean track forecast

There are many ways to characterize the information contained in an ensemble forecast. The simplest of these involves calculating the total ensemble mean, which is the expected value obtained by averaging the forecasts of all ensemble members. The total ensemble mean provides a zeroth-order measure of the information in the ensemble, giving just one forecast track. In the process of making such a drastic summarization, all information about the ensemble spread, skewness, and other higher moments is lost. Thus, the ensem-



As an example, the tech identifier of the kilo-ensemble control member is A5K0. This designates that this member uses the control forecast of the GFS ensemble, uses the standard deep layer mean (850-200 hPa), has a medium-sized vortex with v_m = 30 m s^{-1}, uses an equivalent phase speed of c_{eqv} =100 m s^{-1}, and has no motion vector perturbation.

Figure 5.1: The systematic naming system for the kilo-ensemble tech identifiers.

ble mean forecast can tell nothing about the potential forecast uncertainty or probability distribution.

For ensemble member i, the forecast position $f_i(\tau)$ is given by a latitude/longitude pair (ϕ_i, λ_i). The total ensemble mean forecast position $\bar{f}(\tau)$ is given by the mean latitude $\bar{\phi}$ and the mean longitude $\bar{\lambda}$ of all the individual ensemble members at a given forecast time τ. The bar notation ($\bar{}$) is used to designate the mean over all ensemble members:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{5.1}$$

where n is the total number of available member forecasts. For later forecast periods (e.g., $\tau = 72\,\mathrm{h}$ or $\tau = 120\,\mathrm{h}$), n is often less than 1980 because the model representation of the storm can be lost. This occurs whenever a member forecasts the storm to travel into the edge of the model domain, or when the model representation of the storm is too weak to be tracked by the vortex tracking algorithm. In either case, the remaining periods of that member's forecast are assigned a value of 'missing' so that those forecasts can be excluded from the calculation of the ensemble mean. For fast-moving storms, it is not uncommon for most or all of the members to hit the edge of the domain before the end of the forecast period. If a substantial number of ensemble members are missing, the total ensemble mean ceases to be representative of the ensemble, since it is only based on the members which remain in the domain. As the influence of faster-moving member forecasts is lost, the ensemble mean forecast track may experience unphysical changes in direction and speed. The effect of these 'edge issues' can be minimized somewhat by requiring a minimum number of nonmissing forecasts n in order to calculate a valid ensemble mean forecast.

Two cutoff values were tried initially: a cutoff value equal to 90% of all ensemble members and a value equal to 20% of all members. The 90% cutoff reduces the edge bias substantially, but also reduces the number of cases available for verification at the longer forecast periods. Using too high of a cutoff removes too many cases with rapidly moving storms. Since these cases are more likely to possess a higher potential for large forecast errors, removing them from the verification sample could artificially inflate the performance

of the total ensemble mean, especially at later time periods. On the other hand, the 20% cutoff only eliminates the most egregious edge biases, but gives a forecast for many more cases at the longer time periods. As a compromise, a 50% cutoff is also used. Verification is conducted on these means (calculated with 20%, 50%, and 90% cutoffs) and is discussed later in this chapter.

The total ensemble mean forecast is calculated by the postprocessing program at each τ. The resulting forecast is appended to the ensemble output file as a line with tech identifier 'ZTOT' (for the 20% cutoff) or 'ZT50' (for the 50% cutoff).

5.3.2 Subensemble mean track forecasts

In addition to the total ensemble mean, various subensemble mean forecasts can be generated. A subensemble is comprised of a subset of the members in the total ensembles. These members may be selected either randomly, or by certain selection criterion. The subensemble mean forecast is simply the mean of the forecasts of the members that make up that particular subensemble. When the inclusion criterion is based on a certain selection criterion, a subensemble mean forecast represents a collective measure of the behavior of the all the members in that subensemble set.

In the case of the kilo-ensemble, there are 26 possible perturbations across the five perturbation classes. Thus, it is logical to calculate 26 corresponding subensemble means such that all individual members in each subensemble have a common perturbation. Here is another way to say this: one parameter in a perturbation class is held constant while all other perturbations in the other classes are allowed to vary through all possible parameter values. The postprocessing program calculates each subensemble mean and appends the resulting forecast line onto the ensemble output file with a subensemble tech identifier (described in the next paragraph). As an example, one subensemble is comprised of all kilo-ensemble members which use the background wind fields of the GFS control run. Everything in all other perturbations classes (deep-layer averaging, storm size, c_{eqv}, and motion vector) is

allowed to vary through the range of possible perturbations. Since there are 11 different GFS background wind fields, this subensemble has 1980/11 = 180 members. As with the total ensemble mean, each subensemble also requires a minimum percentage of nonmissing forecasts to have a valid subensemble mean forecast. The cutoff value is set to 20% for all subensembles. Under this selection system, the collection of all individual members that comprise the subensembles for all the perturbations of a particular perturbation class is simply the totality of the kilo-ensemble. Since the mean is distributive, it follows that the mean of the subensemble means for all the perturbations in a particular perturbation class is simply the total ensemble mean.

Like the individual members of the kilo-ensemble, each subensemble is given a tech identifier according to a methodical naming convention. All subensemble techs begin with the letter 'S' which stands for "subensemble mean". The following two letters identify the perturbation class that is used for the selection criterion. The five perturbation classes and associated perturbations in each class are: the 11 perturbations to the environment given by the various GFS ensemble runs (this class is designated by 'BF' for 'background fields'), the 4 deep-layer averaging depth perturbations (designated by 'LY' for 'layer'), the 3 perturbations to vortex size ('VM' for 'vmax'), the 3 equivalent phase speeds ('GM' for 'gamma'), and the 5 perturbations to storm motion vector ('MV' for 'motion vector'). The fourth letter designates which of the perturbations in that particular class is held constant. Thus, the subensemble mean example given in the previous paragraph would have a tech identifier of 'SBFA'). The subensemble tech naming convention is summarized in Fig. 5.2.

5.3.3 Forecast track error

The error of any forecast is simply the difference between the forecast and the corresponding observation. In the context of ensemble track forecasting, the error (or direct positional error) is simply the distance between each forecast position and the verifying best track position, for each forecast time period, τ. These calculations are only possible when

Subensemble Techs Explained

Sub-ensembles based on background wind field evolutions

SBFA: All members that use GFS control background wind fields
SBFB: All members that use wind fields of GFS positively-perturbed run #1
SBFC: All members that use wind fields of GFS positively-perturbed run #2
SBFD: All members that use wind fields of GFS positively-perturbed run #3
SBFE: All members that use wind fields of GFS positvely-perturbed run #4
SBFF: All members that use wind fields of GFS positively-perturbed run #5
SBFG: All members that use wind fields of GFS negatively-perturbed run #1
SBFH: All members that use wind fields of GFS negatively-perturbed run #2
SBFI: All members that use wind fields of GFS negatively-perturbed run #3
SBFJ: All members that use wind fields of GFS negatively-perturbed run #4
SBFK: All members that use wind fields of GFS negatively-perturbed run #5

Sub-ensembles based on depth of the layer-mean wind average

SLY4: All members that use the 1000-100 hPa deep layer-mean wind fields
SLY5: All members that use the 850-200 hPa deep layer-mean wind fields
SLY6: All members that use the 850-350 hPa deep layer-mean wind fields
SLY7: All members that use the 850-500 hPa deep layer-mean wind fields

Sub-ensembles based on the storm size/strength (v_m)

SVM1: All members that use v_m = 15 m s^{-1}
SVM2: All members that use v_m = 30 m s^{-1}
SVM3: All members that use v_m = 50 m s^{-1}

Sub-ensembles based on the equivalent phase speed (c_{eqv})

SGM1: All members that use c_{eqv} = 50 m s^{-1}
SGM2: All members that use c_{eqv} = 150 m s^{-1}
SGM3: All members that use c_{eqv} = 300 m s^{-1}

Sub-ensembles based on the motion vector perturbation

SMV0: All members that use do not have a motion vector perturbation
SMV1: All members that are perturbed fast and to the right
SMV2: All members that are perturbed slow and to the right
SMV3: All members that are perturbed slow and to the left
SMV4: All members that are perturbed fast and to the left

Figure 5.2: The systematic naming system for the subensemble mean tech identifiers.

the best track verification becomes available at the end of the season.[4]

The track error E is defined as the magnitude of the distance vector $\vec{d}$ between the forecast position f and the observed position o at each τ. In practice, these distances are calculated using an approximation to the Great-Circle distance formula:[5]

$$E = \|\vec{d}\| = \sqrt{d_x^2 + d_y^2} = \sqrt{\left(a\,\cos\tfrac{\phi_f+\phi_o}{2}\,(\lambda_f - \lambda_o)\right)^2 + \left(a\,(\phi_f - \phi_o)\right)^2}\,, \tag{5.2}$$

, where d_x and d_y are the x- and y-components of $\vec{d}$, a is the radius of the earth in nautical miles (n mi, and ϕ and λ are in radians. In keeping with the practice of NHC, distance is expressed in nautical miles, rather than kilometers. This approximate distance formula gives very small errors for most distances under 1000 n mi. Errors of up to 1-2% can occur when calculating distances on the order of 2000 n mi and both the latitude and longitude of the two points are very different. For more on verification procedures, see Neumann and Pelissier (1981).

5.3.4 Ensemble spread

The spread of an ensemble gives a measure of the variability of the individual ensemble member forecasts. The ensemble spread (or ensemble standard deviation) S is defined as the mean distance d_i between the individual member forecasts f_i and the ensemble mean forecast $\bar{f}$:

$$S = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\left(a\,\cos\tfrac{\phi_i+\bar{\phi}}{2}\,(\lambda_i - \bar{\lambda})\right)^2 + \left(a\,(\phi_i - \bar{\phi})\right)^2}\right). \tag{5.3}$$

Ensemble spread provides a first-order estimate of the uncertainty of the forecast situation. Larger spreads imply greater forecast uncertainty (or decreased predictability), and smaller spreads imply smaller forecast uncertainty (increased predictability). Ensemble

[4] It is possible to calculate provisional track errors during the season using the subsequent operational positional estimates, but these track errors are not considered to be official.

[5] The actual Great-Circle distance formula between two points (λ_1, ϕ_1) and (λ_2, ϕ_2) is given by:

$$d = a\ \arccos\left[\cos\phi_1\cos\phi_2\cos\left(\lambda_1 - \lambda_2\right) + \sin\phi_1\sin\phi_2\right]$$

.

spread is almost always greater at later forecast periods than at early times periods, in line with the expectation that predictability decreases at longer lead times. This very basic interpretation of spread is generally true as long as the ensemble can simulate at least some of the sources of uncertainty. If the ensemble simulates the uncertainty incorrectly (due to structural deficiencies in the model dynamics and/or parameterizations) or adds false sources of uncertainty (due to improper ensemble design), then this spread-uncertainty interpretation is not as valid. More discussion on the interpretation of spread and forecast uncertainty is given later in the chapter.

5.3.5 Spatial strike probability forecasts

A great advantage of large ensembles is that they allow the estimation of the probabilistic density functions (PDF) for the predicted variables. In this study, the storm track is the forecast variable of interest, so it is reasonable to calculate a map of storm strike probabilities. This spatial map is simply a two-dimensional PDF of future storm location: each region of the map is assigned a probability associated with the predicted likelihood of a storm strike. If the ensemble can accurately assess the forecast uncertainty, then the probability for a given point in the domain should represent the actual likelihood of the storm striking that point.

The kilo-ensemble spatial strike probabilities are calculated on a $1^\circ \times 1^\circ$ grid which corresponds to a domain slightly larger than the model domain. Each grid point corresponds to the intersection of an integral parallel of latitude with an integral meridian of longitude. For the purposes of this calculation, a storm strike occurs when the center of the storm passes within 75 statute miles of the grid point. Thus, the strike probability for a given point is equal to the number of individual member forecasts that pass within 75 statute miles of the grid point divided by the total number of ensemble members (which is always taken to be 1980) multiplied by 100%.

The passage criterion of 75 statute miles is used in this study because it is the

threshold value used by the NHC's experimental strike probability product. Even though the NHC probabilities are not based on the same type of explicit calculation, these probabilities should be directly comparable to those of the kilo-ensemble strike probabilities. The 75 mile distance roughly corresponds to the radius of the core of damaging winds in a hurricane.

5.3.6 Ensemble Clustering

In some forecast situations, such as recurvature, it is possible for the forecasts of ensemble members to fall into two or more distinct camps – a bifurcation. As an example, some members may forecast a TC to recurve, while others might maintain a westward track. A clustering analysis seeks to find groups of individual ensemble members that have similar solutions. This work does not attempt a rigid cluster analysis, but the generation of spatial strike probabilities does lend itself to a simple cluster analysis. By tracking distinct regions of the highest local instantaneous strike probability, several synthetic cluster forecasts could be generated. These track forecasts would be verified in the same manner as the ensemble mean track forecasts. Due to lack of time, this type of cluster analysis is left for future work.

5.3.7 Graphical products

Several graphical products are created from the kilo-ensemble forecasts. One product is a 'swarm' animation which displays color contours of the instantaneous ensemble strike probability at each τ in the forecast period (out to 120 h, at 12-h intervals). These swarm animations show the forecast positions of the total ensemble mean and 26 subensemble means, and their tracks up through time τ. The swarm plots also include the instantaneous forecast positions of all the individual ensemble members. This allows the user to see the entire raw ensemble forecast output at a glance. Of course, there is a close correspondence between the density of the individual forecasts and the computed strike probabilities. Finally, if the plot is made after the best tracks are available (generally a month or two after

the end of the season), the best track positions and track are also included. Figure 5.3 shows the $\tau = 72\,\mathrm{h}$ frame of a swarm animation for Hurricane Isabel for the forecast starting at 0000 UTC on 17 September 2003.

Cumulative spatial strike probability plots are also created by finding the maximum strike probability at each grid point for all forecast times from $\tau = 0\,\mathrm{h}$ to $\tau = 72\,\mathrm{h}$, and $\tau = 0\,\mathrm{h}$ to $\tau = 120\,\mathrm{h}$.[6] The resulting maps of cumulative strike probability are designed to mimic the NHC's experimental strike probability product. The total ensemble mean forecast (ZTOT) and the best track are plotted as black and red lines, respectively. An example of the cumulative strike probability map for the Hurricane Isabel case is given in Fig. 5.4. For comparison, Fig. 5.5 shows the NHC product.

Several other plots are created to allow easy comparison of the kilo-ensemble track forecasts with the other early- and late-cycle operational forecast aids.[7] Early-cycle forecast aids include statistical-dynamical aids (like A90E), the five-day Climatology and Persistence (CLP5), the extrapolation forecast (XTRP), the Beta and Advection (like BAMD) forecasts, and the barotropic models (like LBAR).[8] To allow the early-cycle products to be produced within a few moments of the forecast verifying time, old model forecast fields are used to initialize and run these products (for example, LBAR uses the 6-h old AVN forecast fields). Late products are generally available 3 or 6 h after the forecast verifying time. Examples include the full physics 3-D models (like GFDL) and the track forecasts obtained from the global models (like the NCEP GFS, the Navy NOGAPS, and the UK Met Office). To allow comparison and display of late-cycle forecasts alongside the early-cycle forecasts, the late-

[6] Before calculating the cumulative spatial strike probabilities by taking the maximum probability over all time periods, the ensemble forecast tracks are first interpolated to a spatial resolution of 1.5 h. This is done to ensure a smooth and continuous spatial field of strike probability, even for fast-moving storms.

[7] Based on their timeliness, the operational forecast aids fall into two camps. If the forecast of a given operational aid is available at the verifying time of the forecast ($\tau = 0\,\mathrm{h}$), it is considered an early aid. If the forecast does not become available until after the forecast verifying time, then it falls into the late-cycle suite of forecast aids.

[8] As a reminder to the reader, DeMaria and Gross 2003 offer a detailed description of all NHC models through 2002.

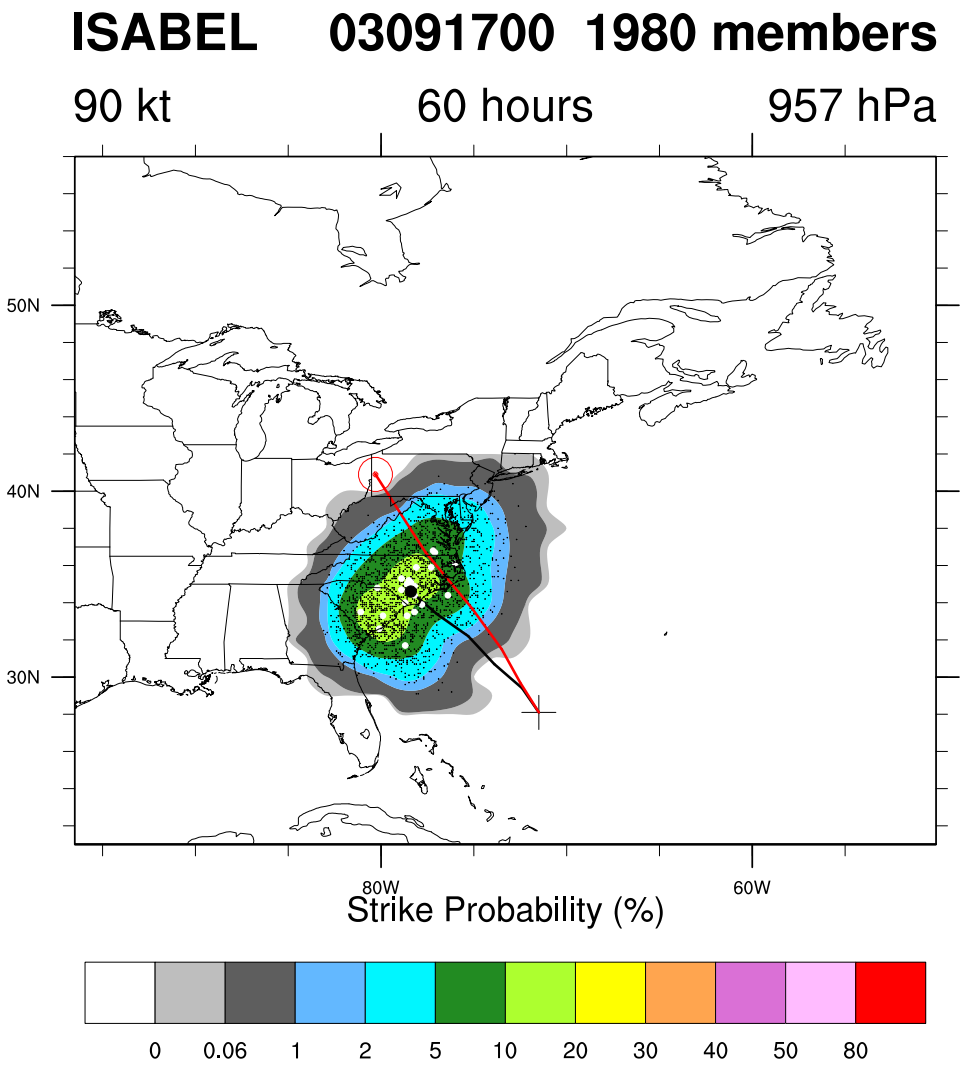


Figure 5.3: The $\tau = 72$ h frame of the swarm plot for Hurricane Isabel for the forecast period starting at 0000 UTC on 17 September 2003. A black crosshatch designates the initial storm position at $\tau = 0$ h. The black line indicates the track of the total ensemble mean forecast, while the red line denotes the verifying best track. White dots show the forecast positions of the 26 subensemble means, while small black dots indicate the forecast positions of the 1980 individual ensemble members. The spatial strike probabilities are indicated by color-filled contours. The first line of the plot header gives information about the storm, including the name, the date (in YYMMDDHH format), and the number of nonmissing individual forecasts at time τ. The second line of the header contains the operationally-estimated storm maximum sustained wind speed (in kt) and minimum central pressure (hPa) at the initial forecast time, as well as the forecast period τ of the frame.

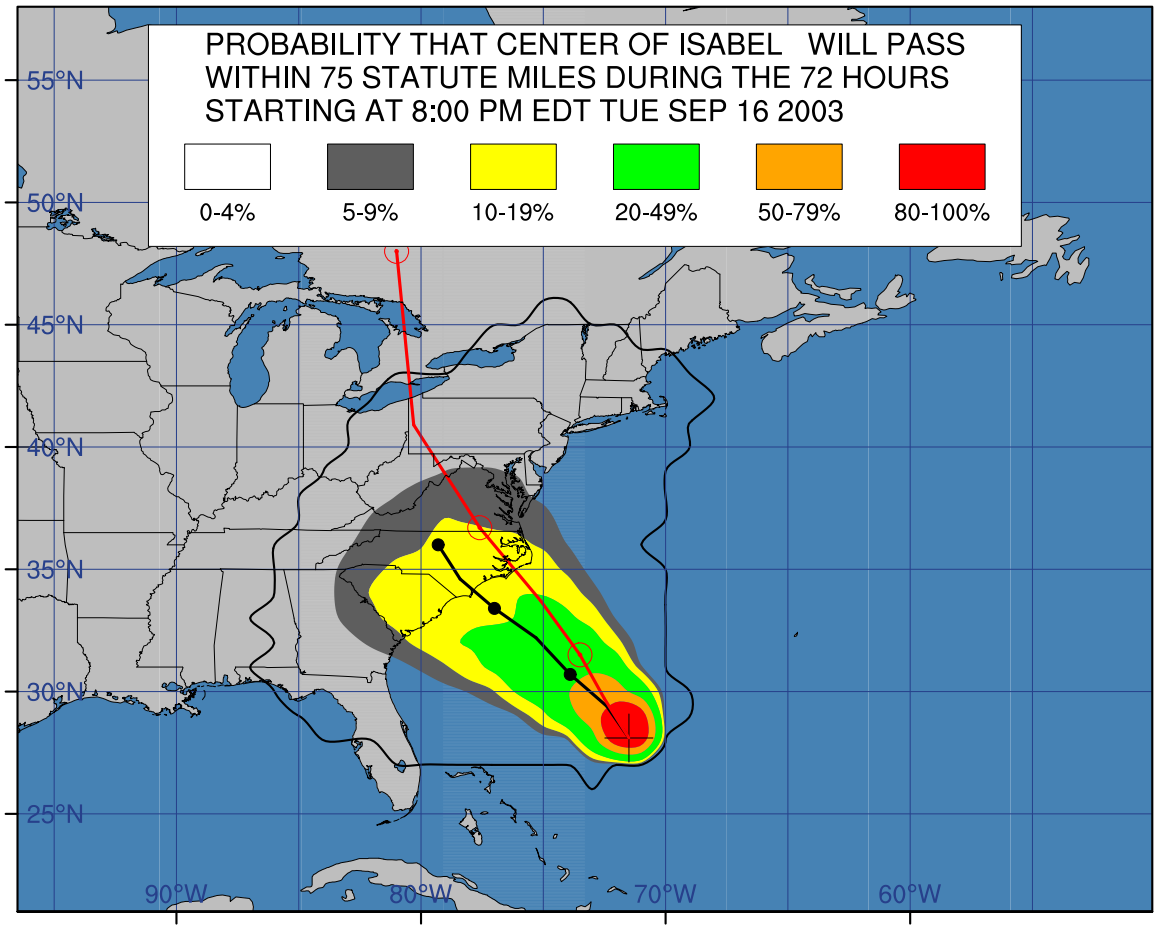


Figure 5.4: The cumulative strike probabilities through $\tau = 72\,\mathrm{h}$ for Hurricane Isabel for the forecast period starting at 0000 UTC on 17 September 2003. The tracks of the total ensemble mean forecast and the verifying best track are indicated by the black and red lines, respectively.

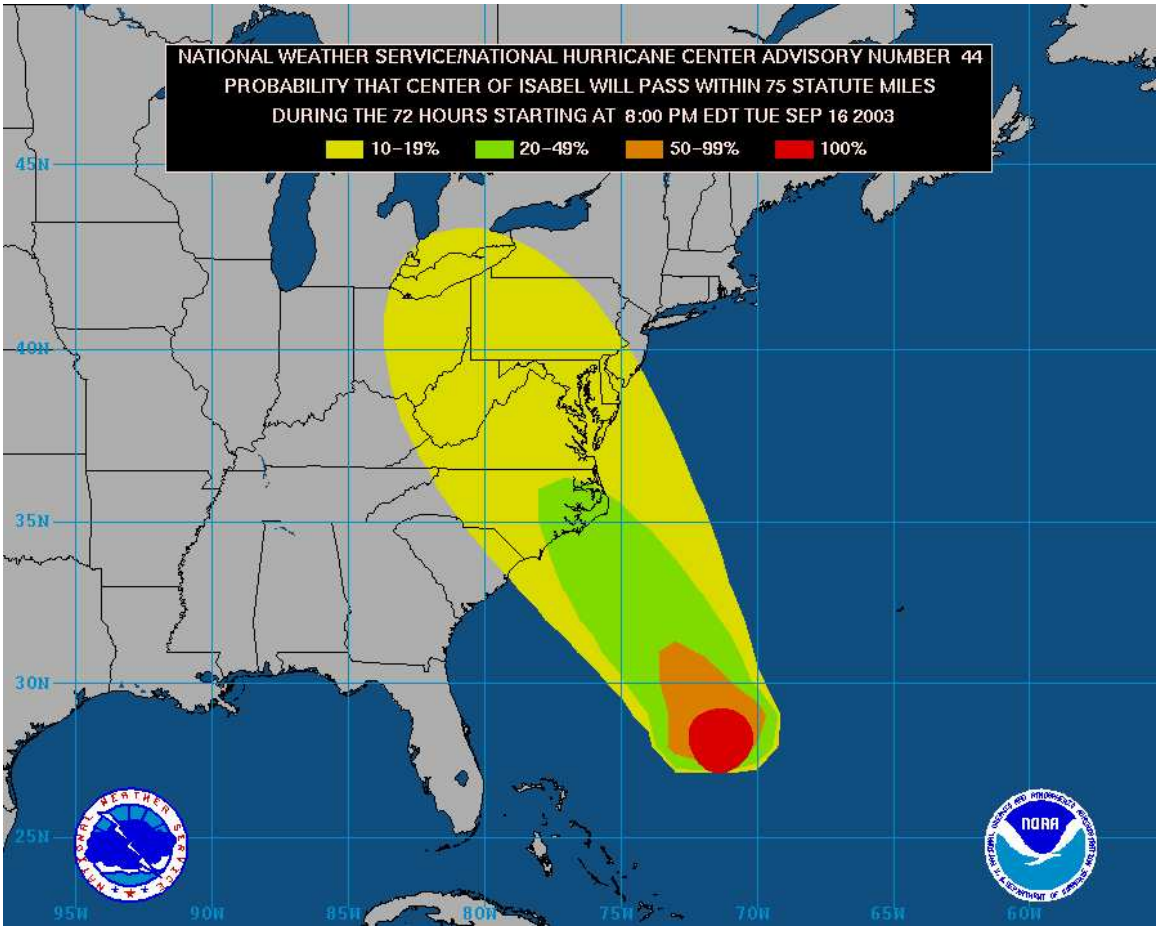


Figure 5.5: The NHC experimental cumulative strike probabilities through $\tau = 72\,\mathrm{h}$ for Hurricane Isabel for the forecast period starting at 0000 UTC on 17 September 2003.

cycle products are also framed as early-cycle products by interpolating and extrapolating the late-cycle forecast from the 6-h old forecast. When the late-cycle forecast has been interpolated, the tech identifier ends in 'I' (e.g., AVNI indicates the interpolated track forecast from the Aviation model). Figure 5.6 shows the early-cycle forecast tracks for Hurricane Isabel for the forecast period starting at 0000 UTC on 17 September 2003; Fig. 5.7 shows the late-cycle forecast tracks for that same case. Finally, since the kilo-ensemble is initialized from the GFS ensemble background fields, it is of interest to see the track forecasts that result from those runs. These are shown in Fig. 5.8.

In keeping with the goal of making the kilo-ensemble system as similar as possible to an automated operational system, a scripting system was created to generate the graphical products when they become available. The resulting plots are posted to the World Wide Web at the following URL: *http://euler.atmos.colostate.edu/∼vigh/ensembles/output_plots.htm*.

5.4 Verification

The job of any verification scheme is to assess the goodness of a given set of forecasts. What constitutes a good forecast? In an essay Murphy (1993) discusses three types of goodness that a forecast may possess: consistency, quality, and value. *Consistency* is the correspondence between the forecasts and the judgments made by the forecaster. *Quality* is the correspondence between the forecasts and the observations. *Value* is the incremental benefit that accrues to users of that forecast. It turns out that each type of 'goodness' is multifaceted – it is impossible to measure the goodness of a forecast through just one or two measures. Verification schemes can be employed for different purposes: scientific, economic, or administrative. Whatever the purpose, some verification measures[9] are better than others, and each scoring scheme possesses different attributes. A verification scheme should use measures that are optimized to the quantities of interest and the purpose of evaluation.

[9] A 'measure' is simply an objective technique for evaluating some aspect of the forecast system's performance.

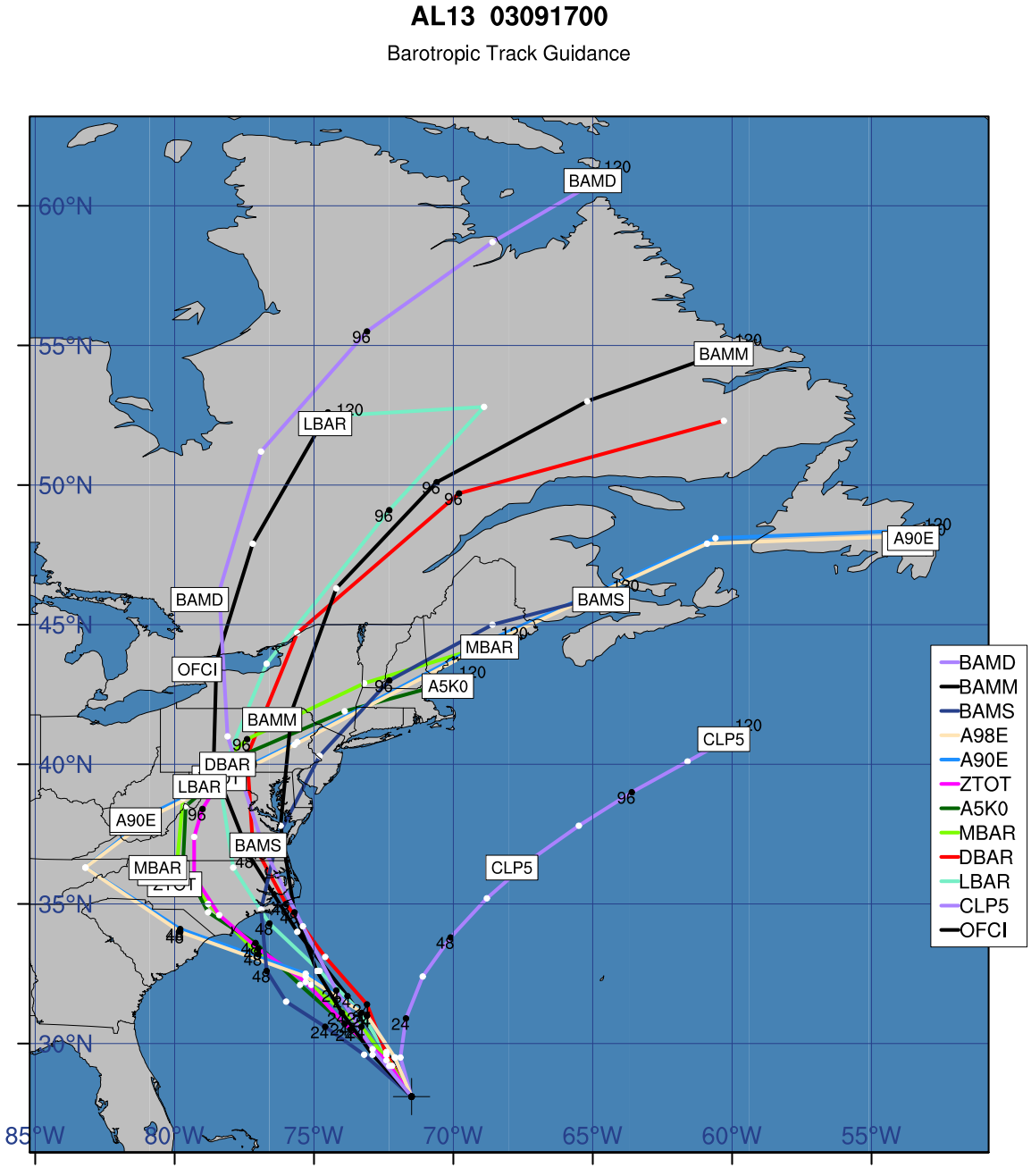


Figure 5.6: The forecasts of the early-cycle aids through $\tau = 120\,\mathrm{h}$ for Hurricane Isabel for the forecast period starting at 0000 UTC on 17 September 2003.

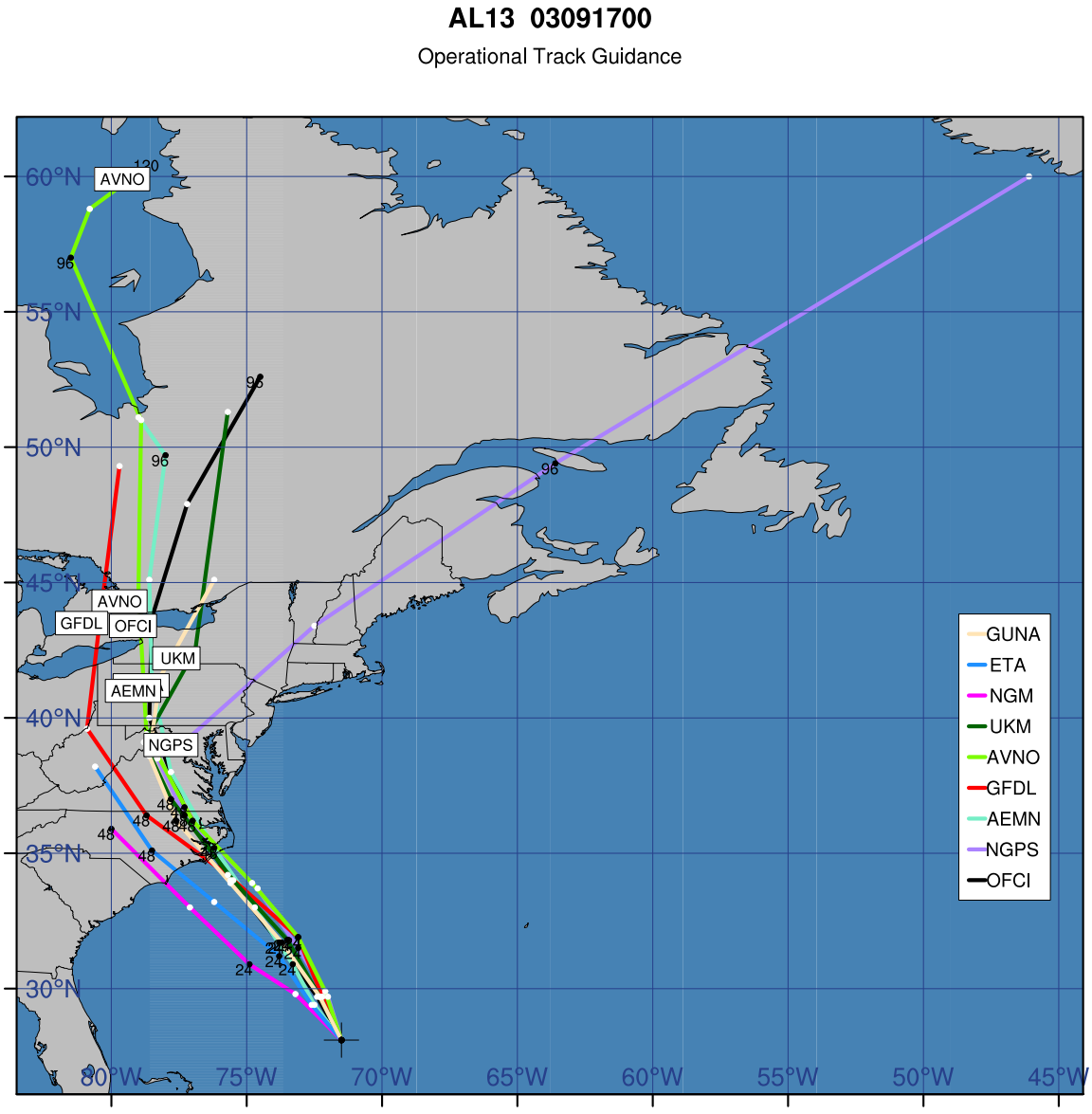


Figure 5.7: As in Fig. 5.6 but for the late-cycle forecast aids.

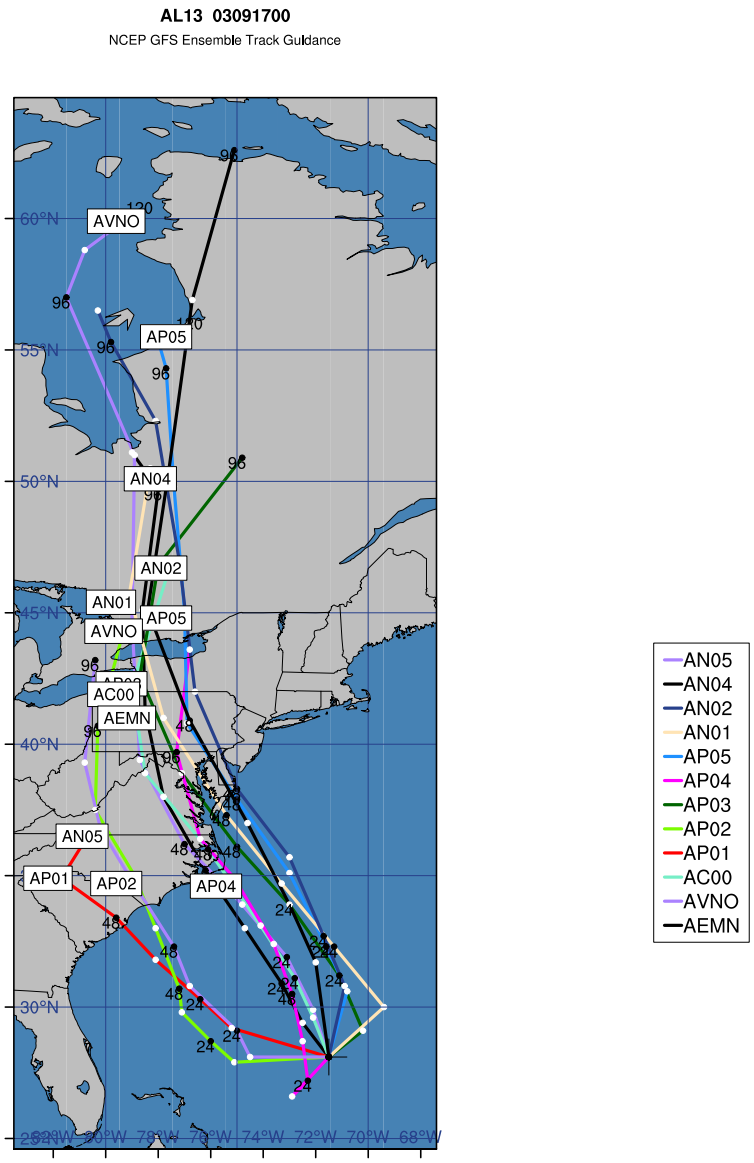


Figure 5.8: As in Fig. 5.6 but for the GFS ensemble forecast aids.

This chapter concerns the performance of one specific forecast aid, the kilo-ensemble, so matters of type-1 goodness (consistency) are not relevant, since that is more a function of the forecaster and the inclusion of his/her judgments in the forecast.[10] Measures of type-3 goodness (value) require information about the forecast users and how their decisions depend on the forecast (i.e., a forecast for hurricane landfall will result in different actions by different users). Measures of value typically involve economic considerations for the cost/loss scenario, where a protect/no protect decision framework is used based on the forecast value. The value is typically presented in terms of economic terms, but could also include the human cost as well. The importance of forecast value is obvious, but this chapter does does not dwell on this subject (for a brief treatment of the economic value of ensemble forecasts, see Zhu et al., 2002), except to note that probabilistic forecasts generally have higher value than deterministic or categorical forecasts.[11] This is one of the motivating factors behind the development of the ensemble strike probabilities. For the purposes of this chapter, it is assumed that more accurate forecasts have higher value than less accurate ones – although this is not necessarily true, it is generally so. Hence, the remaining portion of this chapter is concerned with measuring type-2 goodness: the quality of the kilo-ensemble forecasts.

There are two basic approaches to constructing verification measures: the *measures-oriented approach* and the *distributions-oriented approach*. The measures-oriented approach involves the design of summary measures of the correspondence between the forecasts and the observations, focusing on some overall measure of the quality of the forecast, such as accuracy or skill. The result, obtained after calculation, is a score. Conclusions about

[10] If the ensemble system is considered to be the forecaster, it is obviously completely consistent with its 'judgments', since it is an objective forecast aid.

[11] As an example, consider the hypothetical situation in which a cold front passage may occur the next day, but the strength of the frontal surge is uncertain. The forecaster may feel that it is equally likely that tomorrow's high could be 60°F or 90°F. If the forecaster *hedges* with a forecast high of 75°F, he/she does a disservice to the forecast users, since the forecast does not have any information about the range of equally-likely temperatures. This example of hedging is also a gross violation of type-1 goodness: consistency. For more on the subject of hedging and consistency see Murphy (1978) and Murphy and Epstein 1967.

the forecast performance can then be drawn based on the value of that score. Measures-oriented scores are often easy to calculate and interpret, however they are quite reductionistic, collapsing all the information about the vagaries of the forecast down into just one number. Also, there are numerous scores which can be employed, each with different attributes. The distributions-oriented approach instead seeks to conduct verification based on the joint probability distribution between the forecasts and the observations. Thus, a distributions-oriented verification scheme seeks to assess the statistical characteristics of the joint distribution (summarized from Jolliffe and Stephenson 2003, pp. 30, 58). This framework, outlined by Murphy and Winkler (1987), is extremely useful for the verification of probabilistic forecasts. Measures from both approaches will be presented after a discussion of the cases included in the verification and how forecast availability from domain edge issues impact the verification.

5.4.1 Cases included in the verification

Before discussing the various verification metrics and their merits, it is important to first consider what is being verified. This section provides an overview of the cases included in the verification, and factors that may introduce artificial bias into the verification statistics. Since most tropical cyclone forecast aids are optimized specifically for the forecasting of tropical cyclones, official seasonal verification practices at NHC include only the cases for which a tropical or subtropical cyclone is present (including tropical depressions). These storms are characterized by convective vortex dynamics that operate through a primary energetics pathway involving the extraction of energy from the underlying warm ocean. In general, such systems have a closed circulation in the wind field and a warm-core aloft. The extratropical cyclone, tropical wave, and remnant low stages are excluded from the official seasonal track verification statistics to ensure a fair evaluation of forecast performance.

During the 2001-2003 Atlantic seasons there were a total of 246 valid forecast cases. To be a valid forecast case in the context of this study, the following conditions must be met:

the GFS model fields must be available, a MBAR forecast must exist for at least the $\tau = 12$ h forecast period, and there must be a verification position in the best track through at least 12 h. Out of the 246 total valid cases, there were 230 cases that possessed at least a 12-h total ensemble mean forecast (ZTOT), giving an availability of 93.5%. Because some storms are short-lived and because some forecasts are made near the end of a storm's lifetime, the number of forecast cases with a best-track verifying position decreases at the longer forecast periods. For example, there were 137 valid forecast cases at 72 h. By $\tau = 120$ h, there were only 75 cases. In contrast, ZTOT (using a 20% cutoff) produced forecasts for 125 cases at $\tau = 72$ h and 66 cases at $\tau = 120$ h (availability of 91.2% and 88.0%, respectively).

Figure 5.9 shows the best tracks of the cases included in the official verification during the relatively active 2001 Atlantic season (for a summary of the 2001 Atlantic season, see Beven et al. 2003). For comparison, Figure 5.10 shows the best tracks of all cases for the season, including the extratropical cyclone, wave, and remnant low stages that are excluded from the verification. Two long-lived hurricanes, Erin and the late-season Olga, provided many of the cases with verification at the longer time periods.

Figure 5.11 shows the best tracks for the 2002 Atlantic storms. Due to unfavorable atmospheric conditions associated with El Niño (Gray 1984), the 2002 Atlantic season featured suppressed formation in the deep tropics, with many weak subtropical systems forming at relatively high latitudes (for a summary of the 2002 Atlantic season, see Pasch et al. 2004). Long-lived hurricanes Kyle and Isidore provided many of the cases with verifications at the longer time periods for this season.

The very active 2003 Atlantic season featured a wide range of storms, from very intense and long-lived Cape Verde hurricanes Fabian and Isabel, to several meandering storms of subtropical origin. The season was especially unusual with three out of season storms (Ana, 20-24 April 20-24; Odette, 4-7 December; and Peter, 7-11 December). Figure 5.12 shows the best tracks for the 2003 Atlantic storms.

To summarize the effect of the differences in seasonal storm characteristics on ensem-

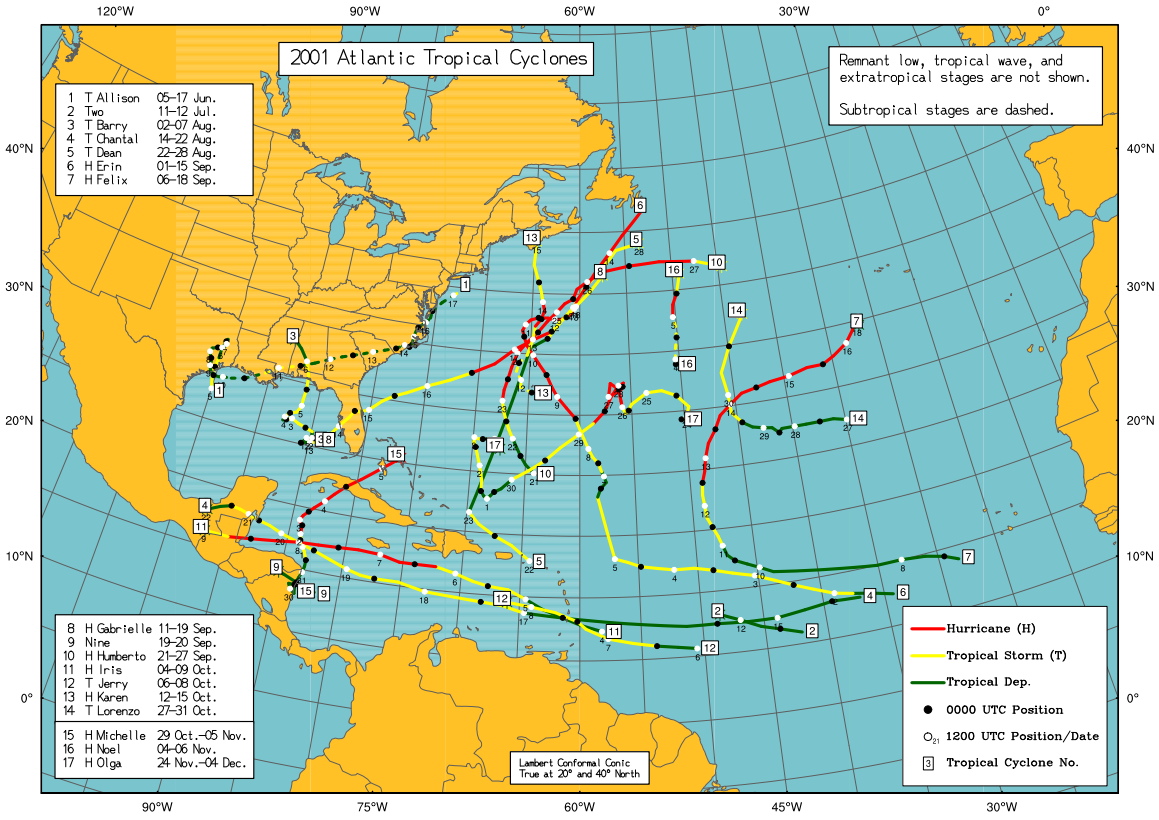


Figure 5.9: Tracks of the 17 tropical cyclones from the 2001 Atlantic hurricane season. Numbers in squares identify the storm, and filled and open circles give the storm position at 0000 UTC and 1200 UTC, respectively, on the indicated day. Tropical stages are shown as solid lines and subtropical stages are shown as dashed lines. The line color indicates the intensity (hurricane, tropical storm, or tropical depression based on wind thresholds of ≥ 64 kt, < 64 kt and ≥ 34 kt, and < 34 kt respectively). Remnant low, tropical wave, and extratropical stages are not shown, since these stages are not included for purposes of calculating official track verification statistics.

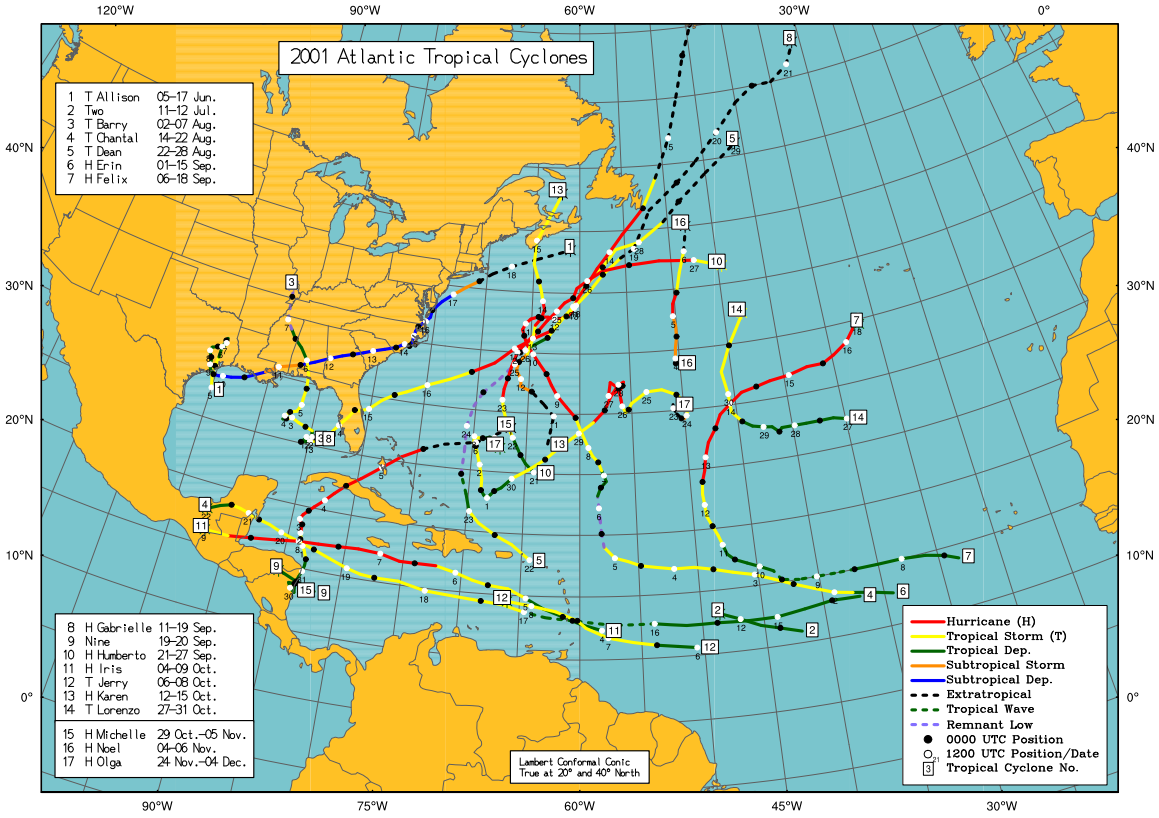


Figure 5.10: As in Fig. 5.9, but this figure also shows the stages that are not included in the official verification statistics. The tropical and subtropical storm stages are shown as solid lines, with line color indicating the intensity (hurricane, tropical storm, or tropical depression) and stage (tropical or subtropical). Remnant low, tropical wave, and extratropical stages are indicated by dashed lines.

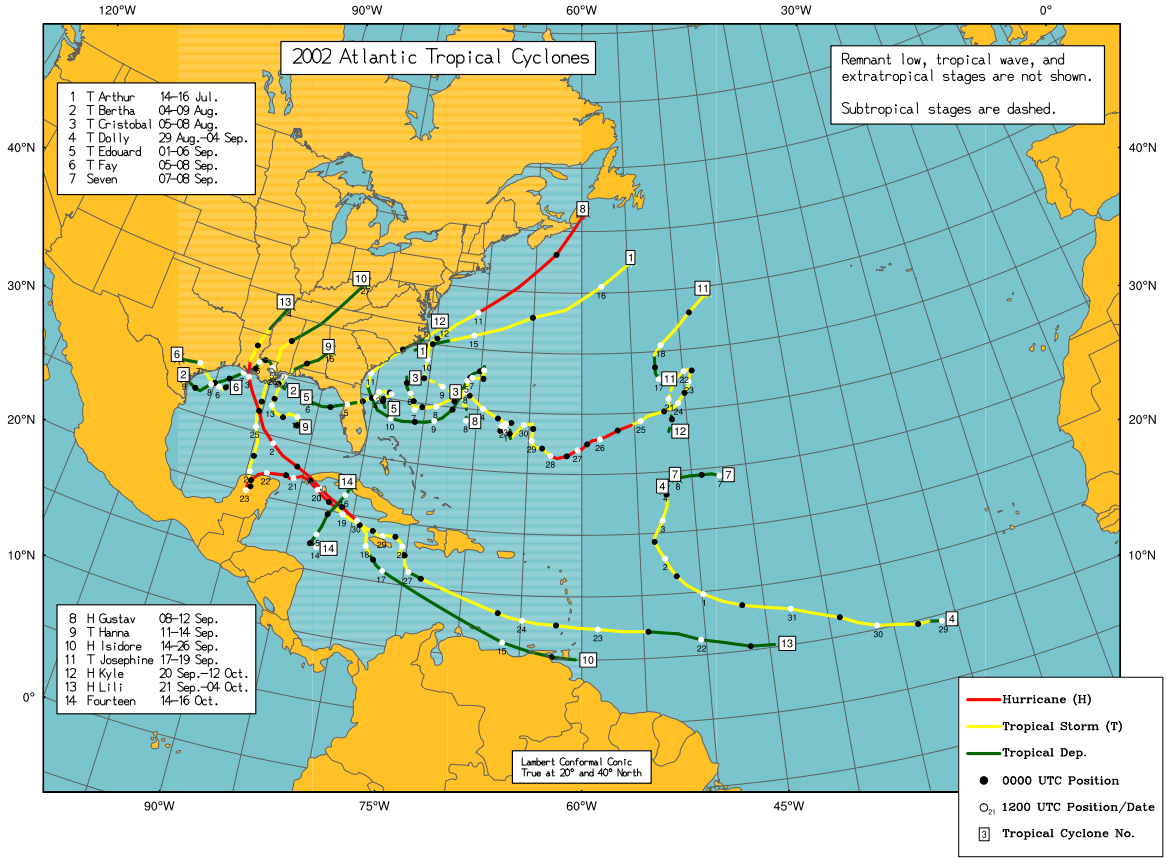


Figure 5.11: As in Fig. 5.9 but for the 14 tropical cyclones of the 2002 Atlantic hurricane season.

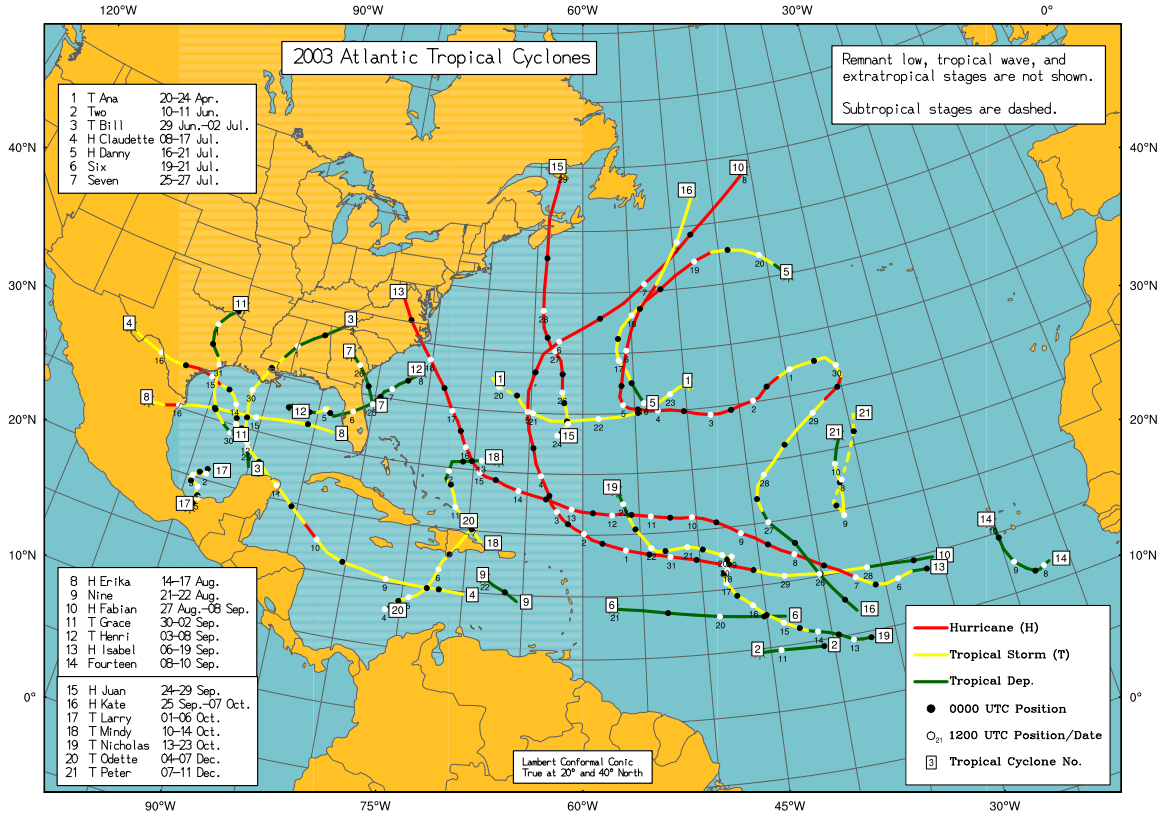


Figure 5.12: As in Fig. 5.9 but for the 21 tropical cyclones of the 2003 Atlantic hurricane season.

ble performance, Table 5.1 is given. This table shows the seasonal mean track errors for 5-day CLIPER model (CLP5) described by Aberson (1998) and the total ensemble mean forecast (ZTOT) using a 20% cutoff value. The seasonal mean errors are shown at each τ for each of the 2001, 2002, and 2003 seasons, and the average error over all three seasons is given below each section of the table. To give an overview of the total ensemble forecast availability, the number of cases with a verifying best track position are indicated in parantheses.

Marked differences in performance occur for CLP5. This may seem surprising, especially since the the configuration of CLP5 was static through the period (Sim Aberson 2004, personal communication). These differences are almost certainly due to differences in the characteristics of the storms in each year. The relative performance of a CLIPER model from one year to the next provides an estimate of the forecast difficulty associated with the cases of a given year (Neumann 1981; Pike and Neumann 1987). With this interpretation, the 2002 season featured the most difficult forecast cases on average, while 2003 had the easiest cases overall. In general, active seasons typically feature many formations in the deep tropics. Storms that develop in this region tend to behave in line with climatological norms, so CLP5 tends to perform better in active years.

In contrast, the ZTOT track errors have a secular decreasing trend at early and intermediate forecast periods. During the period of interest, many incremental upgrades have been made to the GFS global model and data assimilation scheme. These upgrades have resulted in improvements to the general model forecasts, which in turn may have lead to better kilo-ensemble forecasts. The GFS ensemble forecasts may have also improved due to changes in the generation of perturbations (Zoltan Toth 2004, personal communication). Whatever the cause, the trend towards decreasing errors in the kilo-ensemble forecasts (summarized by ZTOT) is almost certainly due to the improvements in the underlying wind fields of the GFS ensemble members. The performance trend is less clear for the later time periods, with 2002 putting in the best performance at the 96- and 120-h periods.

τ	0	12	24	36	48	60	72	96	120
CLP5									
2001	7.9	56.9	115.4	177.7	244.2	307.4	357.7	515.0	674.6
2002	7.6	52.3	113.7	188.1	261.8	338.9	415.9	593.3	754.5
2003	7.3	44.0	96.1	155.9	207.8	267.6	325.7	455.8	633.6
ALL	7.6	50.5	107.0	171.9	234.3	301.0	361.5	515.6	686.0
ZTOT									
2001	7.9	50.3	95.2	152.1	202.1	267.2	304.1	374.0	431.0
2002	7.6	47.4	86.6	140.3	184.2	238.3	273.5	345.0	416.5
2003	7.3	40.6	78.8	121.0	170.3	215.8	264.5	388.4	492.9
ALL	7.6	45.6	86.1	136.3	184.4	239.0	280.2	370.8	449.2
# Cases									
2001	(75)	(74)	(63)	(58)	(51)	(48)	(43)	(30)	(23)
2002	(65)	(64)	(54)	(51)	(43)	(42)	(36)	(31)	(26)
2003	(92)	(92)	(82)	(73)	(64)	(57)	(50)	(40)	(29)
ALL	232	230	199	182	158	147	129	101	78

Table 5.1: Season mean track error characteristics (in n mi) of the total ensemble mean forecast (ZTOT) and the 5-day CLIPER (CLP5) for the 2001, 2002, and 2003 Atlantic tropical cyclone seasons. The number of cases for each season is given in parentheses. The average error over all three seasons is given on the bottom line of each section. The total number of cases over all three years is given on the last line.

However, this anomaly may just be a consequence of sampling variability due to the low number of cases with verifying best track positions at these later forecast periods.

5.4.2 Domain Edge Issues

Several factors related to the verification domain and storm behavior may introduce artificial biases in the verification statistics, even though they do not relate directly to biases of the model or ensemble design. Briefly, these issues of artificial bias revolve around the problem of a finite model domain and the fact that there may be a nefarious exclusion of cases in certain forecast situations due to model unavailability. This problem can cause difficult forecast cases to be excluded from a homogeneous comparison. Despite the fact

that a 6000-km domain is used, there were many cases during the 2001-2003 seasons when the domain size was insufficient for a 120-h forecast. Two types of edge issues can result. Type-one edge problems occur when many individual ensemble members hit the domain wall and are 'lost', resulting in a dropped forecast for that member. Type-two edge problems happen when the verifying best track travels outside the domain. Often these issues occur jointly, as illustrated by the later forecast periods of the Hurricane Iris 5 October 2001 case shown in Fig. 5.13. In this case, the verifying best track traveled beyond the domain boundary, and many of the ensemble members were lost when they hit the domain wall. The members that remain (mostly, the incorrect recurving forecasts) slow the speed of the ensemble mean forecast and pull it to the north, resulting in an unrealistic track forecast.

The first edge issue was briefly discussed in Section 5.3.1. To summarize that discussion, some reduction of the type-one edge issue can be achieved by requiring that a minimum number (based on the cutoff percentage value) of ensemble member forecasts be available in order to calculate and produce a valid forecast for the ensemble mean. In type-one cases, it is usually obvious to a user when most of the ensemble members have hit the domain – the user can correctly infer that the ensemble mean forecast was trending in a given direction by visually extrapolating outside of the domain. But for objective measures-oriented verification (e.g., the skill or accuracy of the ensemble mean forecast), this issue is not so easily resolved. Using a stringent cutoff results in fewer forecast cases and may still artificially inflate the skill by excluding potentially bad forecasts (fast-moving storms tend to have larger errors). For example, using a 20% cutoff yields a total of 66 ZTOT forecasts at $\tau = 120\,\mathrm{h}$ (out of 75 possible cases). A more stringent 90% cutoff reduces the number of available forecasts down to 39. The compromise cutoff value of 50% leaves a total of 63 cases. Another potential solution is to produce an extrapolated forecast for all ensemble members whose forecasts have crossed the domain edge, then calculate the ensemble mean forecast. This is a messy solution however, and it is not clear that such an extrapolation would add any additional value to the ensemble forecast. The value of the

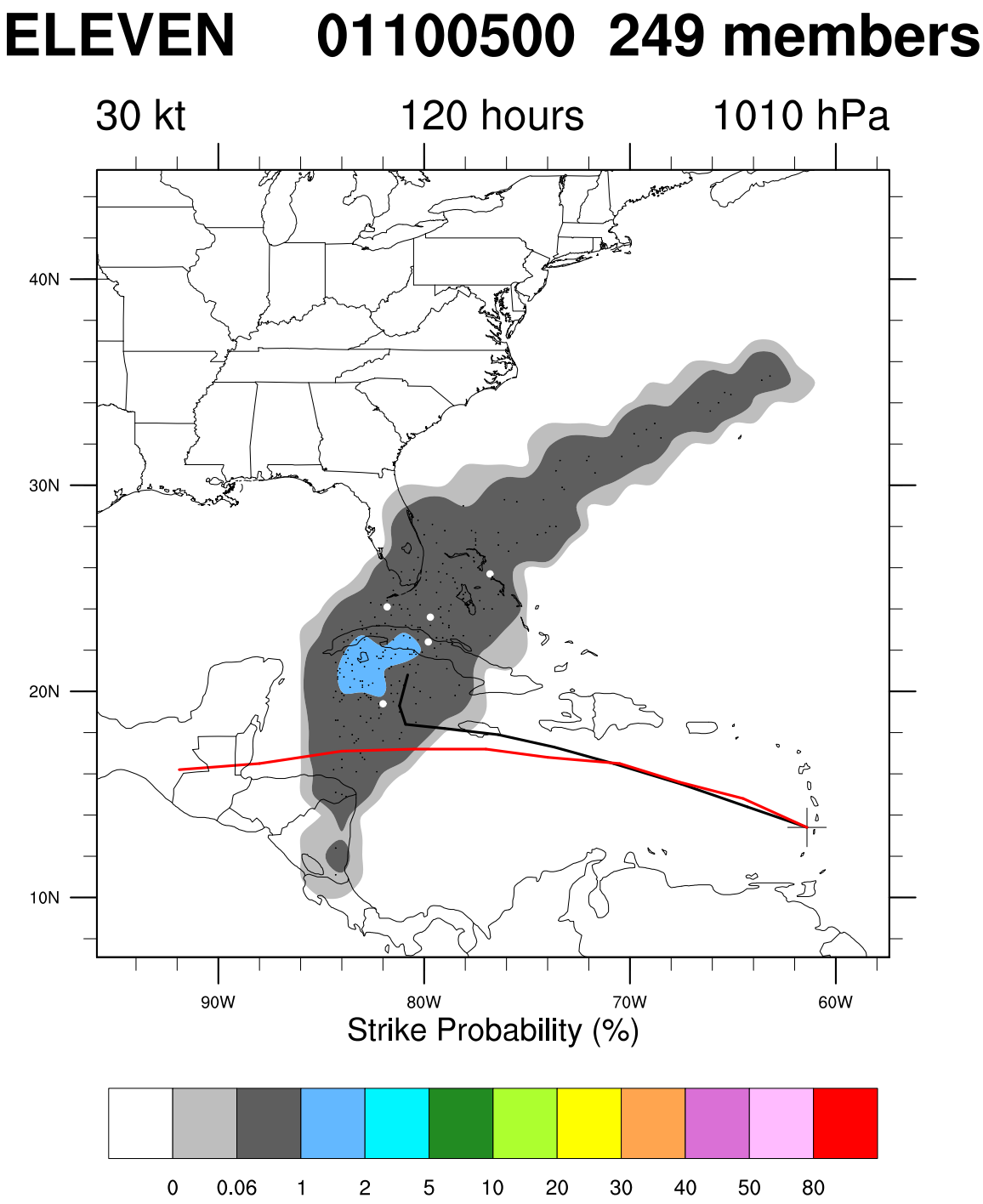


Figure 5.13: The 120-h frame of a swarm animation of the Hurricane Iris forecast starting at 0000 UTC on 5 October 2001.

ensemble forecasts at the later forecast periods may be of a dubious nature anyway, so it might be appropriate to only consider forecasts out to through $\tau = 84$ h or 96 h. However, this may be the time-frame during which the environmental ensembles begin to provide real benefit, so this is not a favorable solution either.

The most appropriate solution to this edge issue would be to increase the model domain size. The domain size for this study was chosen based on sensitivity experiments, as discussed in Chapter 3. Larger model domains were shunned in the ensemble design due to their larger track errors, probably because MUDBAR's barotropic dynamics degrade the GFS model's synoptic information as it travels into and across the domain. To allow a domain size that is appropriate for 5-day forecasts (perhaps 9000 km instead of the current 6000 km), a nudging term could be applied in a buffer zone near the edge of the domain to force the MUDBAR fields towards the global model deep-layer mean fields. This approach is currently used in LBAR. There could still be problems if the storm were to enter the buffer zone. The nudging solution would eliminate all or most of the edge issues and still retain the most possible number of forecast cases. This solution was not implemented in this study, but if the kilo-ensemble were to be turned into an operational track guidance product, it would be important to accomplish this.

It should be noted that the ensemble spatial strike probabilities are also affected by this edge problem, as seen in the 120-h cumulative strike probabilities for Hurricane Iris, given in Fig. 5.14. The strike probabilities end abruptly at the edge of the domain, although their trend is clear for a user looking at the swarm animation. Distribution-oriented verification measures will have biases since the total probability of the storm remaining in the domain at later time periods drops below 100% as members hit the domain edge and go missing. One stop-gap solution would be to employ a similar cutoff percentage of available members to determine when to exclude a forecast case from the verification.

The second edge issue occurs when the verifying best track travels outside the domain. One obvious solution is to not verify the forecast periods that do not have a verifying best

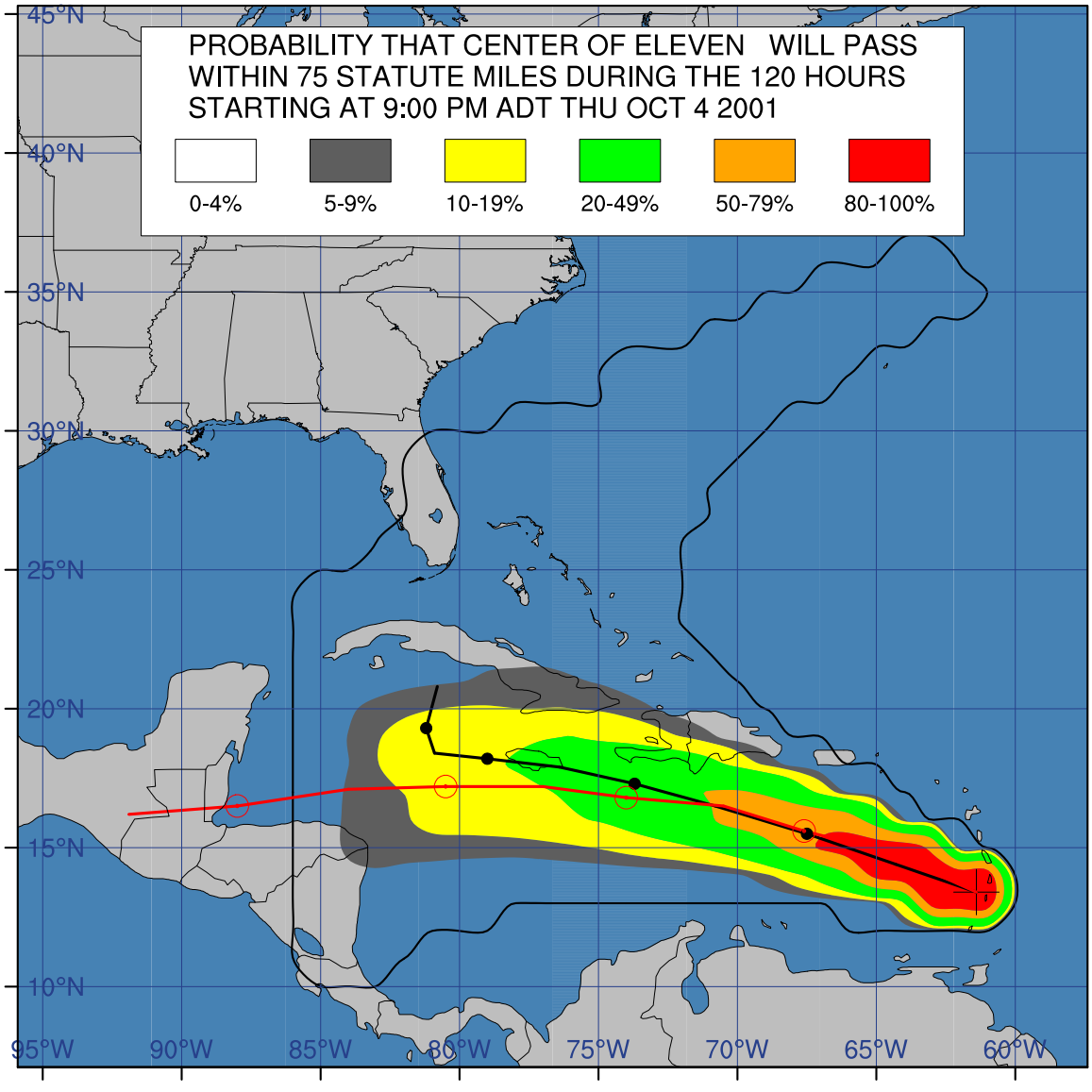


Figure 5.14: The cumulative strike probabilities through 120 h for the Hurricane Iris forecasting starting at 0000 UTC on 5 October 2001.

track within the model domain (this has not been done in this research, mainly due to time constraints). This issue is probably not as important as the first issue, given the following line of reasoning. If the best track travels outside the domain, most of the ensemble members *should* be outside as well, provided that the ensemble forecast is good. But if the ensemble forecast is poor and most of the members are still inside the domain while the verifying best track is outside, then the forecast is penalized as it should be. Thus, type-two edge issues present less of a problem in the verification statistics. The skill may still be artificially lowered however, because these measures cannot reward the ensemble for members which would have been in the vicinity of the actual verifying position but were lost at the boundary.

Both of these issues tend to occur for rapidly moving storms. These cases generally break down into two types of forecast situations: a rapidly recurving storm that is caught up in the westerlies, or a storm embedded in fast easterly flow in the deep tropics (e.g., the 5 October 2001 Iris case). Whatever solution is chosen in dealing with these edge issues, it is important to realize that any solution which affects forecast availability may itself introduce artificial bias by deselecting forecast periods for cases that would have been verified under a different solution. Because the x- and y-bias statistics are strongly affected by outliers, these statistics are especially vulnerable to the forecast availability or lack thereof. The upshot of this discussion is that extreme care should be taken in interpreting any ensemble results at the later forecast periods. In general, the domain size limitations probably limit the validity of kilo-ensemble forecasts beyond 84 h unless the domain size is increased with the use of a nudging term.

To determine how much uncertainty is introduced into the verification statistics by these edge issues, the errors and x- and y-biases are examined for three cutoff values: 20%, 50%, and 90%. The details are not shown here, but the conclusions are that the 50% cutoff offers a decent compromise between keeping as many cases as possible, while keeping the introduction of artificial biases to a minimum. The mean track error and bias statistics both seem to be robust for all cutoff values through $\tau = 72$ h. By 120 h, the mean track

errors vary by 4-6%, with the 90% cutoff forecasts possessing the largest errors. The 120-h x- and y-bias statistics are much more strongly affected, with large uncertainties apparent. The 90% cutoff forecasts have a much smaller x-bias than the 50% cutoff forecasts (ZT50), but a much larger y-bias. This suggests that the bias statistics are unreliable beyond 72 h – it is difficult or impossible to separate the bias due to forecast unavailability from the actual model bias. Therefore, for the remainder of this study, the 50% cutoff is used as *the* standard total ensemble mean forecast (ZT50). For some analyses and plots, ZTOT (20% cutoff) was used, but the two sets of forecasts are for all practical purposes the same, since they have very similar error and bias characteristics.

5.4.3 Measures-oriented Verification Methods

The following measures-oriented measures are used to characterize the ensemble performance: accuracy, skill, and bias.

Accuracy is the average correspondence between individual pairs of forecasts and observations. In this context, it is expressed by the mean track error over a set of forecast cases. In TC track verification, it is customary to determine the seasonal mean track error by averaging the track errors over all cases in a basin for that particular storm season. This is done for each τ.

Skill is the measure of accuracy relative to some baseline forecast such as climatology, persistence, or a forecast randomly selected from the distribution of previous observations. For TC track forecasting, it is customary to use the climatology and persistence forecast as the baseline for skill. This work uses the 5-day CLIPER model (CLP5) as the baseline for skill. Skill (or relative error) is calculated using:

$$Skill = \frac{E_{model} - E_{CLP5}}{E_{CLP5}} \times 100\%, \tag{5.4}$$

where E_{model} is the seasonal mean track error of the model forecast and E_{CLP5} is the seasonal mean track error of the CLIPER model forecast.

Bias is the correspondence between the mean of the forecasts and the mean of the

observations. Since storm tracks occur in a 2-dimensional domain, the bias is decomposed into x- and y-components. Thus, the x-bias is the average (over the cases under consideration, such as in a season, or several seasons) of the x-component of the model's track error, and similarly, the y-bias is the average of the y-component of the track error.

Verification of the Total Ensemble Mean Forecast

Verification statistics of mean track errors, x- and-y biases, and frequency of superior performance are compiled using a verification program similar to the one used by TPC/NHC.[12] The program also conducts a statistical t test for differences between each model in the suite, following Franklin and DeMaria (1992). The level of significance is taken to be 95%. For more on taking serial correlation into account in tests of the mean, see Zwiers and Storch (1995).

To eliminate errors caused by sampling variability, it is imperative that all seasonal forecast verifications be homogeneous – that is, for a given suite of models to be verified, a forecast case is included in the calculation of the verification statistics *if and only if* valid forecasts exists *for all* the models in that suite. If one or more model forecasts is missing, that forecast case is excluded. This can cut down the number of valid cases if there are many missing cases from one or more models in the verification suite, so verification of too many models at a time can be troublesome. Unless noted otherwise, all the following results are based on a homogeneous comparison for the models in that suite.

Figure 5.15 shows the mean track errors for all available cases of the 2001-2003 Atlantic hurricane seasons. Four models are compared: the 5-day Climatology and Persistence Model (CLP5), the MUDBAR forecast that is run from the GFS control member (MBAR), the control member of the kilo-ensemble (A5K0), and finally, the total kilo-ensemble mean forecast (ZTOT) calculated using a 20% cutoff. All three MUDBAR-based forecasts have similar error characteristics, with 72-h errors of approximately 260 n mi and 120-h errors of

[12] A slight modification was made: the exact Great-Circle distance formula is used instead of the approximate distance formula given earlier in the chapter.

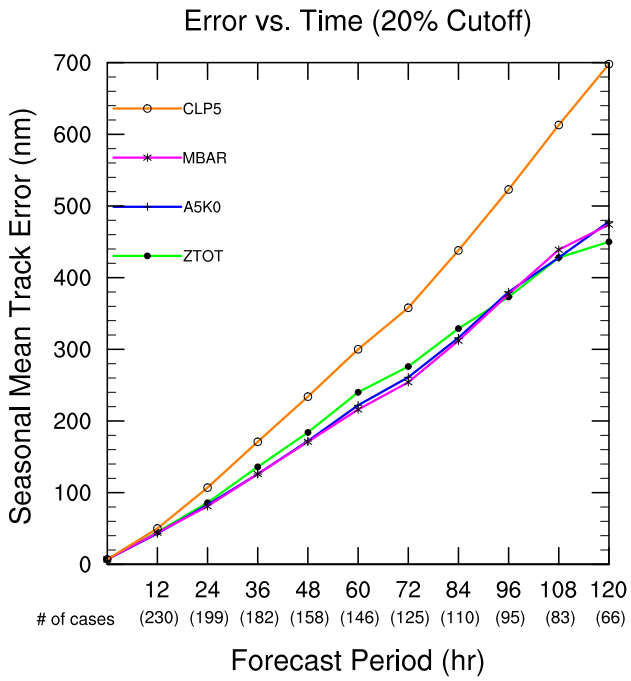


Figure 5.15: Mean track errors (in n mi) for all available cases from the 2001-2003 Atlantic hurricane seasons for 5-day Climatology and Persistence model (CLP5), for MBAR which is the run of MUDBAR using the GFS control forecast, for the kilo-ensemble control forecast (A5K0), and for the total kilo-ensemble mean forecast (ZTOT).

450 n mi. The CLP5 errors are significantly worse, as expected.

One of the key questions of this research is whether the total ensemble mean forecast can improve over a single deterministic forecast, such as the ensemble control forecast. To answer this question, the skill of ZTOT is calculated using the ensemble control forecast (A5K0) as the baseline. This is shown in Fig. 5.17. For comparison, Fig. 5.16 shows the skill relative to CLP5 for these models. The ZTOT forecasts were not able to improve over the control forecast until the end of the forecast period. The 2-sided t test indicates that model differences were not significant for any forecast periods beyond 60 h. The differences were significant up through 60 h (except at 24 h), or more properly, the null hypothesis that the two means are not different can be rejected at the 99% significance level for these time periods. Thus, the apparent differences between the models at later forecast periods are not significant – there is a high likelihood that they are caused by sampling variability.

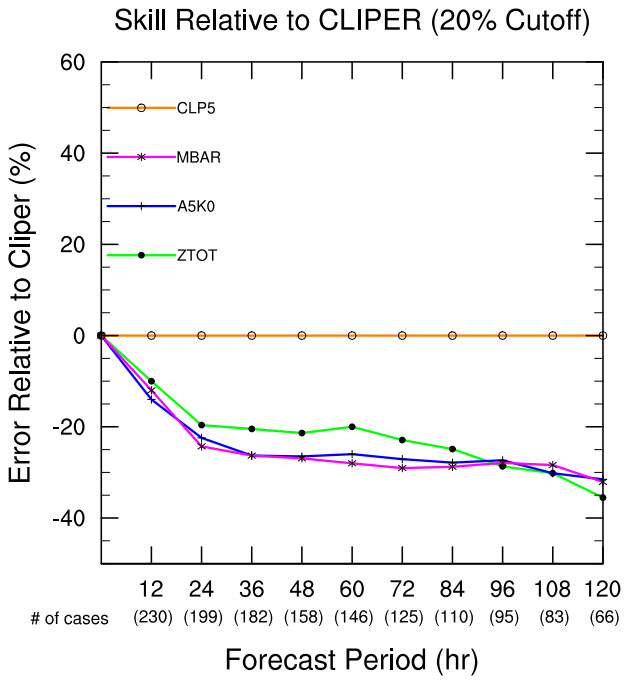


Figure 5.16: As in Fig. 5.15 but for error relative to CLP5. Negative values indicate positive skill.

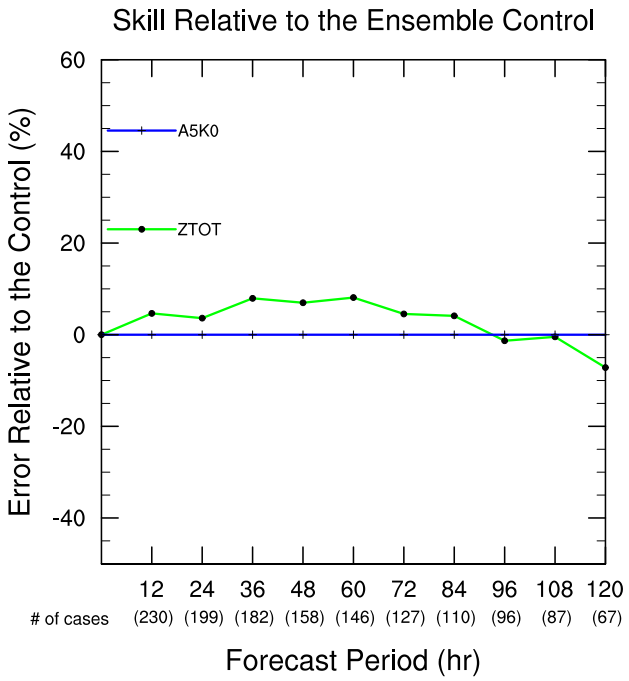


Figure 5.17: The total ensemble mean forecast (ZTOT) error relative to the control forecast (A5KO) for all available cases from the 2001-2003 Atlantic hurricane seasons.

This failure of the total ensemble mean to improve over the control forecast is not really so surprising. The ensemble mean is actually a hedged forecast. By its very nature, it represents a geographical 'center of mass' for all the ensemble members, but there is no guarantee that the actual highest mass density will be found at the ensemble mean forecast. The ensemble mean is easily biased by outliers. This simple example illustrates one of the weakness of using the total ensemble mean forecast as a measure of ensemble performance.

Another way to compare ZTOT and A5K0 is to determine how often one model was better than the other. This 'frequency of superior performance' is shown in Fig. 5.18. The control forecast was better than ZTOT about 60% of the time for all forecast times.

The bias statistics show how close the mean of the forecasts correspond to the mean of the observations, averaged over all forecast cases. Figure 5.19 shows the mean track x-bias over the three seasons, while Fig. 5.20 shows the mean track y-bias.

Another way to judge the kilo-ensemble performance is to compare it to the perfor-

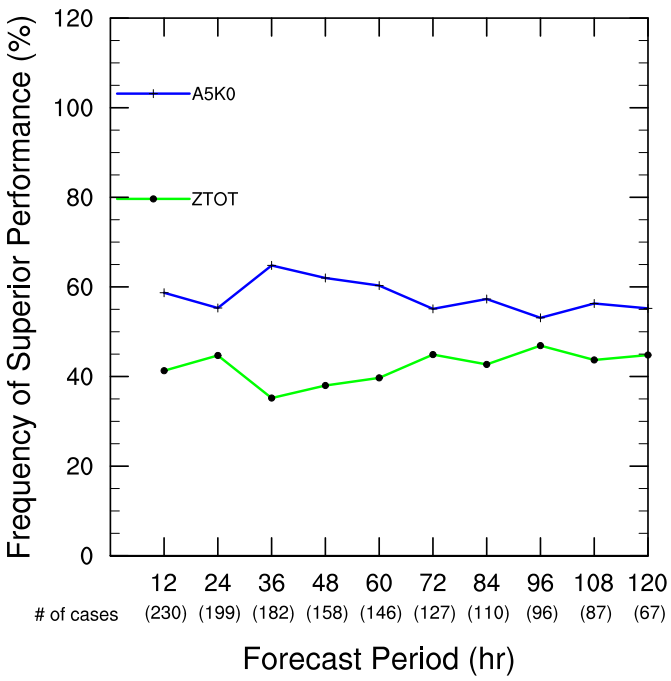


Figure 5.18: The frequency of superior performance for ZTOT and A5KO for all available cases from the 2001-2003 Atlantic hurricane seasons.

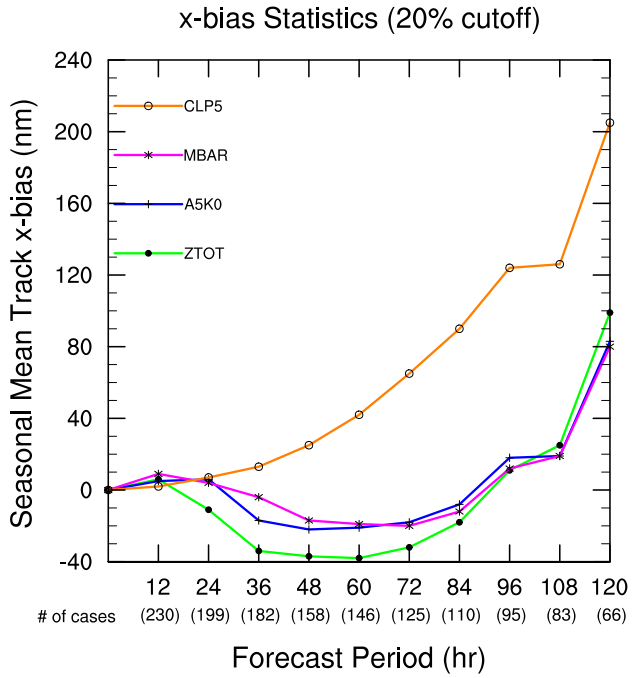


Figure 5.19: The mean track x-bias for all available cases from the 2001-2003 Atlantic hurricane seasons. Positive values of x-bias indicate that the average forecast location was further west than the average observed storm location.

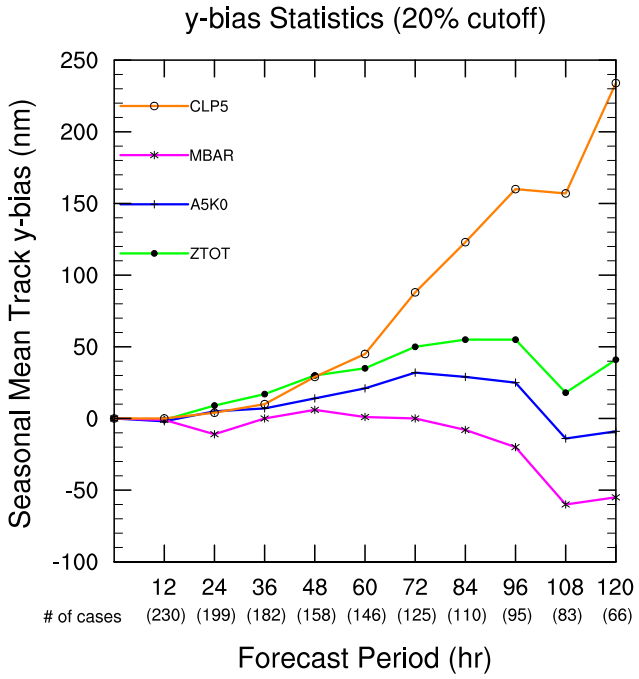


Figure 5.20: The mean track y-bias for all available cases from the 2001-2003 Atlantic hurricane seasons. Positive values of y-bias indicate that the average forecast location was further north than the average observed storm location.

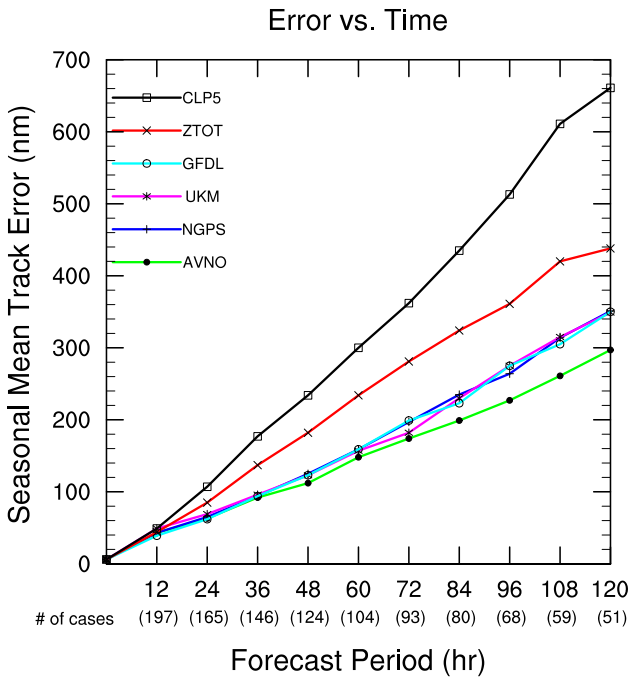


Figure 5.21: As in Fig. 5.15 but for the skillful global and regional models.

mance of other models. Figure 5.21 shows the combined seasonal mean track errors for the regional GFDL hurricane model (GFDL) and the track forecasts from three global forecast models: the NCEP GFS model (AVNO), the Navy NOGAPS (NGPS) model, and the UK Met Office model (UKM). The performance of these models represents the current practical predictability limit, since they are the current state-of-the-art TC track forecast models. There is no expectation that the kilo-ensemble should outperform these models, given the great simplification in using barotropic dynamics. For convenience, the track errors of CLP5 and ZTOT are also included.

Plots of skill relative to some baseline provide another means to judge forecast performance. CLP5 is used as the baseline for skill, since it only uses information available from climatology (previous history) and persistence (current trend of the storm). The skill is calculated as percent difference from the baseline, so negative values indicate errors that are smaller than the baseline: positive skill. Figure 5.22 shows the skill relative to CLP5 for ZTOT and the current state-of-the-art track models.

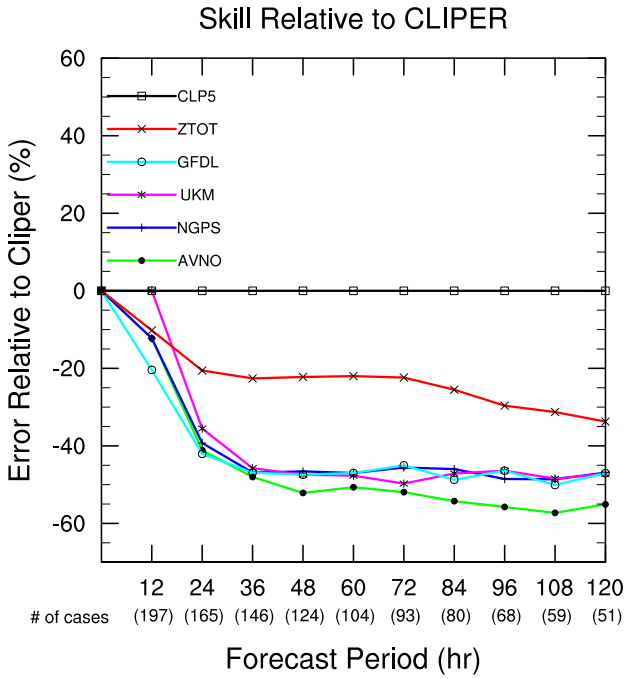


Figure 5.22: As in Fig. 5.21 but for the error relative to CLP5. Negative values indicate improvement over the CLP5 forecasts.

Finally, I would be remiss not to examine the performance of the underlying GFS ensemble members, since the kilo-ensemble uses the wind field forecasts from these ensembles. Figure 5.23 shows the errors for the 2002-2003 seasons. There are no cases available after 84-h, probably because the GFS ensemble members were truncated to T62 resolution at 84 h, so vortex tracking becomes difficult (the resolution of the ensemble members was increased to T126 for the first 180 h in late 2003). The AVNO and GFS control member forecasts are far superior to the other perturbed members. Interestingly, the MBAR forecast is also superior to all of the perturbed GFS members through 48 h. This is likely due to the lack of vortex relocation in the GFS ensemble members, whereas MBAR uses the GFS control which does use vortex relocation. Figure 5.24 shows the skill relative to CLP5. This figure dramatically illustrates the low skill of the perturbed GFS members at early forecast times.

Verification of the Subensemble Mean Forecasts

Verification was conducted for each of the subensemble mean forecasts, as was done for the total ensemble mean forecast above. Most subensembles have skill comparable to their respective peers of the same class. Since these error and skill statistics are similar to the overall ensemble mean results, these plots are not given. As expected, the plots of x- and y-bias were generally in line with the sensitivity experiments of Chapter 4. Plots of the frequency of superior performance were somewhat surprising at first glance. As an example, Fig. 5.25 shows the frequency of superior performance for the subensembles based on perturbations to the motion vector. Surprisingly, the total ensemble has superior performance very little of the time. But the subensemble based on no motion vector (SMV0, a control subensemble in this regard), has even lower frequency of superior performance. After a little thought, this result is not really surprising after all – it is simply a reflection of the fact that the total ensemble mean and the SMV0 control subensemble forecasts must always be centered in the midst of the respective perturbed subensembles. Since it

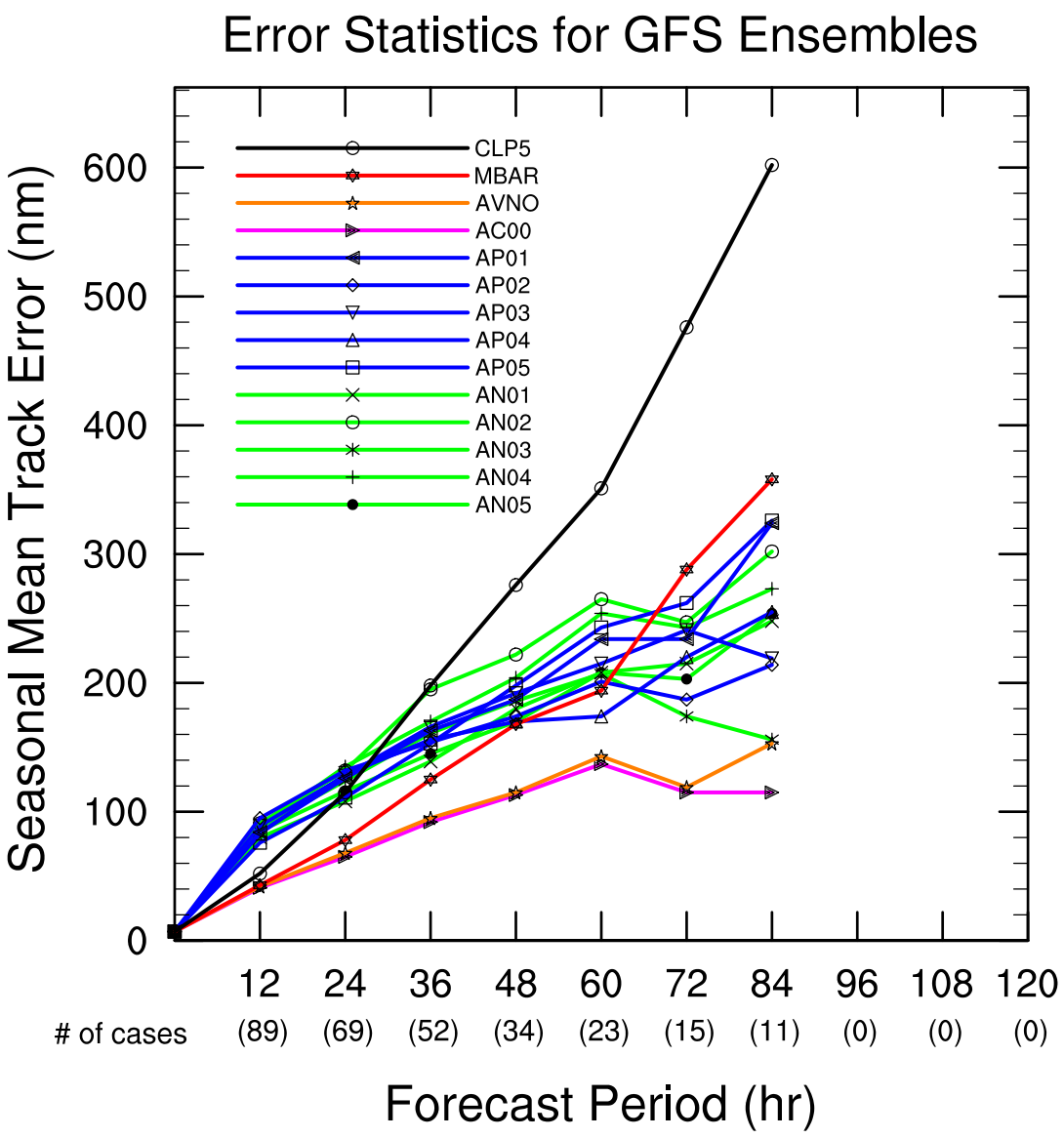


Figure 5.23: Mean track error for the 2002-2003 seasons for the GFS ensembles, Aviation (AVN0), and MBAR models.

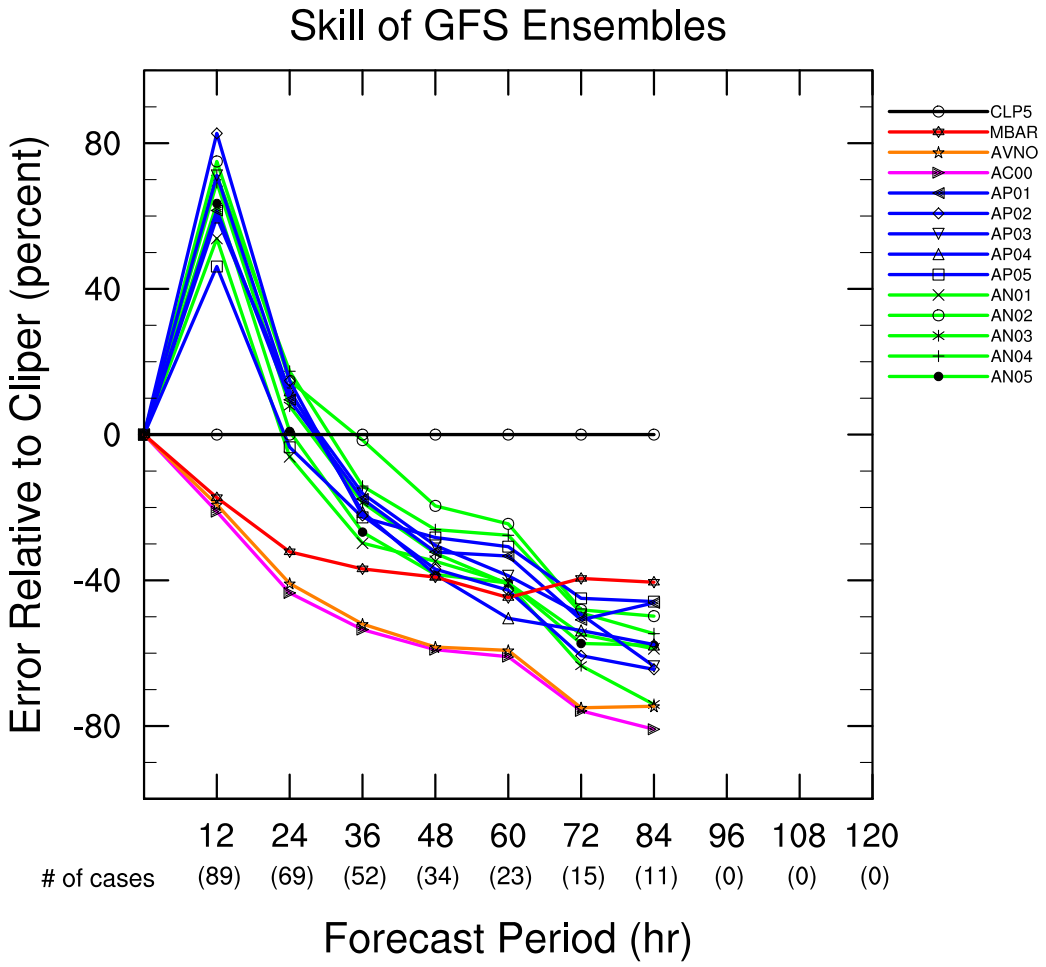


Figure 5.24: The skill (relative to CLP5) of the GFS ensembles, the Aviation (AVN0), and MBAR.

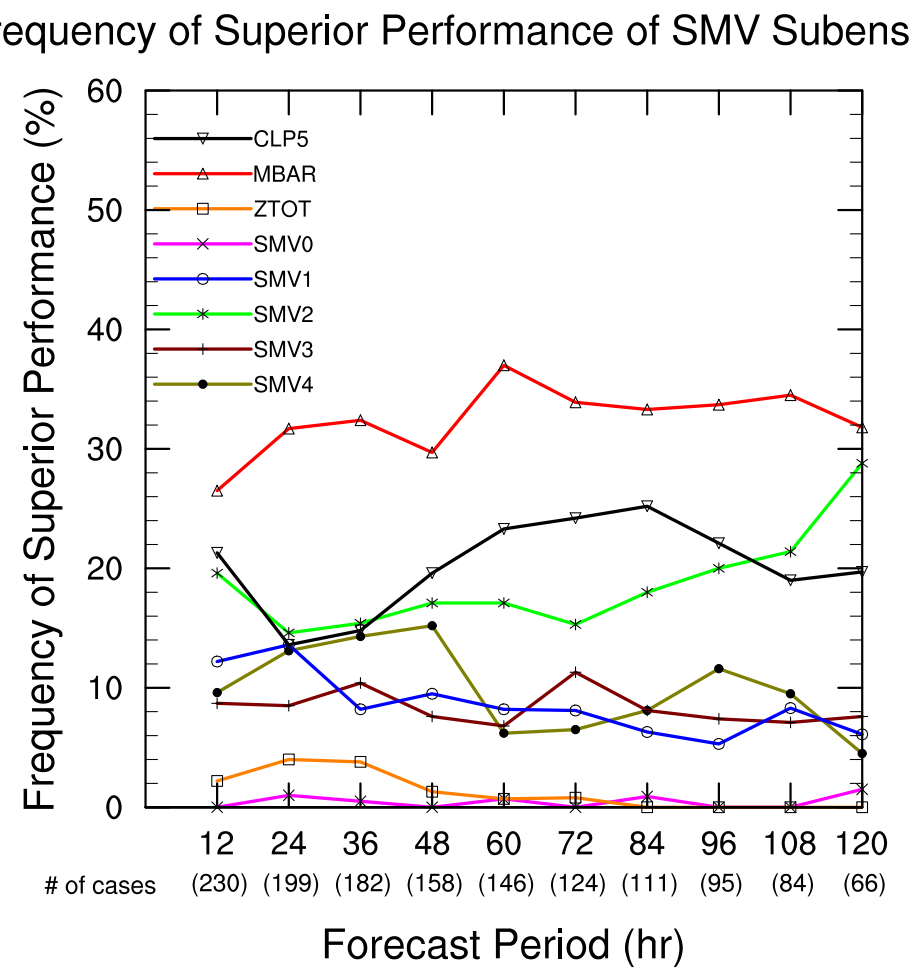


Figure 5.25: The frequency of superior performance for the subensembles based on perturbations to the motion vector.

is extremely rare for the verification to fall right at the center of the ensemble (that would be a near-perfect forecast), there is very little chance for forecasts in the center to have a superior forecast – one of the perturbed subensembles on the outside edge will almost always capture the prize.

One important result of the subensemble mean verification is the profound lack of skill for the subensemble based on the shallow layer mean perturbation (SLY7). This is summarized by Fig. 5.26. The mean forecasts of the SLY7 subensemble was worse than the other subensembles in that perturbation class. Since 495 of the 1980 total ensemble members use the shallow layer mean wind, this perturbation is substantially degrading

the overall ensemble forecasts when averaged over all forecast cases. However, when the frequency of superior performance is examined, a different story emerges, as shown by Fig. 5.27. The SLY7 gave superior forecasts about 20% of the time, which was just as good as any of its peers. This suggests that in some forecast cases, the SLY7 subensemble is beneficial. These are, of course, the cases in which the lower depth of mean steering current is more appropriate. This suggests that the ensemble performance could be improved if the inclusion of perturbations was based on the forecast-specific situation rather than fixed parameter values.

5.4.4 Distribution-oriented Verification Methods

Measures-oriented verification methods are often too simplistic to provide a good picture of performance, given the multifaceted aspects of quality. For instance, the seasonal mean track error gives no information about the distribution of errors other than its center. It is also interesting to ask questions like: how often did the ensemble produce forecasts with extremely large errors. To capture a greater understanding of the true performance characteristics, it is necessary to use verification methods that examine the entire distribution of the forecasts and observations. Such approaches are known as distribution-oriented methods. One example is discriminant analysis of the joint distribution between forecasts and observations. Other measures use conditional distributions. In the verification of probabilistic forecasts, there are many factors to consider (see Murphy and Epstein 1967). Some scoring measures include:

- probability score (Brier Score)
- 'reliability' and 'resolution' scores
- information quantity
- information ratio
- 'validity' measure

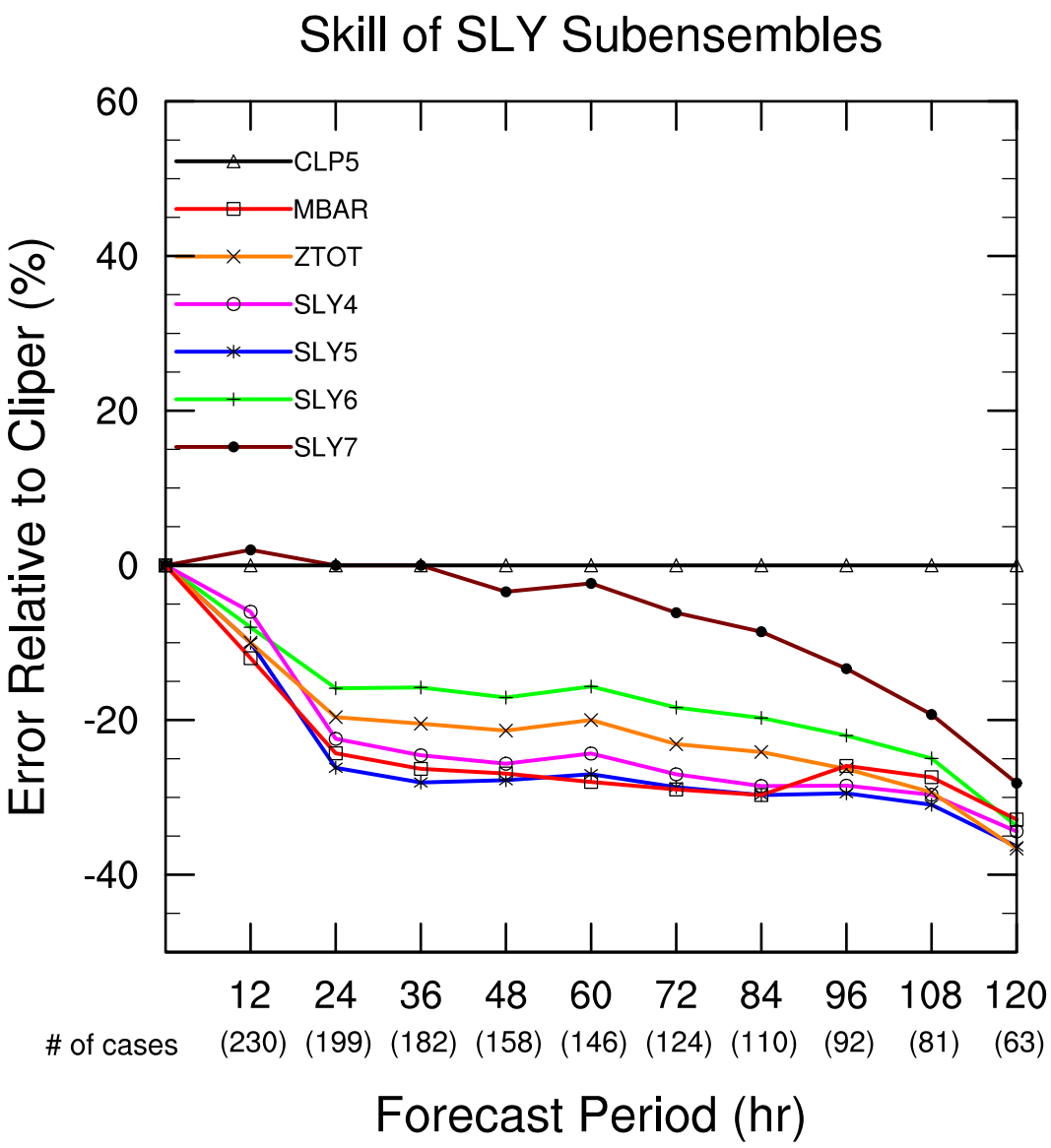


Figure 5.26: The skill of the subensembles based on perturbations to the deep layer-mean wind averaging.

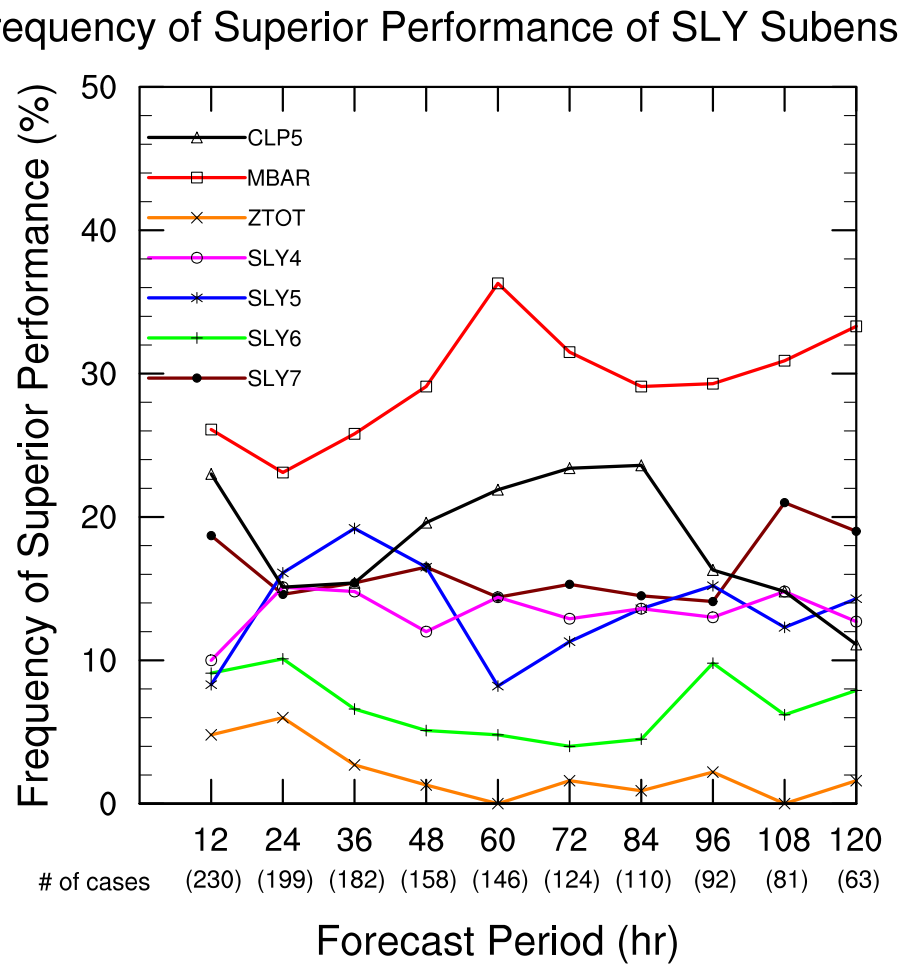


Figure 5.27: The frequency of superior performance for the subensembles based on perturbations to the averaging depth of the deep layer-mean wind.

- 'skill' score
- distance measures

The Brier Score (Brier 1950) is one of the most widely-used distributions-oriented measures for verification of probability forecasts. It is particularly advantageous because it is a *proper* score – one that does not reward 'hedging'. It is also easy to understand and calculate. For binary events (in this case, an event is the passage of a TC within 75 statute miles of a given location), the Brier Score (BS)[13] is defined as:

$$BS = \frac{1}{n}\sum_{j=1}^{n}(f_j - x_j)^2, \tag{5.5}$$

where n is the number of forecast occasions, f_j is the forecast probability for the occurrence of the event, and the observation x_j takes the value 1 or 0 depending on whether the event occurred or not. A forecast occasion is a single forecast of strike probability at a given location. In this study, there are two types of probability: 'instantaneous' strike probabilities[14] or cumulative strike probabilities from $\tau = 0$ to 72 or 120 h. The number of forecast occasions n is simply the total number of gridded strike probability forecasts.[15] The summation can also be taken over all the forecast occasions for an entire season (or multiple seasons), i.e., the summation is taken over all gridded strike probabilities (instantaneous or cumulative) for each forecast case and over all cases.

When the Brier Score is calculated, the result is just a number. A perfect forecasting system, one which always forecasts an event to occur when it occurs, and never forecasts it when it doesn't occur, has a BS of 0. A system which is always wrong, other other hand, will have a score of 1. The determination of the skill of a forecasting system requires a comparison with the BS of some other probability forecasting system. One of the difficulties

[13] Actually, this is the *half* Brier Score, since for binary forecasts, it is only necessary to sum over one category of forecasts. If the event occurrence matrix was categorical (with more than 2 categories) instead of binary, it would be necessary to sum over each category as well.

[14] In reality, these strike probabilities are for a 12-h forecast window ending at time τ.

[15] In this study, the probabilities are calculated on a $1^\circ \times 1^\circ$ grid, but the total number of grid points varies since MUDBAR uses the Mercator projection

in applying probability-based skill scores to the hurricane track problem is that it is not clear what to use as a baseline for skill. To create such a baseline, it is necessary to generate probabilities that somehow correspond to the expectation of zero skill. If climatology and persistence are used as such a baseline, a crude method would be to generate probabilities using the error distribution for a CLIPER-type forecast. More elaborate statistical schemes could be used as well. Of course, other baselines could be used. For example, it would be reasonable to compare the ensemble probabilities to probabilities generated from a single deterministic track forecast such as the control forecast. A nomogram-type approach could be used to generate the probabilities for such a track forecast using a method similar to the original NHC strike probability methodology (Sheets 1985). When such a reference forecast is created, then the Brier Skill Score (BSS) can be defined (following Jolliffe and Stephenson 2003, p. 145) as:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{5.6}$$

where BS is the Brier Score of the forecasting system (in this case, the kilo-ensemble), and $\mathrm{BS}_{\mathrm{ref}}$ is the Brier Score of the reference forecasting system, which could the climatological probability of occurrence, the probability forecast based on a single deterministic forecast, or another ensemble forecasting system. Unlike the Brier Score, positive values of Brier Skill Score indicate better forecasts: a perfect forecasting system will have BSS = 1, while a system which only performs as well as the reference will have BSS = 0. Negative values of BSS indicate a forecast system that is worse than the reference system.

A disadvantage of the Brier Score is that it does not give any indication as to how *useful* the forecasts are, just the correspondence between the forecast probabilities and the event occurrence. To get an idea of how useful the forecasts are, it is necessary to know something about the frequency at which forecasts are made for each of the different categories (or in this case, probability thresholds). It is also instructive to look at the conditional probability of event occurrence given the forecast of a certain probability category (Brier 1950, Jolliffe and Stephenson 2003). For instance, if the event was 'rain' or 'no rain', it

would be useful to know the percentage of times it rained when the forecast probability for rain was 90%. If it only rained 30% of the times when a 90% forecast was issued, this indicates that the forecasts aren't very useful at that probability threshold. Ideally, the forecast probability of event occurrence for a given probability threshold should exactly match the event occurrence frequency (i.e., if the forecast calls for 90% chance of rain, it should rain 9 times out of 10. Similarly, if the forecast calls for a 20% chance of rain, then it should only rain 20% of the time). If this is true, the forecast is said to be perfecly *reliable*. Other measures of forecast usefulness include *resolution* and *sharpness*. The Relative Operating Characteristics (ROC) score is a measure which is particularly suited to summing up the performance of a probability-based forecast system. The ROC score is described by Mason and Graham (1999) and Kharin and Zwiers (2003), and is essentially an extension of the two-by-two contigency table which involves the number of hits and misses, as compared to the number of 'yes' and 'no' forecasts. The result is the ROC curve, which allows one to see many important aspects of the joint distribution between the forecast probabilities and the observations of event occurrences. Due to time constraints, the computation of the Brier Skill Score and ROC Score is left for future work.[16]

Reliability of the Ensemble Envelope

One key question that should be asked of any ensemble is: "how well can the ensemble envelope encompass reality?" If the ensemble is perfectly reliable, then the verification should lie within the ensemble-predicted range at all times (this is a somewhat narrower definition of reliability than discussed in the previous section). This possibilistic interpretation seeks to determine if the ensemble is able to correctly simulate the subspace of all dynamical pathways available to the system. Of course, if the ensemble spread is always extremely large, the envelope will often contain the verification. For imperfect ensembles, having perfect reliability comes at a price – the envelope must be so large as to become

[16] The author anticipates that these issues will be solved shortly, and that these results will be published in a paper at some point in the not-too-distant future.

meaningless.

Figure 5.28 shows the reliability of various ensemble envelopes based on thresholds of instantaneous (not cumulative) strike probability at each forecast period. 0%, 1%, 2%, 5%, 10%, 20%, and 50% strike probability thresholds are displayed. As an example, the envelope based on the 50% strike probability threshold contains the verifying best track position about 65% of the time at 12 h, but only about 11% of the time at 24 h. Another way to interpret this figure is to draw a line at the 50% frequency threshold on the left axis. The verifying position is more likely to be inside the envelope for all probability thresholds that are above the 50% line, and less likely below that line. As the 50% line is traversed across the figure, one can read off the probability threshold values of the envelope reliabilty curves and determine the forecast period at which the actual verifying best track escapes the ensemble envelope. The envelope escape time was about 15 h for the 50% probability threshold, 26 h for the 20% threshold, 30 h for the 10% threshold, 42 h for the 5% threshold, and 64 h for the 2% threshold. The outer envelope of the ensemble, corresponding to the 0% probability threshold, or equivalently, to the region of 'possibility', contained the verification about two-thirds of the time at 120 h. This reliability information is quite useful for forecasters, since it allows them to know off-hand if the ensemble envelope is indicating higher than normal predictability. For instance, if the 72-h forecast position of the total ensemble mean lies within the 20% contour, then this figure indicates that the forecast uncertainty is much less than normal. Forecast uncertainty often varies significantly from one part of the basin to another, and also by latitude (storms in the deep tropics tend to be more predictable than storms further poleward). Ideally, once a large enough data set has been compiled, it may be beneficial to construct reliability diagrams for different parts of the basin.

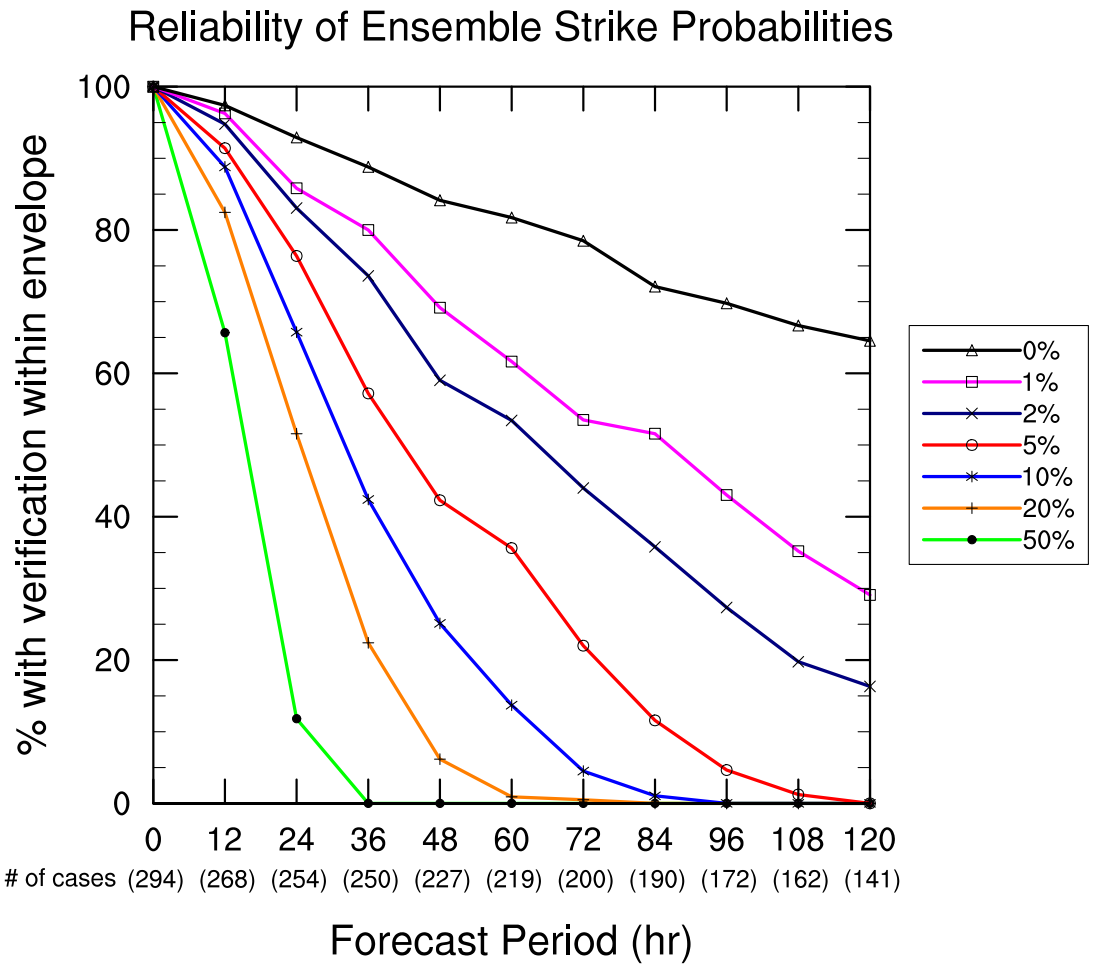


Figure 5.28: The reliability of ensemble envelopes based on instantaneous strike probability thresholds of 0%, 1%, 2%, 5%, 10%, 20%, and 50%, shown for each forecast time. Frequencies are computed based on all available forecast cases for the 2001-2003 Atlantic hurricane seasons.

5.4.5 Spread vs. Error Relationship

The relationship between the spread about the ensemble mean forecast and the error of the ensemble mean forecast is an important indicator of how well the ensemble is able to forecast uncertainty. When the ensemble mean forecast position is indistinguishable from the verifying storm position, the ensemble spread about the ensemble mean forecast will be identical to the mean of the track errors of the individual ensemble members. Thus, for a perfect ensemble, the spread about the ensemble mean gives the value of the expected error associated with the ensemble mean forecast. Figure 5.29 shows the maximum and minimum spreads and track errors for all the individual forecasts of the total ensemble mean (ZTOT) during the 2001-2003 Atlantic hurricane seasons. Also shown are the average spreads and track errors over that period. In general, the average spreads about the ZTOT forecasts are smaller than the average track errors of the ZTOT forecasts, especially at later time periods. This tendency for the ensemble spread to grow more slowly than the actual track error is a typical result of most or all TC ensemble track studies. It indicates that the ensemble is not simulating all the sources of uncertainty in the forecast problem.

If the ensemble is able to accurately simulate the sources of forecast uncertainty, then the case-by-case size of the ensemble spread can be interpreted as an *a priori* estimate of the forecast skill of the ensemble mean forecast. In real (imperfect) ensembles, the interpretation of the spread-error relationship becomes quite murky. Before discussing the two ways that the interpretation can be wrong, let's first consider the positive side: there are two ways that the spread-error relationship can be correct: (1) the case of small spread and small observed error and (2) the case of large spread and large error. The following statistics are for the spreads and errors based on the total ensemble mean with a 50% cutoff: ZT50. Figure 5.30 shows an example of the first type of 'correctness' for the 9 September 2001 case of Tropical Storm Erin. The 72-h spread is quite small, measuring 122 n mi (compared to the average 72-h spread of 206 n mi), and the track error of 105 n mi is minimal compared to the 72-h average track error of 309 n mi. These types of cases are apparently highly predictable, and

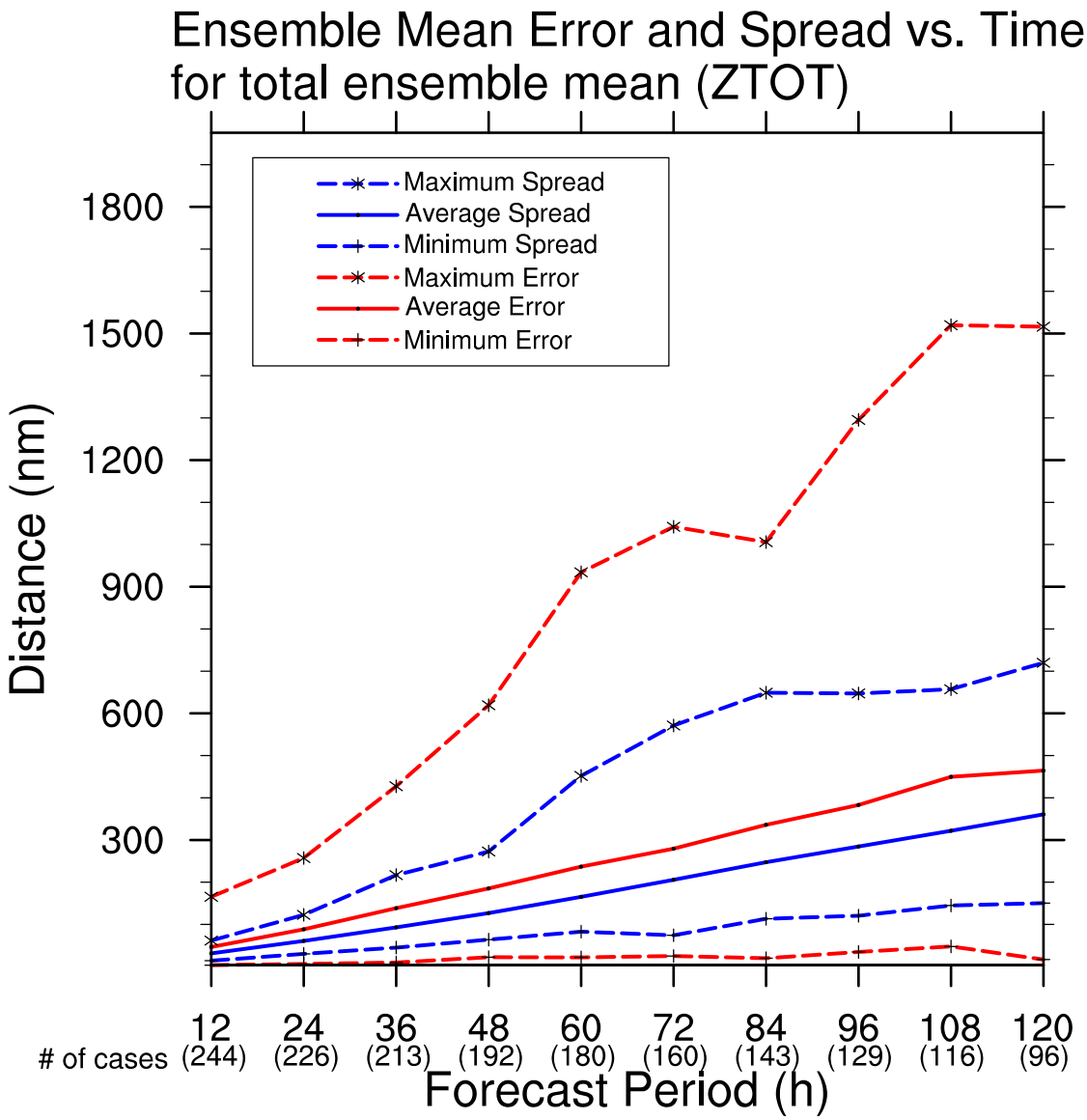


Figure 5.29: The maximum, mean, and minimum spread and error for the total ensemble mean forecasts (ZTOT), by forecast period, for all available cases from the 2001-2003 Atlantic hurricane seasons. The maximum and minimum values of spread and error are the maximum and minimum of all individual forecast cases, respectively, whereas the mean spread and error is the average over all cases.

are most likely to occur for forecasts of slow moving storms in a constrained environmental flow.

Figure 5.31 shows an example of type-2 correctness for the 27 November 2001 Hurricane Olga case. There is very large spread of 571 n mi – the largest 72-h spread of any forecast case during the 2001-2003 seasons – and a corresponding extremely large error of 1042 n mi, also the largest 72-h track error in the sample. In this case, the spread-error relationship indicated a high degree of uncertainty and it was correct – a very large error resulted.

There are also two ways in which the spread-error relationship can be wrong: (1) the case of small spread with large error or (2) the case of large spread and small error. In the first case, the ensemble mean is far from the verifying position (perhaps even outside the ensemble envelope), but the spread is small indicating low uncertainty. Cases with type 1 'badness' are more common than the type 2 'badness' cases. When they occur, it is serious because they represent the totality of the ensemble being wrong about the forecast. This obviously occurs when the ensemble is unable to correctly simulate the uncertainty of the forecast situation. The ensemble-predicted forecast uncertainty is overconfident in this case, predicting that the errors will be small when they really aren't.

Figure 5.32 shows an example of this small spread/large error type of failure of the spread-error interpretation: the 22 August 2001 case of Tropical Storm Dean, starting at 0000 UTC. The actual track of Dean recurved northward, while the ensemble mean forecast was headed towards South Florida, escaping the ensemble envelope by 72 h. The 72-h spread was a below-average 153 n mi while the observed track error of 659 n mi was more than twice the average 72-h error. Despite the poor error/spread ratio of 4.0 (ideally, it should be 1.0), the spatial strike probabilities are still useful. From the shape of the outer ensemble envelope, it is easy to see that a recurving track is possible, as is a continued westward track. This example shows that the spread/error ratio is not always sufficient to determine the uncertainty associated with possible track scenarios – the spatial strike

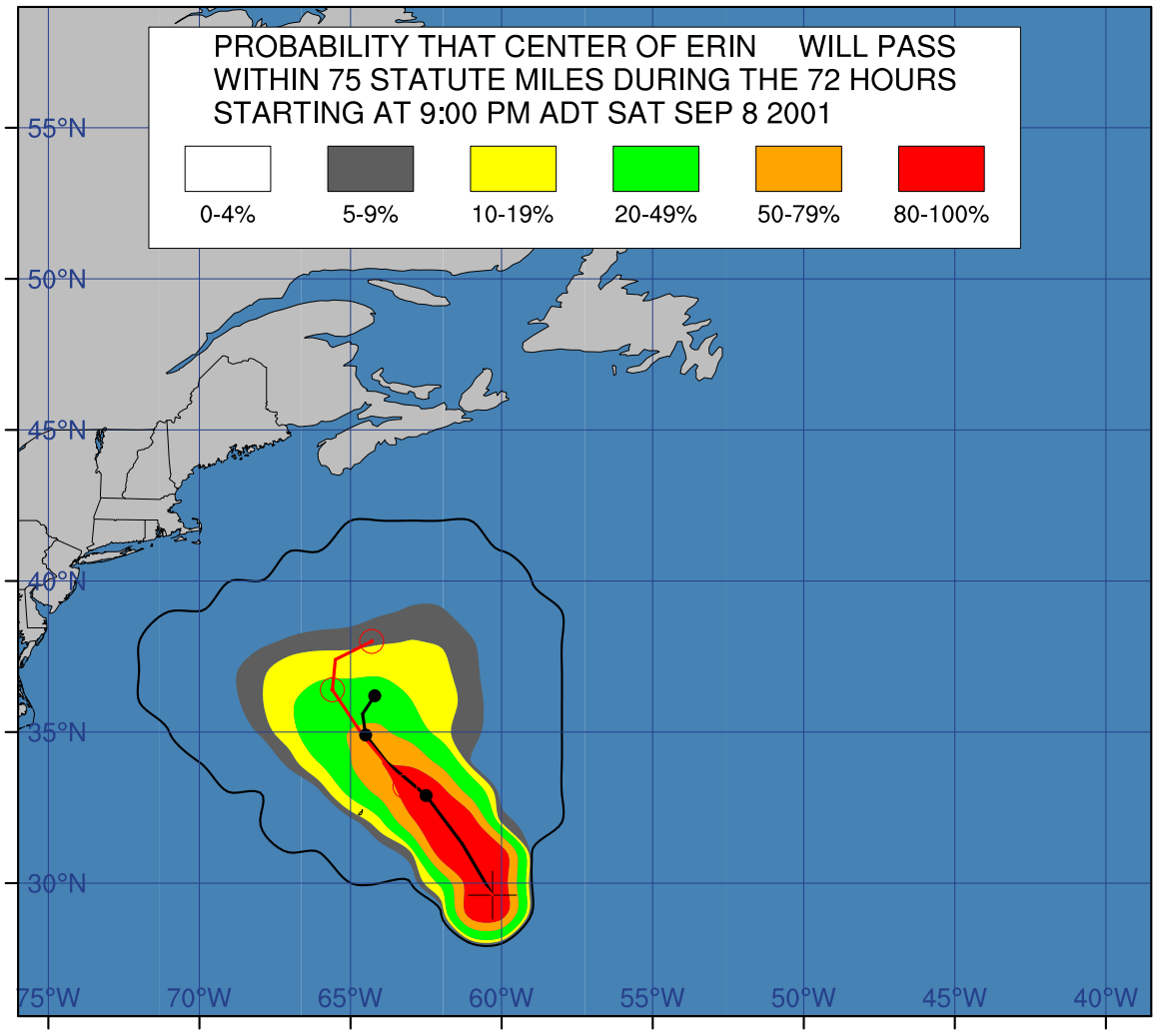


Figure 5.30: An example of small spread and small error. Cumulative strike probabilities through $\tau = 72\,\mathrm{h}$ are shown for Tropical Storm Erin for the forecast period starting at 0000 UTC on 9 September 2001. The tracks of the total ensemble mean forecast and the verifying best track are indicated by the black and red lines, respectively.

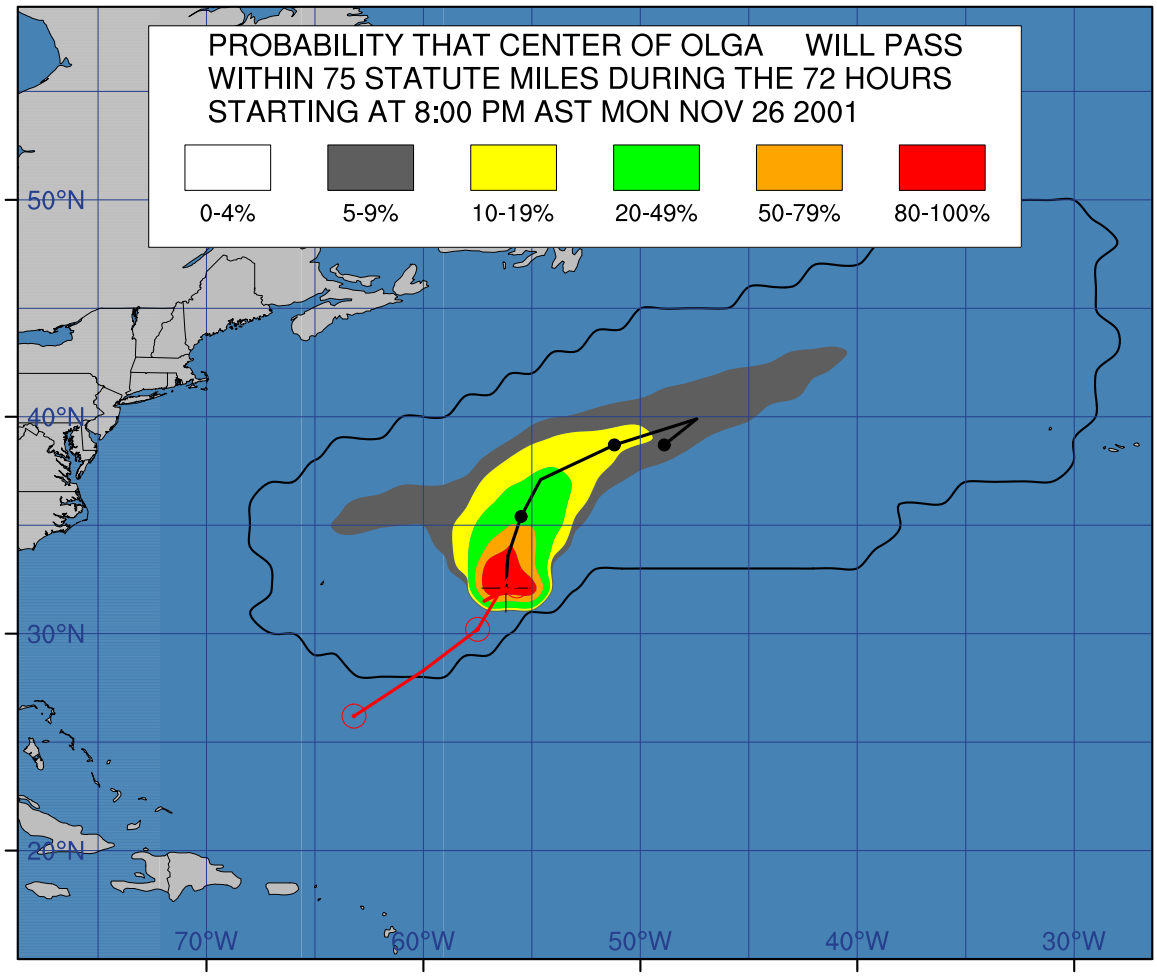


Figure 5.31: An example of large spread and large error. The cumulative strike probabilities through $\tau = 120\,\mathrm{h}$ for Hurricane Olga for the forecast period starting at 0000 UTC on 27 November 2001.

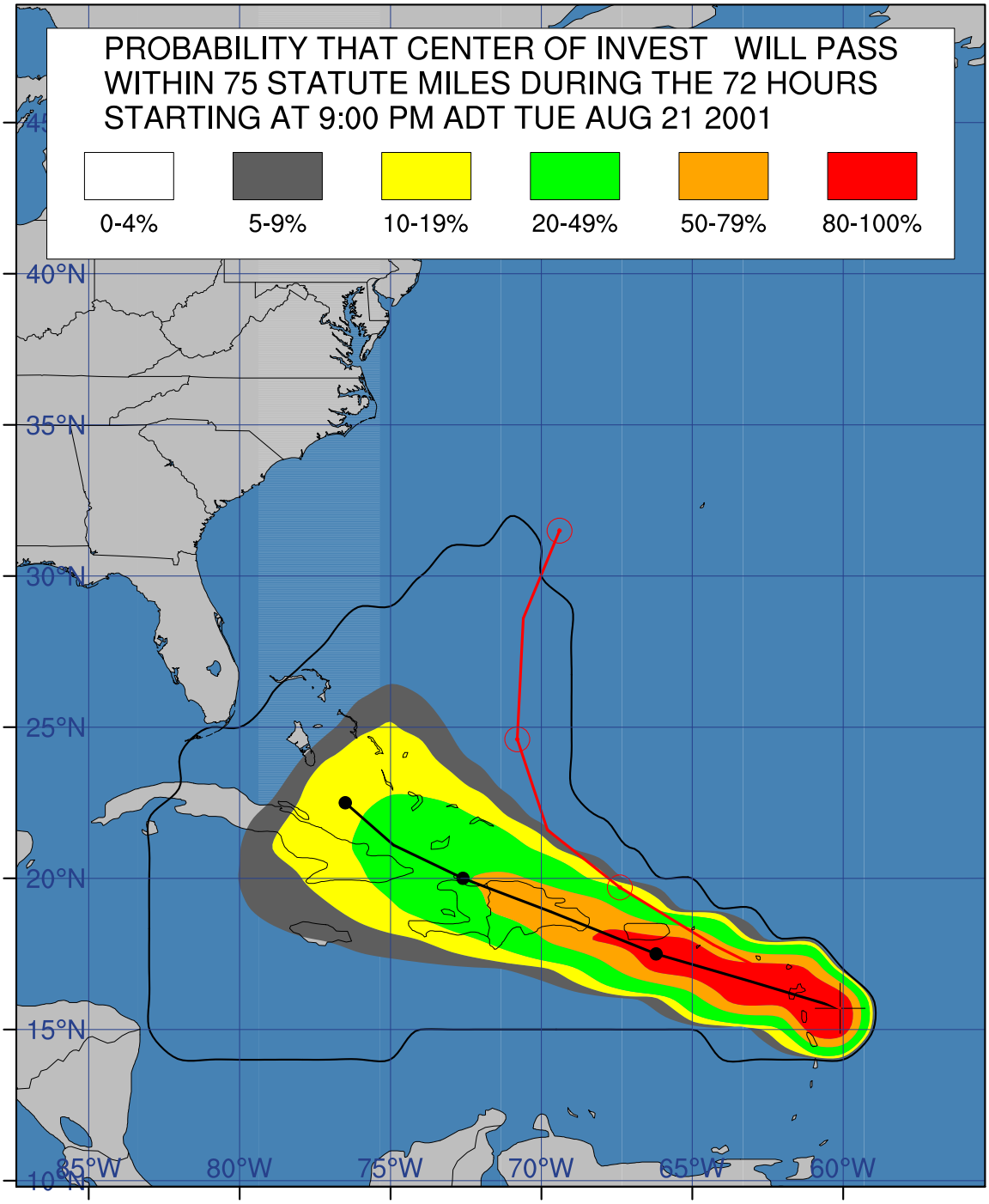


Figure 5.32: An example of small spread with large error. The cumulative strike probabilities through $\tau = 120\,\mathrm{h}$ for Tropical Storm Dean for the forecast period starting at 0000 UTC on 22 August 2001.

probabilities are much more useful in this regard.

The second problem (type 2 'badness') occurs when the ensemble spread is large, but the error is small. This could happen when the ensemble gets 'lucky' and the ensemble mean forecast just happens to be near the best track. The application of the spread-error interpretation implies potential errors larger than the actual observed error – the potential forecast error is overestimated. This type of 'badness' errs on the side of safety which could lead to overwarning. Figure 5.33 shows an example of this large spread/small error type of failure: the 29 November 2001 case of Tropical Storm Olga, starting at 0000 UTC. The 72-h spread for this case was an above average 280 n mi, while the track error was only 136 n mi. Although the ensemble-indicated uncertainty was large, the ensemble forecast was still a good one (not necessarily just by chance, as can be seen in other cases like the 28 September 2003 case for Tropical Storm Kate – in that case which is not shown here, the ensemble mean track just happened to be crossing the best track at 72 h, but by 120 h, was far from the best track).

An exhaustive review of all 294 cases was not conducted, but the error/spread ratio was calculated at 72 and 120 h. The results generally fall between the extreme examples presented here. The error/spread ratio was calculated for each case. At 72-h, there were a total of 154 cases with a verifying best track and a forecast. Out of the 154 cases, 50 cases had error/spread ratios of greater than 2 (large error/small spread, or type 1 'badness'), while 13 had error/spread ratios of less than 0.5 (large spread/small error, or type 2 'badness'). This means that there were egregious failures of the spread-error relationship in 63 out of the 154 cases (41% of the cases). But this also means that the spread-error relationship was at least mildly successful in the remaining cases (consisting of both type 1 and 2 'goodness'). To put all of these statistics into perspective, it is important to realize that the average error at 72 h was 309 n mi, while the average spread was 206 n mi, so the error/spread ratio for all cases was 1.5 – on average, the ensemble-estimated forecast of uncertainty is overestimated – the ensemble is too confident. In other words, the actual forecast is less predictable than the size of the ensemble spread size would indicate. This is

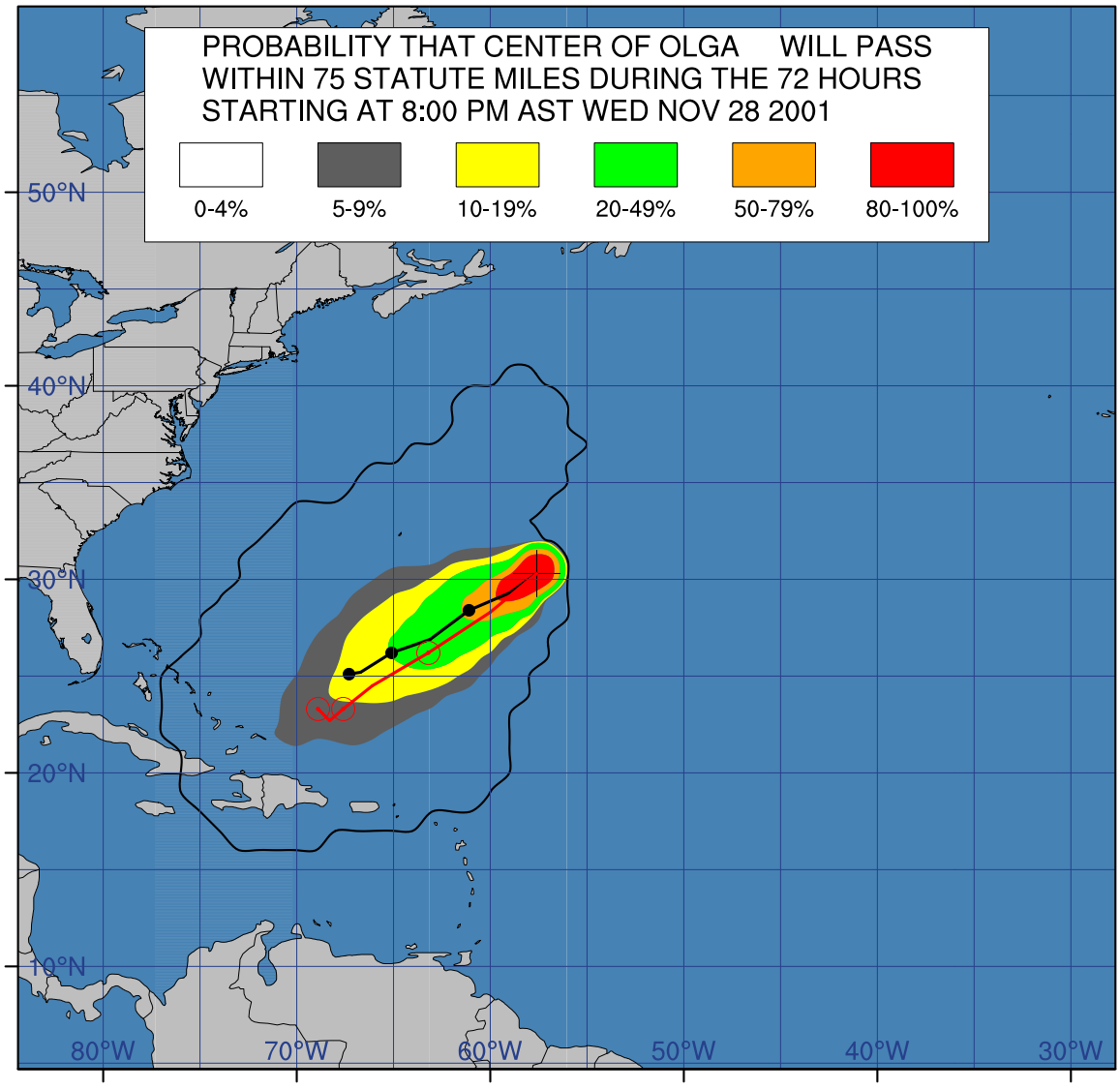


Figure 5.33: An example of large spread with small error. The cumulative strike probabilities through $\tau = 120\,\mathrm{h}$ for Tropical Storm Olga for the forecast period starting at 0000 UTC on 29 November 2001.

a general result of most TC ensemble track studies.

If there is a strong correlation between spread and error, despite a bias, then it should be possible to calibrate the ensemble-predicted errors to match up with the ensemble spread. The kilo-ensemble spreads were regressed onto the observed mean errors to measure the strength of the the spread-error relationship. Figure 5.34 shows scatter plots of the spread vs. the error at selected forecast time periods. The kilo-ensemble was found to have a rather weak spread-error relationship that peaks at 60 h. The relationship was practically nil at the earliest and latest time periods. This suggests that forecasts of ensemble skill are only applicable at the intermediate time periods. The strength of the relationship and % variance explained are nicely summarized by Fig. 5.35. Similar plots for the spread vs. error relationship were also generated for the subensemble means, but the results are generally similar. The less skillful subensembles (SLY7) and (SVM1) had little or no relationship.

The spread-error relationship was conflicting in previous studies. In an early study study of ensemble track forecasting with the GFDL model, Aberson et al. 1998) found no clear correlation between spread and error, but noted that the cases which had large forecast error also tended to have large ensemble spreads – also, it seemed that there was a lower limit to the spread as forecast error increased. Later results from ensemble forecasting with global models suggests a more significant spread-error relationship, with spread generally underestimated (see below). For more on the interpretation of the ensemble spread-error relationship, please refer to: Ch. 7 of Jolliffe and Stephenson (2003); also see Toth and Kalnay (1997) and Buizza (1997).

5.4.6 Comparison with other operational ensembles

To gain perspective on the performance of the kilo-ensemble, it is useful to compare it to the other ensembles systems are used for track forecasting. There are two ensembles which are routinely available and included in the operational A-decks: the GFS ensemble and the hybrid/consensus known as GUNA.

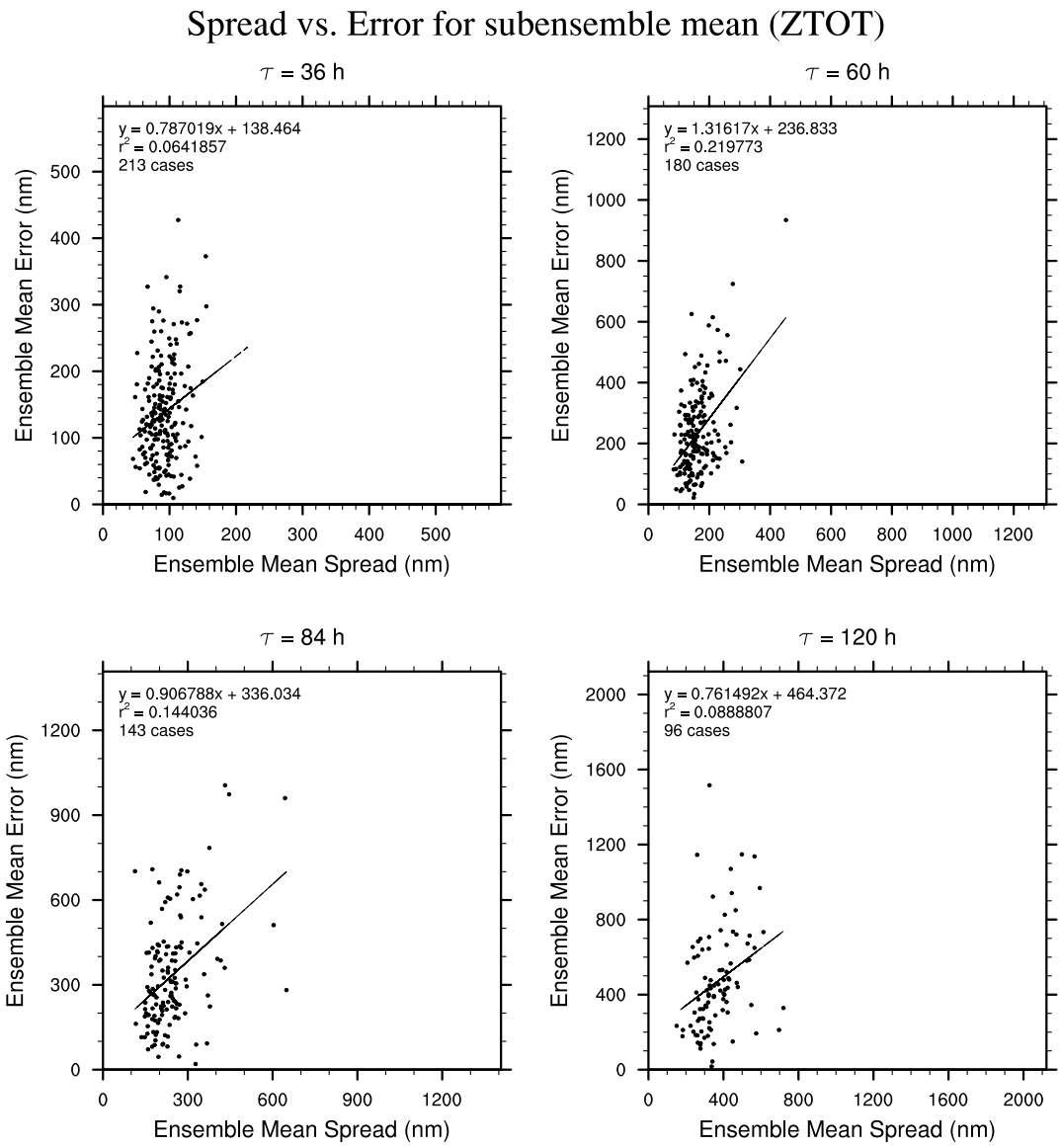


Figure 5.34: Scatter plots of spread vs. error at $\tau = 36$, 60, 84, and 120 h for the 293 forecast cases from 2001-2003. The regression equation is indicated along with the percentage of total variance explained.

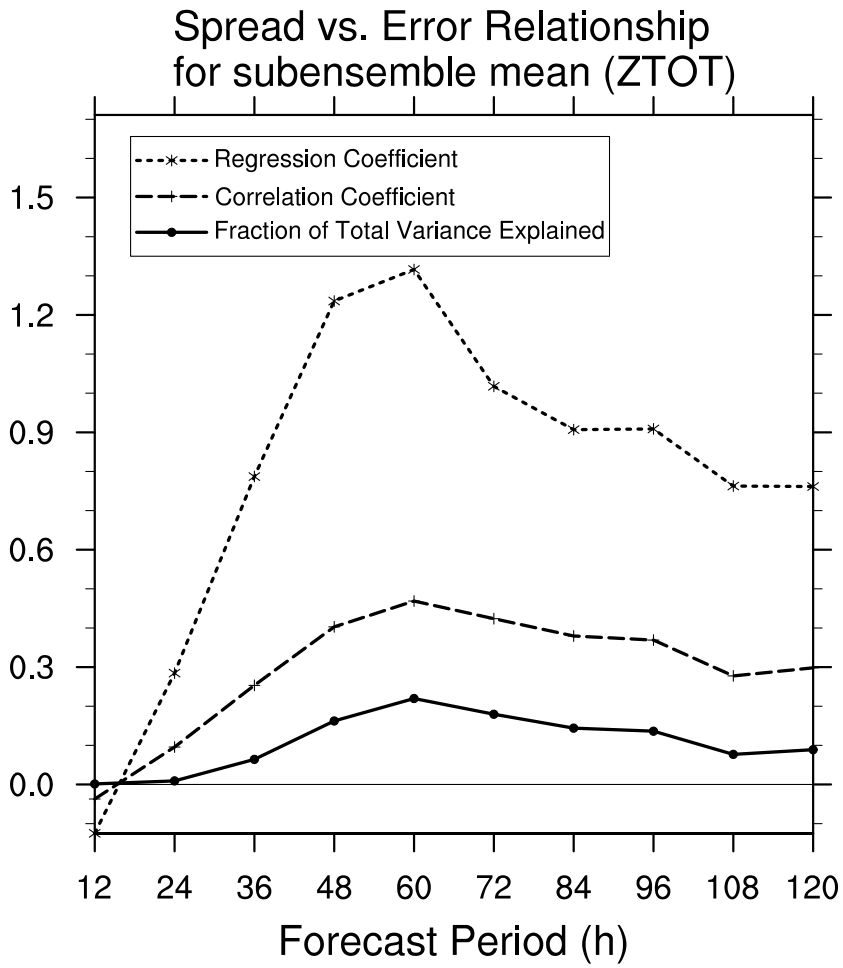


Figure 5.35: The regression and correlation coefficients for the regression of spread onto error, at each forecast time. The percentage of total variance explained by the relationship is also given.

GFS Ensemble

A description of track forecasts of the GFS ensemble members is given in Marchok et al. (2002). On average, the GFS ensemble mean (AEMN) was found to outperform the individual member forecasts. AEMN did not outperform the high resolution operational AVNO until day 4, however. This suggests that much of the utility of the GFS ensembles may occur at the later forecast periods. It would be very interesting to analyze the spread-error relationship of the GFS ensemble, but there are only a limited number of available cases that have forecasts for all 11 members. This may be due in part to the fact that until 2003, the GFS ensemble resolution at intermediate time periods was not high enough to permit reliable vortex tracking (in 2003, the resolution was increased from T62 to T126). Also, the reliability of the ensemble members is not that high – often, at least one of the members will forecast an early dissipation of the vortex. Thus, homogeneous comparisons are difficult: there were only 26 cases with forecasts for all 11 members at the 72-h forecast period. If the requirement that all members be present were relaxed to a cutoff criterion similar to the kilo-ensemble (e.g., 50%), there would be quite a few more cases. This has not be done yet, and is left for future work.

GUNA Ensemble

A hybrid ensemble consisting of the GFDL, UKMO, NGPS, and the AVN is found to outperform the other guidance much of the time. The spread of this small ensemble has proven useful for estimating forecast uncertainty. See Goerss (2000), Mundell and Rupp (2002) for further details.

ECMWF Ensemble

Heming et al. (2004) describes TC track forecasting with the ECMWF Ensemble Prediction System. Neither the ensemble mean or control was able to outperform the deterministic forecast. In general, the EPS seems overconfident at the higher probabilities

(possibly due to positional errors at the initial time), but approaches perfect reliability at forecast probabilities less than 20%.

5.5 Summary

This chapter discussed the post-processing and verification of the kilo-ensemble forecasts for the 2001-2003 Atlantic hurricane seasons. A number of products are generated during the post-processing, including the total ensemble mean forecast, various subensembles, and ensemble-based strike probabilities. Methods of determining the 'goodness' of the ensemble forecasts were discussed, and several measures- and distributions-oriented verification measures are calculated. The characteristics and performance of the ensemble forecasts were examined and compared to several other suites of guidance models and ensembles. In general, the ensemble mean forecast did not perform very well, failing to beat the control forecast. The spread-error relationship was also found to be quite weak, limiting the ability to make confident forecasts of forecast skill. Qualitatively, the ensemble-based strike probabilities seem to offer more information than the first or second moment statistics of the ensemble forecasts, but further work is needed to show this quantitatively.

Chapter 6

CASE STUDIES

6.1 Introduction

This chapter gives a few brief case studies and seeks to offer insight with regard to various aspects of the kilo-ensemble performance, as well as how a forecaster might use and interpret the kilo-ensemble output.[1]

6.2 Hurricane Iris, 2001

Hurricane Iris was a small, but intense storm embedded in deep easterly flow. These type of forecast situations often have low uncertainty/high predictability, with a few caveats. The storm motion is often tightly constrained in the cross-track direction, but significant errors can occur in the along-track direction, which translate into timing errors. The cumulative strike probabilities for the 6 October 2001 case of Hurricane Iris, as seen in Fig. 6.1, show this quite well. There is very little spread (only 140 n mi at 72 h). The resulting error turned out to be very small as well (106 n mi at 72 h). The next day, some of the ensemble members 'got off the track' and erroneously forecast the storm to move northward, resulting in a bifurcation in the spatial strike probabilities, seen in Fig. 6.2. As is usual in a

[1] A web archive of the plots for each of the cases in this study (the 2001-2003 Atlantic seasons) is maintained at:
http://euler.atmos.colostate.edu/∼vigh/ensembles/output_plots.htm.

The archive includes the swarm animations, plots of cumulative strike probability at 72 and 120 h, and the NHC experimental strike probability plots, when available. The archive also includes plots of the forecast tracks of the operational guidance models used by NHC, including the barotropic models, the full-physics and global models, and the track forecasts of the GFS ensemble members.

bifurcation, the ensemble mean track goes right in the middle of the two possibilities (this is typical of the behavior of a hedged forecast), although the map of spatial strike probabilities suggest that this is actually a less likely track scenario than either of the higher probability lobes on each side of the ensemble mean forecast. For comparison, the global model guidance is shown for the 7 October 2001 case in Fig. 6.3. None of the global models support the northward track scenario, so the northward scenarios are likely due to the barotropic dynamics of the MUDBAR model. The best track went further south than the southernmost lobe.

6.3 Hurricane Michelle, 2001

Hurricane Michelle presented one of the most difficult forecast situations of the storms studied. Indeed, in several of the forecast cases, the storm was completely outside of the ensemble envelope for nearly the entire forecast period! As a cat. 4 hurricane, Michelle was a substantial threat to the Miami metropolitan area. General operational model guidance indicated great uncertainty, as did the kilo-ensemble. Yet, based on Aviation model (now the GFS model) forecasts of recurvature, the forecasters at TPC/NHC were able to take Miami out of the imminent threat area two days before potential landfall, even though the storm was still forecast to pass within a couple hundred miles. Some of the 'good' model guidance (i.e., the GFDL and the UK Met Office model) were still showing a close passage to Miami at the 2 and 3 November 2003 forecasts (not shown).

Figures 6.4, 6.5, and 6.6 show maps of the kilo-ensemble cumulative strike probabilities for 2-4 November 2003. The ZTOT forecast for 4 November 2001 passes right over the Miami metropolitan area. ZTOT and some of the other guidance models were still indicating that it would be at least 3 days before Michelle would reach Miami. Yet, the kilo-ensemble outer envelope shows a strong bifurcation – many westward moving cases expanded the envelope westward, but the eastern cases expanded the envelope much further to the east. This suggested that if the easterly track scenario was correct, the storm would

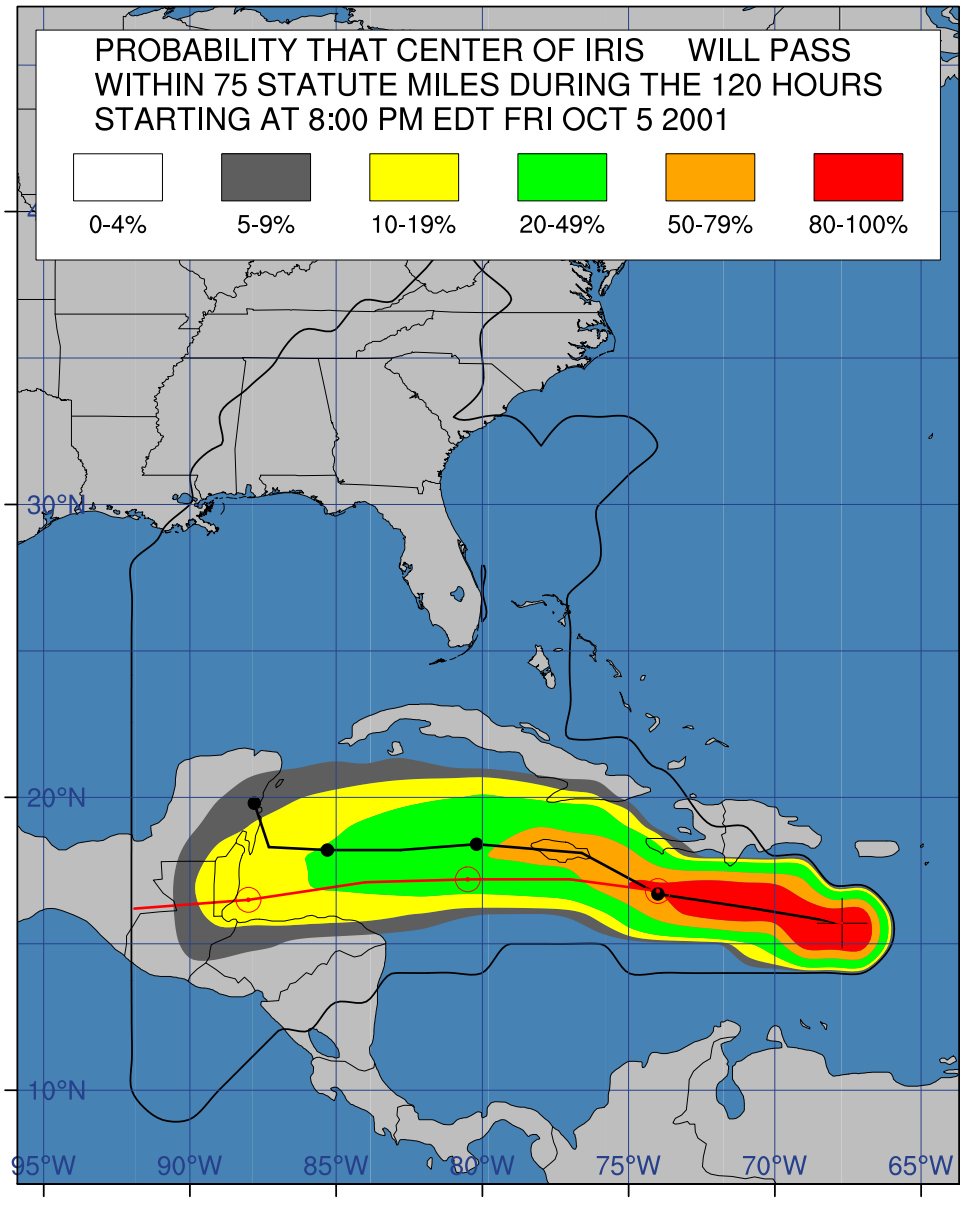


Figure 6.1: The cumulative strike probabilities through $\tau = 120$ h for Hurricane Iris for the forecast period starting at 0000 UTC on 6 October 2001.

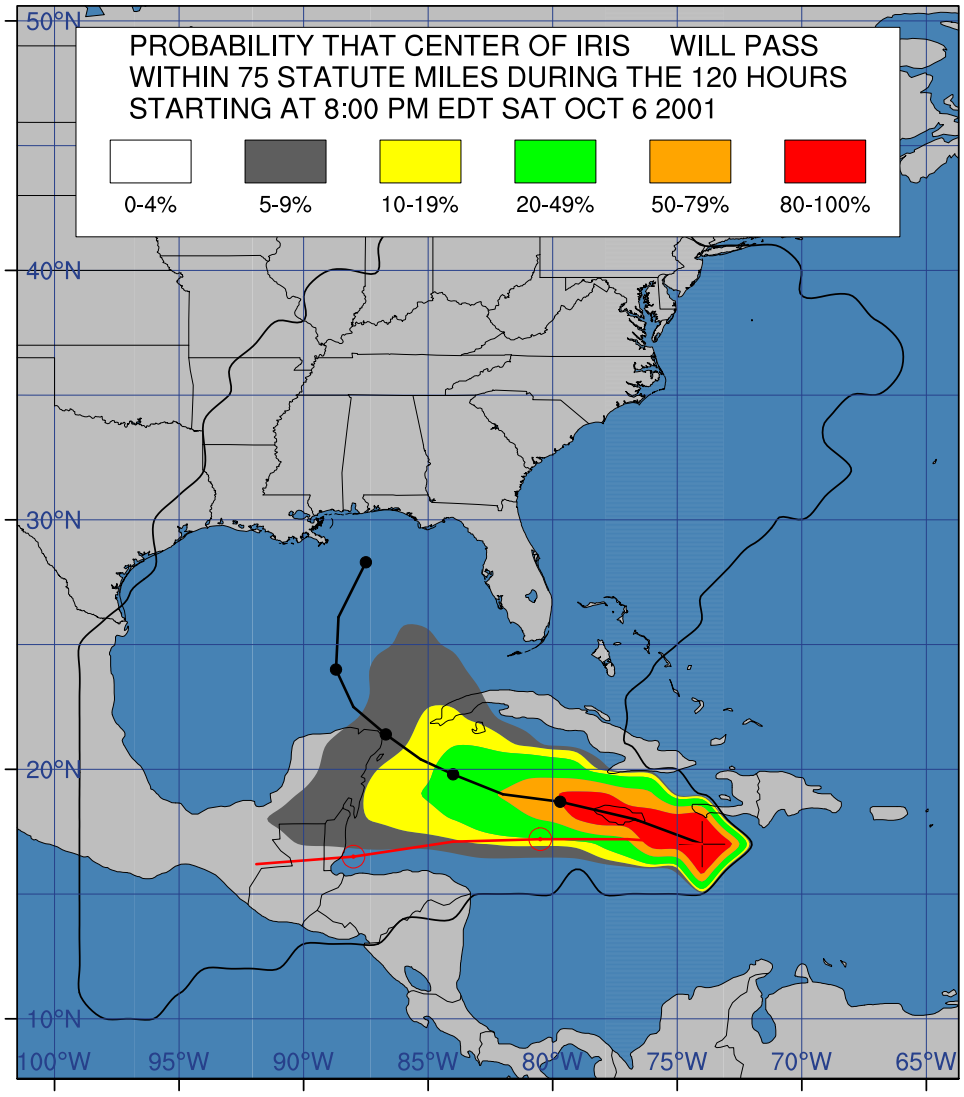


Figure 6.2: The cumulative strike probabilities through $\tau = 120$ h for Hurricane Iris for the forecast period starting at 0000 UTC on 7 October 2001.

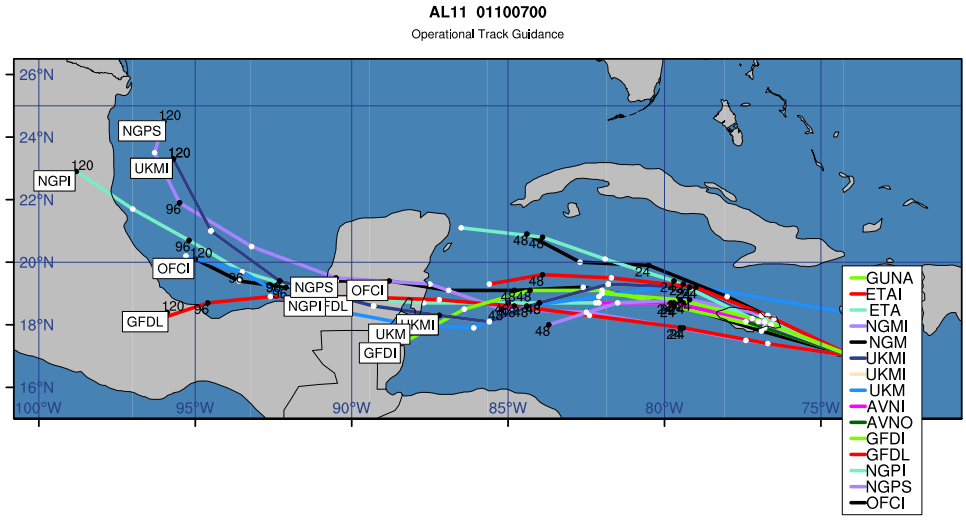


Figure 6.3: The global and regional model guidance for Hurricane Iris for the forecast period starting at 0000 UTC on 7 October 2001.

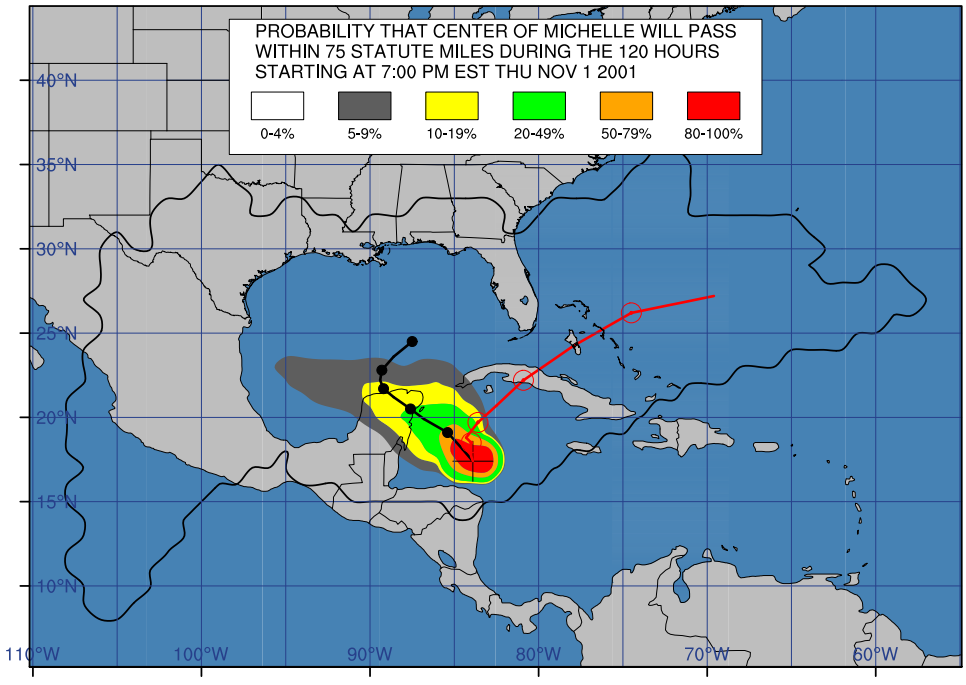


Figure 6.4: The cumulative strike probabilities through $\tau = 120$ h for Hurricane Michelle for the forecast period starting at 0000 UTC on 2 November 2001.

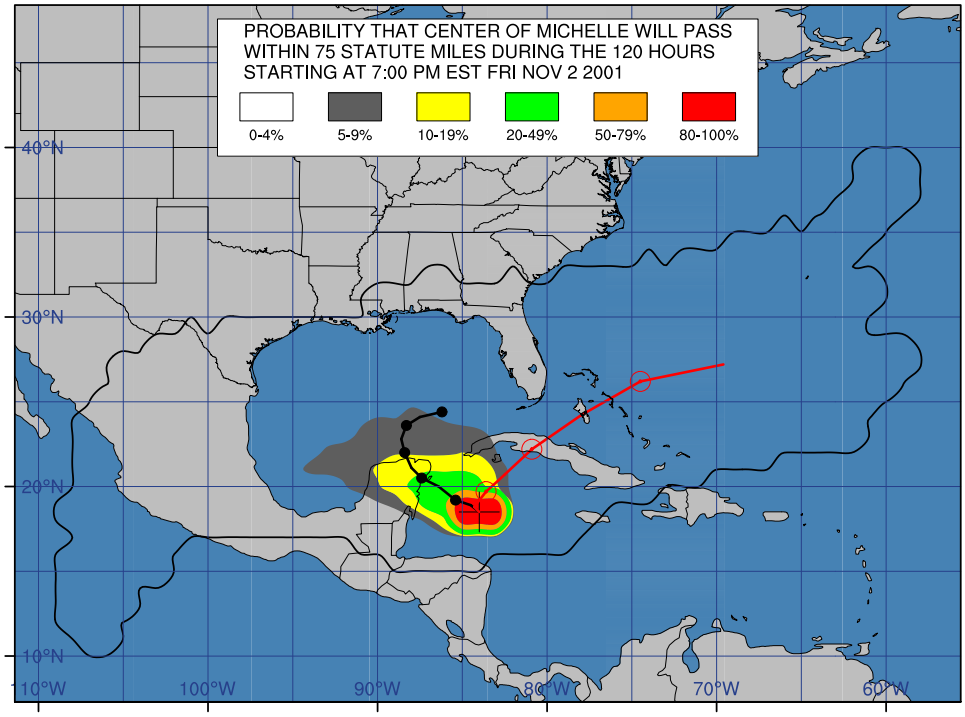


Figure 6.5: The cumulative strike probabilities through $\tau = 120$ h for Hurricane Michelle for the forecast period starting at 0000 UTC on 3 November 2001.

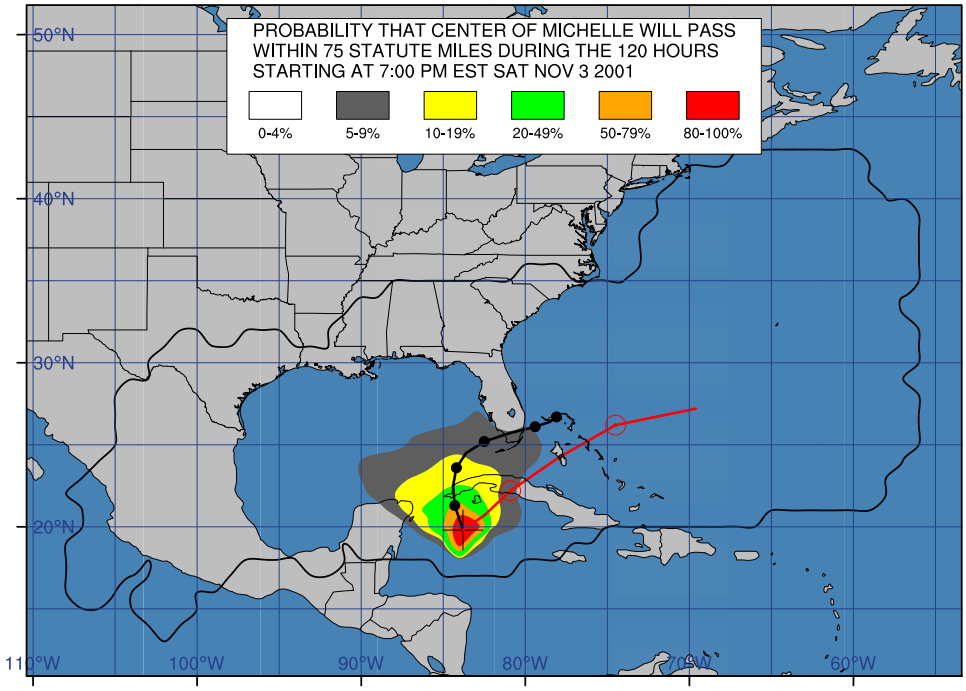


Figure 6.6: The cumulative strike probabilities through $\tau = 120$ h for Hurricane Michelle for the forecast period starting at 0000 UTC on 4 November 2001.

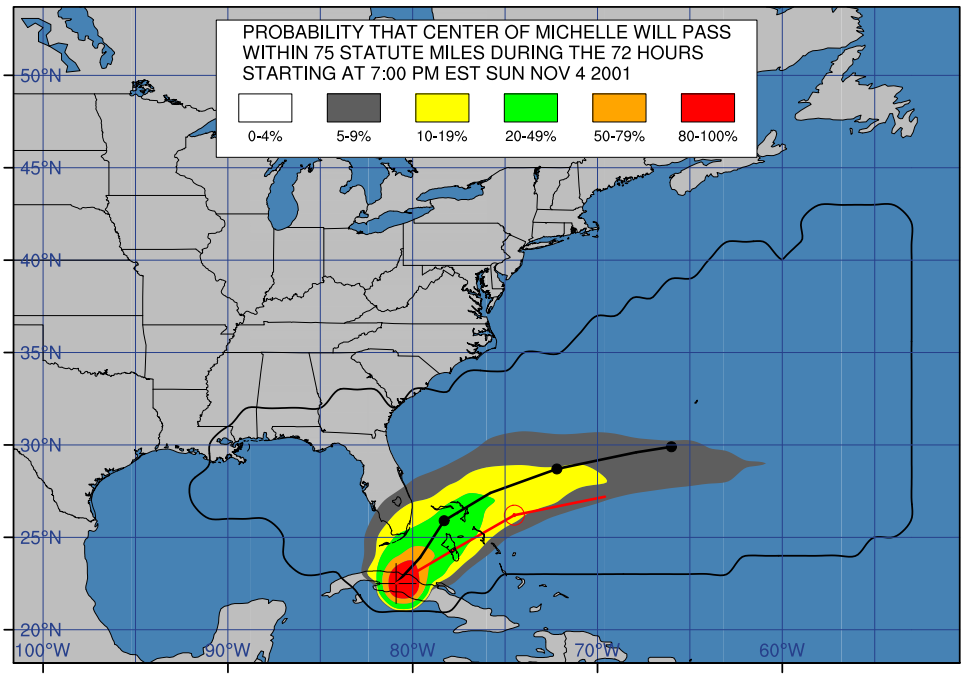


Figure 6.7: The cumulative strike probabilities through $\tau = 72$ h for Hurricane Michelle for the forecast period starting at 0000 UTC on 5 November 2001.

be moving much quicker. At the time of this forecast, the storm was moving north at 2 kt so there was hardly a decisive trend. Shortly thereafter, the storm made the turn to the northeast and began to accelerate. Based on this developing trend, a forecaster could probably have discounted the western scenario – not including those cases would in all likelihood have given an ensemble mean track that would miss Miami. The global model guidance from this crucial forecast time is also interesting. The GFDL track passed directly over Miami, but the AVN and NGPS models insist on a track that is faster and further east.

The 5 November 2001 forecast case was only 12-18 hrs before the point of closest approach to Miami. By this time the threat to Miami could be largely ruled out, as seen in Fig.6.7. All track guidance for this forecast time, Fig.6.8, was now showing a quick recurvature to the east

There are many questions to be answered on this case (and the kilo-ensemble results can't answer some of them) – for instance why did the AVN get Michelle right? Was there

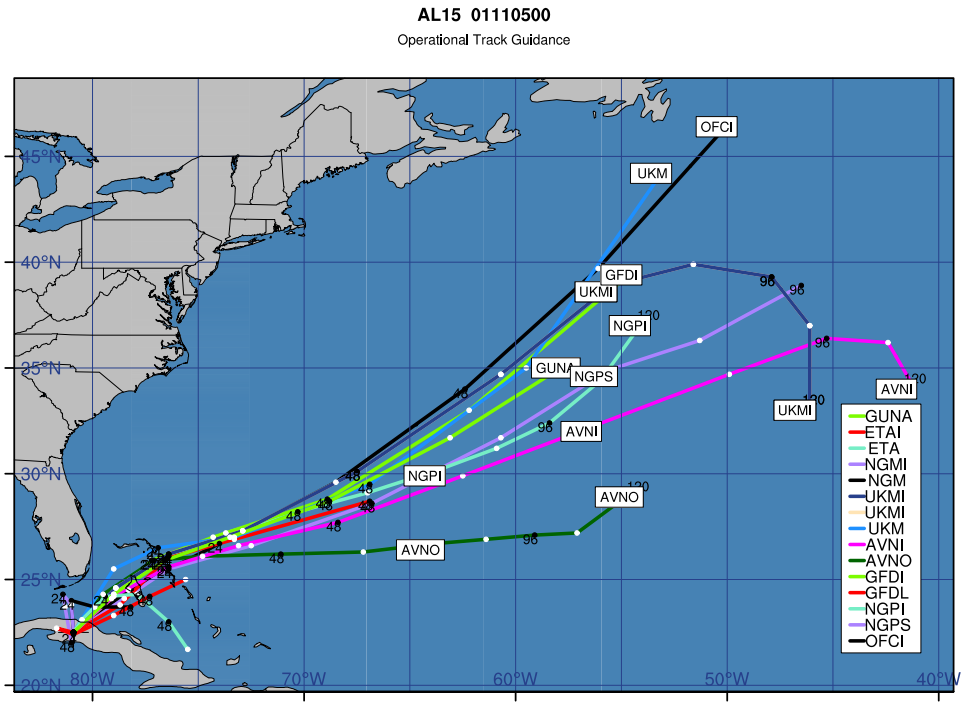


Figure 6.8: The global and regional guidance for Hurricane Michelle for the forecast period starting at 0000 UTC on 5 November 2001.

really a possibility of the storm hitting Miami, as some of the model guidance showed? Why was the model spread so large, and the forecasts so wrong, both for the kilo-ensemble and the global models? Was there a systematic error in the initialization of those models? Did the NHC forecasters get lucky on this one?

What can be concluded from this case study? Michelle represented a case with extreme forecast uncertainty in the early periods, but the implied uncertainty was still useful if portions of the ensemble that were obviously wrong could be discounted – the same goes for the operational guidance. In this case, the use of the AVNO model, which had been forecasting nearly the same track for the life of the storm, was correct – the GFDL also gave a good early forecast (2 November 2001) that agreed with the AVNO scenario, but the GFDL later flip-flopped to the western scenario, which was more in line with the wayward UKM forecasts. This probably caused great consternation to the forecasters in Miami.

6.4 Hurricane Olga, 2001

Tropical Storm (later Hurricane) Olga had one of the most interesting and strange tracks of all the storms studied. Several cases are shown which dramatically illustrate the strengths and weaknesses of the ensemble mean forecasts, as well as the utility of the extra information contained in the spatial strike probability maps.

Figure 6.9 shows the 25 November 2001 forecast case. For comparison, the global and regional track forecasts are shown for the next day's forecast in Fig. 6.10. This case illustrates one of the most extreme examples of forecast bifurcation. The near-field probabilities exhibit not a bimodal track forecast, but a trimodal forecast! The outer envelope covers most of the Atlantic Ocean! Interestingly, the ZTOT forecast is not too bad, but this is partly because the fast recurving members are lost and the mean track becomes a reflection of the southern group, tempered by the norther stragglers. This is one case, where the implicit hedging of the total ensemble mean is beneficial. This case lends support to the

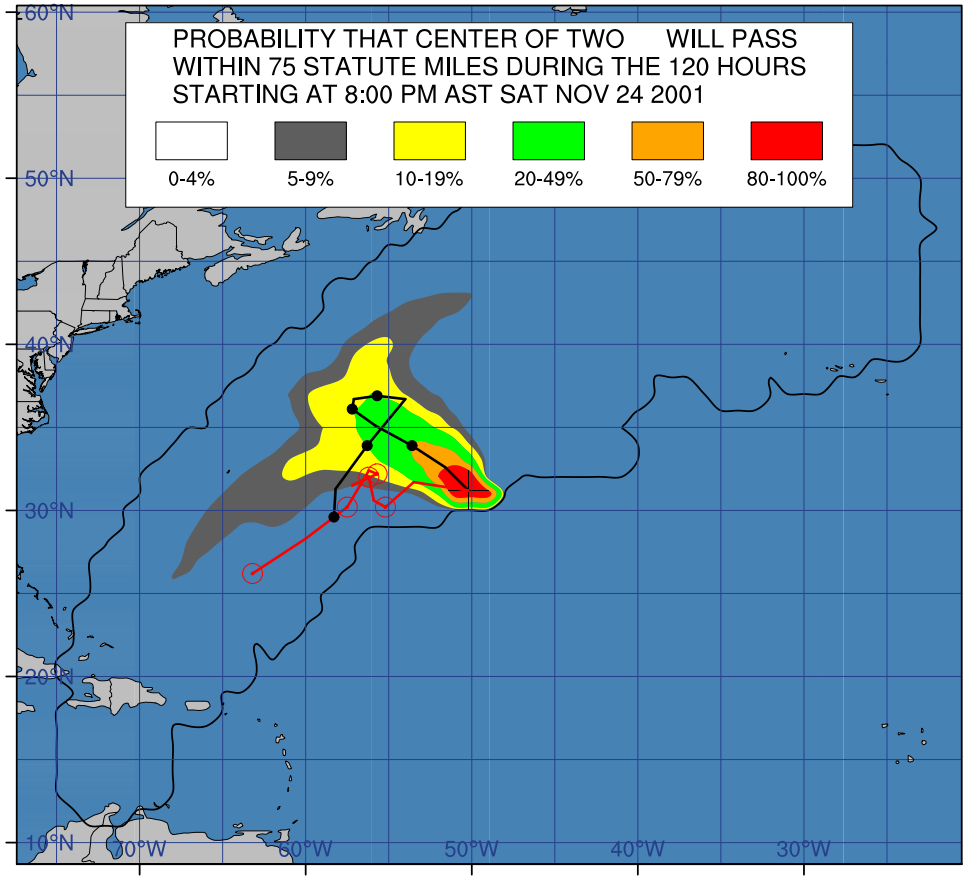


Figure 6.9: The cumulative strike probabilities through $\tau = 120$ h for Tropical Storm Olga for the forecast period starting at 0000 UTC on 25 November 2001.

idea that track forecasts based on the clusters of highest strike probability would be more representative of the possible track scenarios.

6.5 Hurricane Isidore, 2002

Isidore exhibited a very interesting track. Figures 6.11, 6.12, and 6.13 show the cumulative strike probability maps through 120 h. The strong jog to the south, with subsequent landfall on the Yucatan Peninsula (as an intense cat. 4 storm) was not well forecast by barotropic models, but this possibility was strongly advertised by the GFS suite, as seen in Fig. 6.14. The kilo-ensemble and other guidance hinted at this possibility on 18 September 2002, but the 19 September 2002 forecasts took the storm further northwest. By 20 Septem-

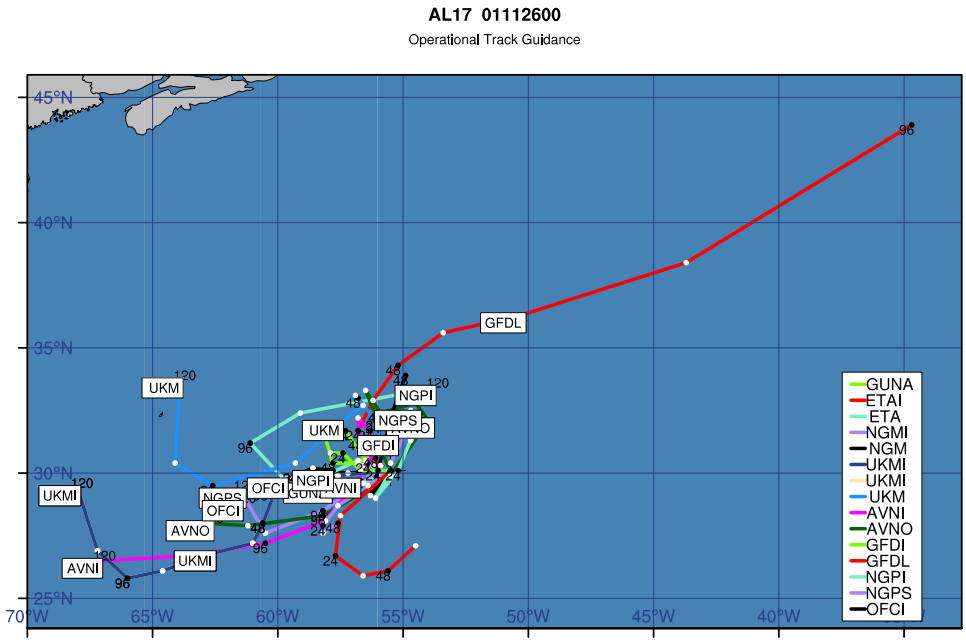


Figure 6.10: The global and regional model guidance for Tropical Storm Olga for the forecast period starting at 0000 UTC on 26 November 2001.

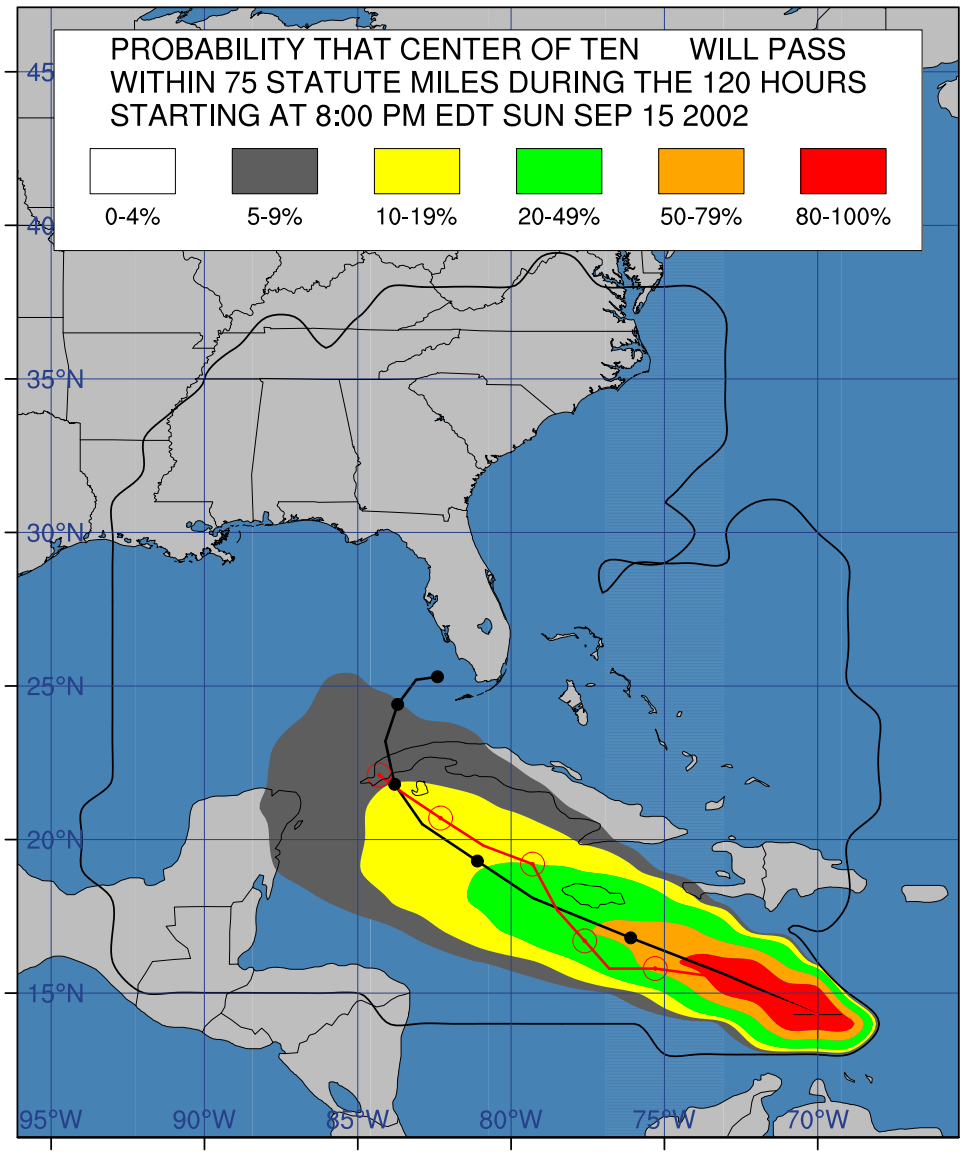


Figure 6.11: The cumulative strike probabilities through $\tau = 120$ h for the Tropical Depression that later became Hurricane Isidore for the forecast period starting at 0000 UTC on 16 September 2002.

ber 2002, some of the skillful global guidance was forecasting the southward sojourn to the Yucatan, as shown in Fig. 6.15, but the kilo-ensemble and other barotropic guidance was still insisting on a northwest track into the central Gulf of Mexico (not shown). The kilo-ensemble's failure to capture this is likely due to the over simplification of the barotropic dynamics.

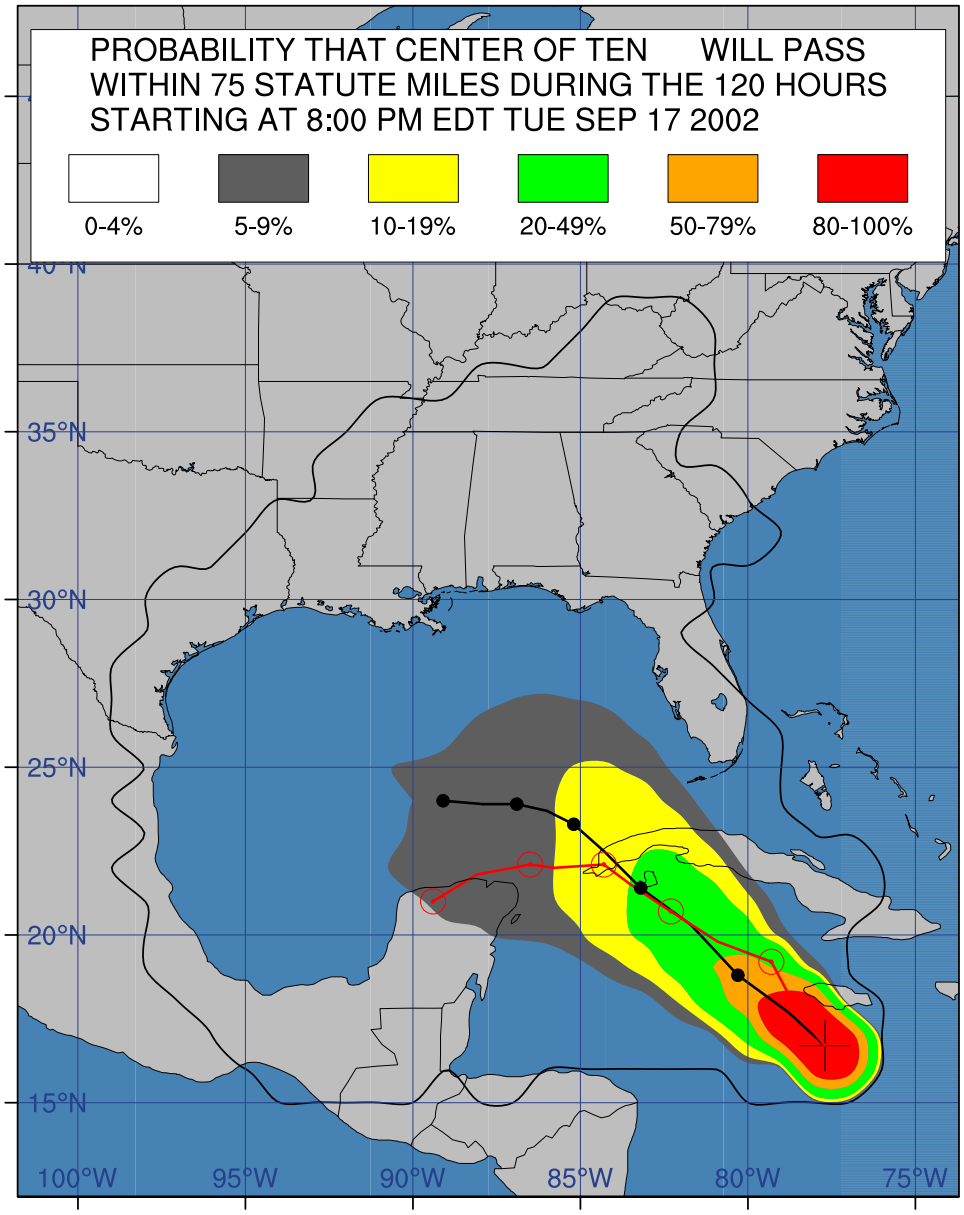


Figure 6.12: The cumulative strike probabilities through $\tau = 120$ h for what later became Hurricane Isidore for the forecast period starting at 0000 UTC on 18 September 2002.

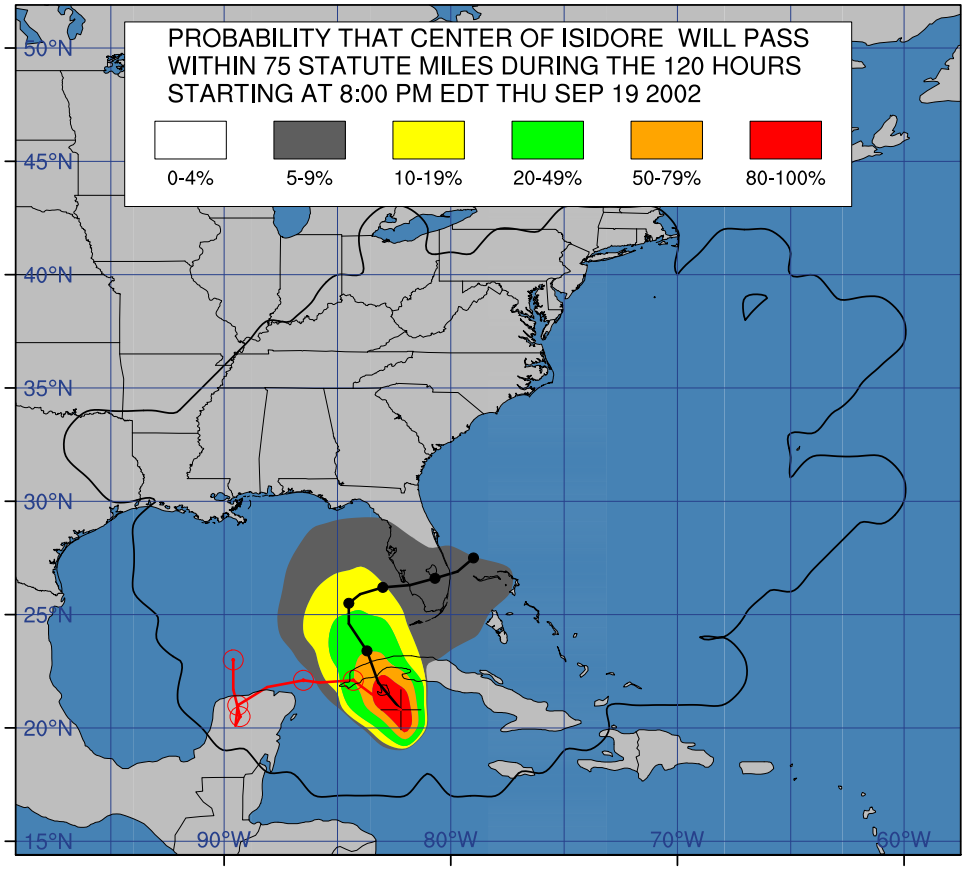


Figure 6.13: The cumulative strike probabilities through $\tau = 120$ h for Tropical Storm Isidore (nearing hurricane strength) for the forecast period starting at 0000 UTC on 20 September 2002.

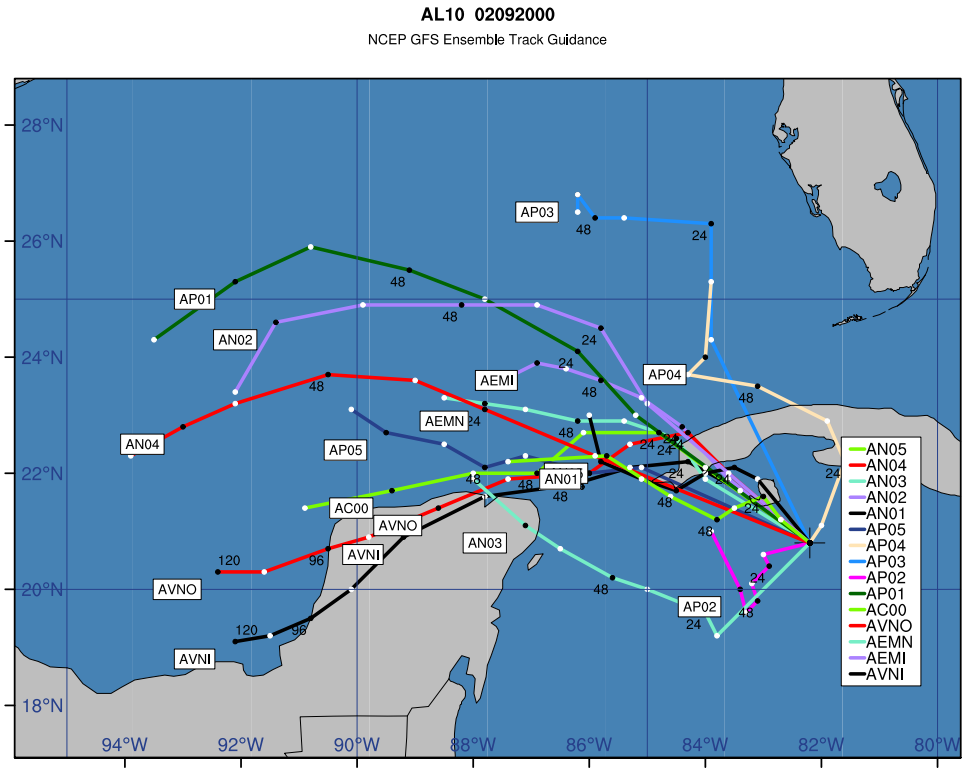


Figure 6.14: The GFS ensemble guidance for Tropical Storm Isidore for the forecast period starting at 0000 UTC on 20 September 2002.

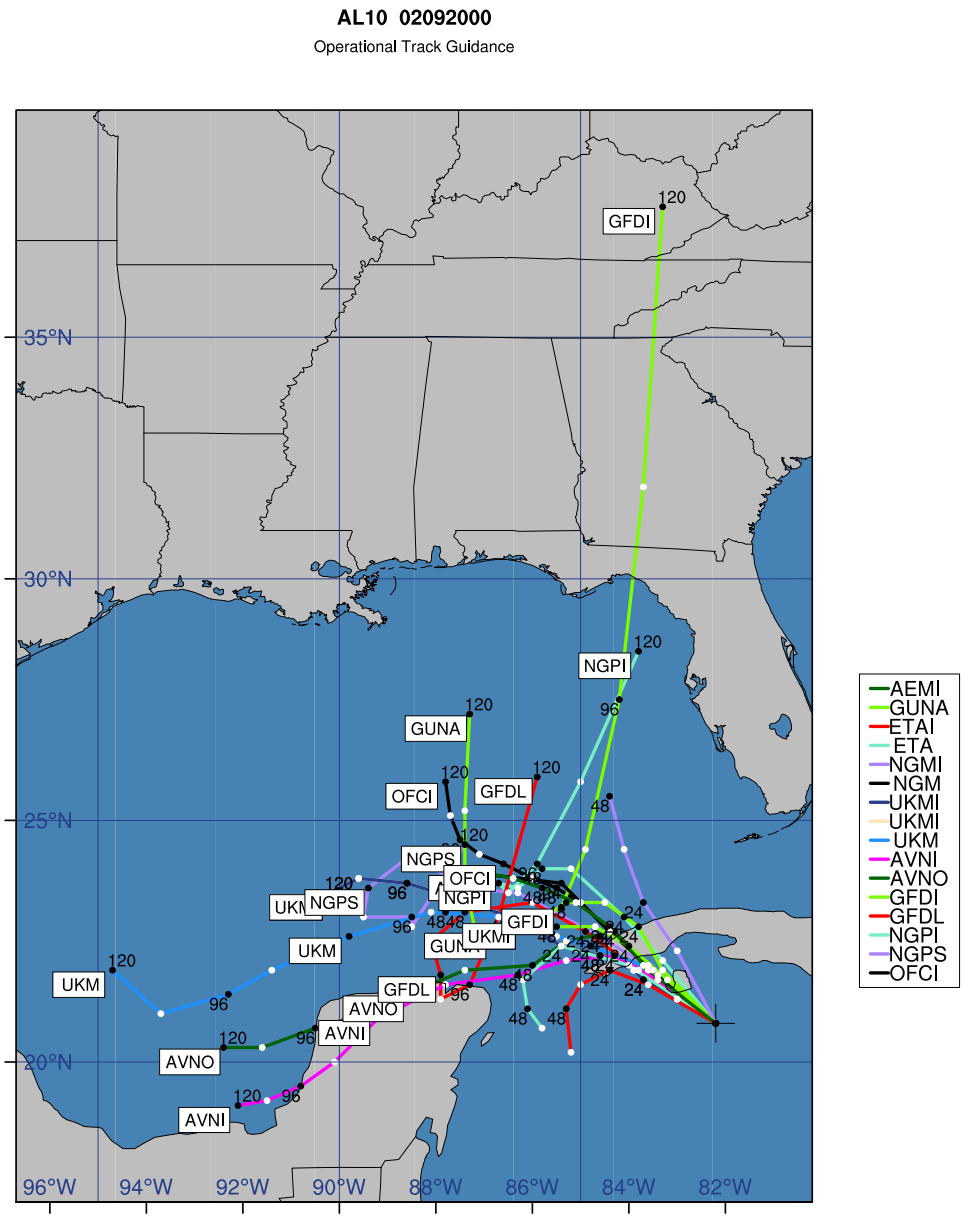


Figure 6.15: The global and regional model guidance for Tropical Storm Isidore for the forecast period starting at 0000 UTC on 20 September 2002.

6.6 Summary

These qualitative case studies show some of the strengths and weaknesses of the kilo-ensemble compared to a single deterministic forecast. In particular, they show the greater amount of information (either correct or incorrect) that is contained in the spatial strike probabilities, compared to just the ensemble mean.

Chapter 7

CONCLUSIONS

This thesis described the development of a kilo-member ensemble using a barotropic model. 26 perturbations were conducted over a total of five different perturbation classes. Various products were generated from the ensemble output – these products were verified using various metrics, and the ensemble performance was examined.

In general, the kilo-ensemble provides a great platform to examine the storm path sensitivity to the initial conditions. The performance of the total ensemble mean was somewhat disappointing, with little or no improvement over the control forecast (A5K0). Furthermore, a single MBAR forecast, initialized with the operationally-estimated intensity and size, was generally able to beat the total ensemble mean forecast most of the time. Also, the results suggest that the applicability of a barotropic forecast model only extends through about 84-h. Various other metrics indicate modest ensemble performance, and the results are in line with expectations, considering the great simplification of using a barotropic model. One positive result of this research is that it lends support to the utility of explicitly-calculated spatial strike probabilities. I feel that the resulting graphical products provide much more information about the dynamical situation than any single deterministic forecast can provide.

7.1 Interpretation of results

7.1.1 Answers to Fundamental Questions

At the beginning of this thesis, the following questions were asked:

(1) Can a well-perturbed ensemble give a better forecast than any single model realization?

(2) How many ensemble members are necessary to get the "right" answer?

(3) Is there a relationship between ensemble spread and forecast error?

(4) Can information about the ensemble spread be used to provide a meaningful forecast of forecast skill?

(5) How accurately does the ensemble envelope of all track possibilities encompass the actual tracks observed?

(6) Can a barotropic model provide a useful framework for ensemble forecasts of TC tracks, or is it necessary to include baroclinic dynamics?

This work did not provide a conclusive answer to the first question, contrary to the predictions of ensemble theory. The reason for this failure is probably due to flaws in the ensemble design (see below). In general, perturbations conducted in a parameter phase space do not appear to have as much merit as those conducted in a dynamically-constrained framework. Previous studies that used barotropic models showed only marginal improvements of the mean forecast over the individual perturbations. The items under future work may shed more light on this question.

For the second question: sometimes an infinite number of ensembles will not give the "right" answer. It seems that the cross multiplication of perturbations adds little if any skill to the total ensemble mean forecast, but it is quite useful way deriving smooth spatial strike probabilities. In that regard, an ensemble size of several hundred to a thousand is probably sufficient to sample the PDF of future storm location.

To answer the third question, a weak relationship was found between ensemble mean error and ensemble spread. The relationship peaks at 60-h, suggesting that the ensemble's ability to predict forecast uncertainty is maximized at intermediate time periods. This

positive result is a confirmation of ensemble theory.

The weak, but positive correlation between spread and error suggests that the kilo-ensemble can provide meaningful forecasts of forecast skill, although with not much 'sharpness'. According to other studies, (Goerss 2000), the spread-error relationship is stronger for the ensemble of operational forecast models (like GUNA). Yet the question still remains – can ensembles provide information on whether an outlier is valid or not. Sometimes the forecasts of an outlier can be discounted as completely unrealistic due to gross errors in the modeling or initialization process. Correct outliers, however, add great benefit to the ensemble. Incorrect outliers, on the other hand, degrade the ensemble skill. More research is needed to determine criteria for discounting or retaining outliers.

The outer ensemble envelope, corresponding to the 0% strike probability, contained the 5-day verifying position about two-thirds of the time – thus, the ensemble envelope was mildly successful at encompassing the actual tracks. The ensemble strike probabilities varied in a reasonable fashion, similar to the current NHC strike probability product. For a complete evaluation of the ensemble strike probability, there is still a need for further tools and baseline for skill, but even without these, it is obvious that there is still plenty of room for improvement.

7.1.2 A Sensitivity Perspective

This work can be viewed as a comprehensive sensitivity experiment regarding the factors of the initial condition as they relate to subsequent storm motion. The results summarize the apparent value (or lack thereof) of the perturbation classes. A given class adds value if it can make a contribution to overall ensemble skill. The results of this thesis suggest that one perturbation in particular does not add much, if any, skill to the ensemble in most cases: the shallow layer-mean wind perturbation (subensemble: SLY7). The weak/small vortex perturbation (subensemble: SVM1) is also less helpful than its peers. If the forecasts of these two members were removed, the total ensemble would be cut

to 990 members, and I would venture to guess that it would be more balanced and accurate for most cases. However, a better solution would be to include only those perturbations which are relevant to the forecast situation. There are many forecast situations when the shallow layer mean wind, or the weak/small vortex is an inappropriate perturbation. But there are a small but significant number where they are appropriate. This line of reasoning suggests that the perturbation classes should be centered on the values that are deemed most appropriate based on the forecast situations. I believe it is possible that an ensemble based on relative perturbations would be more accurate, and possibly beat the control forecast, as theory dictates it should.

Here is a qualitative judgment of the value added by the perturbations of each of the five classes:

(1) SBF – seems to be very valuable

(2) SLY – very important

(3) SVM – important, the forecast track is quite sensitive to the size of the vortex, as well as the vortex profile used (not studied in depth in this work)

(4) SGM – less important

(5) SMV – less important – these perturbations help to add random 'noise' and fill out the spatial probability fields

7.1.3 Possible Reasons for Degraded Ensemble Performance

Although the ensemble performed well in some regards, the failure to improve over the control forecast was disappointing. However, I think some of this disappointment is due to design flaws that can be easily remedied (see below). Some of the reasons for the poor performance include:

- Barotropic dynamics are too simple

- Edge biases (artificial skill degradation through failure to remove certain cases from verification)
- Poor design (should have used relative perturbations, rather than fixed perturbations)
- Possible binary interactions between bogus vortex and GFS-analyzed vortex

In particular, erroneously-located analyzed vortices in the GFS background model fields likely cause significant degradation of the kilo-ensemble forecasts, since the model analyzed vortex can interact spuriously with the bogus vortex that is inserted into MUDBAR at the correct location. Brown et al.(2000) found that false binary interactions occur often in the global forecast models that use synthetic data. In a study of storms during the 1997-1998 forecast season, false binary interactions were found to account for 31%, 28% and 38% of poor forecasts for the NOGAPS, UKMO, and ECMWF models respectively. This problem was also present in the operational AVN/MRF, and is discussed by Liu et al. (2002). Once the vortex relocation scheme of Liu et al. (2000) was implemented, they noted that track errors for the 2000 season were reduced by 30% compared to the 1995-1999 seasons! Some of this dramatic improvement is also the result of substantial model upgrades, as described by Pan et al. (2002).

The perturbed members of the GFS ensemble currently do not experience vortex relocation (Toth 2004, personal communication), although the perturbations are centered on the GFS initial condition which is based on a relocated first guess field. Since the GFS perturbations are derived through a breeding process, I think it is still possible for substantial position differences to be present in the initial condition fields of the GFS perturbed members.

Results from the GFS ensemble suggest that much of the utility may come at the later forecast periods (Days 4, 5, and beyond). Unfortunately, this period also corresponds to a decrease in the validity of using a barotropic model for TC track forecasting. So the

methods of the kilo-ensemble are not able to capitalize on some of the benefits provided by the perturbed GFS model fields, at least in the current barotropic framework.

7.2 Future work

Due to a time constraints, a number of items have been left to future work. Some of these items should be addressed if the kilo-ensemble were ever to be implemented as an operational product. If occasion presents further opportunity, I shall publish the results of this thesis (and further work) as a paper, probably in *Monthly Weather Review*. Here are avenues of future work that I feel would add additional value to this work:

- Increase the MUDBAR model domain to a size appropriate for 5-day forecasts. The best way to do this will probably involve the inclusion of a nudging term which acts in a buffer zone around the edge of the model domain (as in the LBAR model). This term should nudge the regional model wind fields towards the forecast wind fields of the global model. This would ensure that the synoptic information flows into the model in the most accurate manner possible, yet allow the domain to be large enough to always contain the 5-day forecast.

- Determine the extent and effect of binary interactions by looking at the initial fields of the ensemble members runs. Examine ways to reduce this effect, possibly by relocating the storm to the correct location in the GFS ensemble wind fields, or by filtering the storm out entirely throughout the forecast period.

- Calculate ensembles of various sizes by randomly sampling the kilo-ensemble members. This will shed light on the question: "What size should an ensemble be to obtain the 'right' answer?".

- The kilo-ensemble should be redesigned using relative perturbations centered on a MBAR-type control using the information contained in the CARQ lines of the

ATCF files. Relative perturbations are much more appropriate in simulating the actual uncertainties of the forecast problem.

- Conduct a detailed examination of the other ensembles already contained in the ATCF guidance. The performance characteristics of these ensembles (e.g., the spread-error relationship) like the GUNA and GFS ensembles should be compared to those of the kilo-ensemble.

- Derive and verify a simulated cluster analysis, obtained by tracking local maxima in the spatial strike probability field. Such clusters will provide information about the various track scenarios and their likelihood of verifying. This type of information, if accurate, would be extremely useful for forecasters.

- Develop an objective method to identify outlier tracks. Can criteria be developed to decide whether to discount some outliers, while retaining valid outlier scenarios?

- Verify the ensemble-based strike probabilities using the Brier Skill Score and the Relative Operating Characteristics score. This will require the development of an appropriate probabilistic baseline by which to measure skill. It would also be very instructive to compare the kilo-ensemble strike probabilities to strike probabilities generated by other ensembles, such as the GFS or GUNA ensembles, as well as the NHC strike probability forecasts.

- Calibrate the spatial strike probabilities and remove the biases, if possible.

- Using the kilo-ensemble, or the hybrid ensemble of current operational products, develop a MOS-like guidance for the guidance.

7.3 Concluding Comments

In closing, I offer a personal comment on the future of TC ensemble forecasting. Ensemble methods will be vitally important to efficiently harness the ever increasing computer

power, but I feel there are several issues that must be solved before the full potential of ensembles can be exploited. The correct diagnosis of a hurricane's core structure remains one of the most elusive of TC forecasting problems. Much effort should be devoted to developing methods for the real-time determination of this structure. Progress will likely result when the wealth of remote sensing data can be fully incorporated into the simulated initial structure using dynamically-constrained ensemble data assimilation techniques. When reliable structures of the storm's inner core become available, then I think ensemble methods will be able to provide useful information regarding the uncertainty of nearly all aspects of the storm effects, including track, intensity, size, and precipitation.

BIBLIOGRAPHY

Aberson, S. D., 1998: Five-day tropical cyclone track forecasts in the North Atlantic basin. *Wea. Forecasting*, **13**, 1005–1015.

Aberson, S. D., 2001: The ensemble of tropical cyclone track forecasting models in the North Atlantic Basin (1976-2000). *Bull. Amer. Meteor. Soc.*, **82**, 1895–1904.

Aberson, S. D., 2004: Personal communication. re: Has the CLP5 model remained static during the 2001-2003 Atlantic hurricane season? An e-mail exchange.

Aberson, S. D., M. A. Bender, and R. E. Tuleya, 1998: Ensemble forecasting of tropical cyclone tracks. Preprints, *Symposium on Tropical Cyclone Intensity Change*, Phoenix, AZ, Amer. Meteor. Soc., 290–292.

Aberson, S. D. and M. DeMaria, 1994: Verification of a nested barotropic hurricane track forecast model (VICBAR). *Mon. Wea. Rev.*, **122**, 2804–2815.

Aberson, S. D., S. J. Lord, M. DeMaria, and M. S. Tracton, 1995: Short-range ensemble forecasting of hurricane tracks. Preprints, *21st Conf. on Hurricane and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 494–496.

Beven, J. L., S. R. Stewart, M. B. Lawrence, L. A. Avila, J. L. Franklin, and R. J. Pasch, 2003: Atlantic hurricane season of 2001. *Mon. Wea. Rev.*, **131**, 1454–1484.

Bjerknes, V., 1904: Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik. *Meteor. Z.*, **21**, 1–7.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

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

Brown, D. S., M. A. Boothe, Carr, III, L. E., and R. L. Elsberry, 2000: Evaluation of dynamic track predictions for tropical cyclones in the Atlantic during 1997-1998. Preprints, *24th Conf. on Hurricanes and Tropical Meteorology*, Fort Lauderdale, FL, 390–391.

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

Chan, J. C. L. and R. T. Williams, 1987: Analytical and numerical studies of the beta-effect in tropical cyclone motion. Part I: Zero mean flow. *J. Atmos. Sci.*, **44**, 1257–1265.

Chen, Q. and Y. Kuo, 1992: A harmonic-sine series expansion and its application to partitioning and reconstruction problems in a limited area. *Mon. Wea. Rev.*, **120**, 91–112.

Cheung, K. K. W. and J. C. L. Chan, 1999a: Ensemble forecasting of tropical cyclone motion using a barotropic model. Part I: Perturbations of the environment. *Mon. Wea. Rev.*, **127**, 1229–1243.

Cheung, K. K. W. and J. C. L. Chan, 1999b: Ensemble forecasting of tropical cyclone motion using a barotropic model. Part II: Perturbations of the vortex. *Mon. Wea. Rev.*, **127**, 2617–2640.

DeMaria, M., 1985: Tropical cyclone motion in a nondivergent barotropic model. *Mon. Wea. Rev.*, **113**, 1199–1210.

DeMaria, M., 1987: Tropical cyclone track prediction with a barotropic spectral model. *Mon. Wea. Rev.*, **115**, 2346–2357.

DeMaria, M., 2003: Personal communication. re: lack of ensemble improvement related to lack of perturbed time-dependent boundary conditions in Aberson et al. (1995).

DeMaria, M., S. D. Aberson, K. V. Ooyama, and S. J. Lord, 1992: A nested spectral model for hurricane track forecasting. *Mon. Wea. Rev.*, **120**, 1628–1643.

DeMaria, M. and J. M. Gross, 2003: Evolution of prediction models. In Simpson, R., editor, *Hurricane! Coping with Disaster*, chapter 4, pages 103–126. Amer. Geophys. Union.

DeMaria, M., M. B. Lawrence, and J. T. Kroll, 1990: An error analysis of atlantic tropical cyclone track guidance models. *Wea. Forecasting*, **5**, 47–61.

Dudhia, J., 1993: A nonhydrostatic version of the Penn state–NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121**, 1493–1513.

Ehrendorfer, M., 1994a: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.*, **122**, 704–713.

Ehrendorfer, M., 1994b: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part II: Applications. *Mon. Wea. Rev.*, **122**, 714–728.

Ehrendorfer, M. and J. Tribbia, 1997: Optimum prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.*, **54**, 286–313.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Fiorino, M. and R. L. Elsberry, 1989: Some aspects of vortex structure related to tropical cyclone motion. *J. Atmos. Sci.*, **46**, 975–990.

Franklin, J. L. and M. DeMaria, 1992: The impact of Omega dropwindsonde observations on barotropic hurricane track forecasts. *Mon. Wea. Rev.*, **120**, 381–391.

Fulton, S. R., 2001: An adaptive multigrid barotropic tropical cyclone track model. *Mon. Wea. Rev.*, **129**, 138–151.

Gleeson, T. A., 1970: Statistical-dynamical predictions. *J. Appl. Meteor.*, **9**, 333–344.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193.

Goldenberg, S. B., C. W. Landsea, A. M. M.-N. nez, and W. M. Gray, 2001: The recent increase in atlantic hurricane activity: Causes and implications. *Science*, **293**, 474–479.

Gray, W. M., 1984: Atlantic seasonal hurricane activity. part I: El Niño and 30-mb Quasi-Bienniel Oscillation influences. *Mon. Wea. Rev.*, **112**, 1649–1668.

Heming, J. T., S. Robinson, C. Woolcock, and K. Mylne, 2004: Tropical cyclone ensemble forecast product development and verification at the MET Office. Preprints, *26th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, ???–???

Hoffman, R. N. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100–118.

Horsfall, F., M. DeMaria, and J. M. Gross, 1997: Optimal use of large-scale boundary and initial fields for limited-area hurricane forecast models. Preprints, *22nd Conf. on Hurricanes and Tropical Meteorology*, Fort Collins, Colorado, 571–572.

Jarvinen, B. R., C. J. Neumann, and M. A. S. Davis, 1984: A tropical cyclone data tape for the North Atlantic basin, 1886-1983: Contents, limitations, and uses. NOAA Technical Memorandum 22, NWS/NHC. 21 pp. [Available from NOAA/NWS/NHC, Miami, FL 33165.].

Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons Ltd. 240 pp.

Kharin, V. V. and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150.

Krishnamurti, T. N., R. Correa-Torres, G. Rohaly, D. Oosterhof, and N. Surgi, 1997: Physical initialization and hurricane ensemble forecasts. *Wea. Forecasting*, **12**, 503–514.

Kumar, T. S. V. V., T. N. Krishnamurti, M. Fiorino, and M. Nagata, 2003: Multimodel superensemble forecasting of tropical cyclones in the pacific. *Mon. Wea. Rev.*, **131**, 574–583.

Kurihara, Y. M., R. E. Tuleya, and M. A. Bender, 1998: The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Mon. Wea. Rev.*, **126**, 1306–1322.

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

Leslie, L. M., Abbey, Jr, R. F., and G. J. Holland, 1998: Tropical cyclone track predictability. *Meteorol. Atmos. Phys.*, **65**, 223–231.

Leslie, L. M. and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting*, **5**, 158–161.

Leslie, L. M. and G. J. Holland, 1995: On the bogussing of tropical cyclones in numerical models: A comparison of vortex profiles. *Meteorol. Atmos. Phys.*, **56**, 101–110.

Liu, Q. L., S. J. Lord, N. Surgi, H.-L. Pan, and F. Marks, 2002: Hurricane initialization using reconnaissance data in GFDL hurricane forecast model. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, 267–268.

Liu, Q. L., T. P. Marchock, H.-L. Pan, M. Bender, and S. J. Lord, 2000: Improvements in hurricane initialization and forecasting at NCEP with global and regional (GFDL) models. Technical Procedures Bulletin 472, NCEP/EMC. Available from NOAA/NWS Silver Spring MD, or at: `http://205.156.54.206/om/tpb/472.htm`.

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

Marchok, T. P., Z. Toth, and Q. Liu, 2002: Use of the NCEP global ensemble for tropical cyclone track forecasting. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, 176–177.

Mason, S. J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.

McAdie, C. J. and M. B. Lawrence, 2000: Improvements in tropical cyclone track forecasting in the Atlantic basin, 1970-98. *Bull. Amer. Meteor. Soc.*, **81**, 989–997.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

Mundell, D. B. and J. A. Rupp, 1995: Hybrid forecast aids at the Joint Typhoon Warning Center: Applications and results. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 216–218.

Murphy, A. H., 1978: Hedging and the mode of expression of weather forecasts. *Bull. Amer. Meteor. Soc.*, **59**(4), 371–373.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

Murphy, A. H. and E. S. Epstein, 1967a: A note on probability forecasts and "hedging". *J. Appl. Meteor.*, **6**, 1002–1004.

Murphy, A. H. and E. S. Epstein, 1967b: Verification of probabilistic predictions: A brief review. *J. Appl. Meteor.*, **6**, 748–755.

Neumann, C. J. and J. M. Pelissier, 1981: Models for the prediction of tropical cyclone motion over the North Atlantic: An operational evaluation. *Mon. Wea. Rev.*, **109**, 522–538.

Pan, H.-L., Q. Liu, N. Surgi, and S. Lord, 2002: NCEP global model tropical forecast upgrades: Model performance during the 2001 hurricane season. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA.

Pasch, R. J., L. A. Avila, and J. L. Guiney, 2001: Atlantic hurricane season of 1998. *Mon. Wea. Rev.*, **129**, 3085–3123.

Pasch, R. J., M. B. Lawrence, L. A. Avila, J. L. Beven, J. L. Franklin, and S. R. Steward, 2004: Atlantic hurricane season of 2002. *Mon. Wea. Rev.*, **132**, 1829–1859.

Peng, M. S. and R. T. Williams, 1990: Dynamics of vortex asymmetries and their influence on vortex motion on a β-plane. *J. Atmos. Sci.*, **47**, 1987–2003.

Pielke, Jr, R. A. and C. W. Landsea, 1998: Normalized hurricane damage in the United States: 1925–95. *Wea. Forecasting*, **13**, 621–631.

Pike, A. C. and C. J. Neumann, 1987: The variation of track forecast difficulty among tropical cyclone basins. *Wea. Forecasting*, **2**, 237–241.

Powell, M. D., S. H. Houston, L. R. Amat, and N. Morrisseau-Leroy, 1998: The HRD real-time hurricane wind analysis system. *J. Wind Engineer. Ind. Aerody.*, **77**, 53–64.

Puri, K., J. Barkmeijer, and T. N. Palmer, 2001: Ensemble prediction of tropical cyclones using targeted diabatic singular vectors. *Quart. J. Roy. Meteor. Soc.*, **127**, 709–731.

Ramamurthy, M. K. and B. F. Jewett, 1999: Ensemble prediction of hurricane opal's track and intensity. Preprints, *23rd Conf. on Hurricane and Tropical Meteorology*, Dallas, TX, Amer. Meteor. Soc., 834–836.

Richardson, L. F., 1922: *Weather Prediction by Numerical Process.* Cambridge University Press. 236 pp. (Reprint, with a new introduction by Sidney Chapman, Dover Publications, 1965, 236 pp.).

Rogers, E., S. L. Stephen, D. G. Deaven, and G. J. DiMego, 1993: Data assimilation and forecasting for the Tropical Cyclone Motion Experiment at the National Meteorological Center. Preprints, *20th Conf. on Hurricanes and Tropical Meteorology*, San Antonio, TX, Amer. Meteor. Soc., 329–330.

Sanders, F., 1973: Skill in forecasting dailiy temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171–1178.

Sanders, F., A. L. Adams, N. J. B. Gordon, and W. D. Jensen, 1980: Further development of a barotropic operational model for predicting paths of tropical storms. *Mon. Wea. Rev.*, **108**, 642–654.

Sanders, F. and R. H. Burpee, 1968: Experiments in barotropic hurricane track forecasting. *J. Appl. Meteor.*, **7**, 313–323.

Sanders, F., A. C. Pike, and J. P. Gaertner, 1975: A barotropic model for operational prediction of tracks of tropical storms. *J. Appl. Meteor.*, **14**, 265–280.

Sheets, R. C., 1985: The National Weather Service hurricane probability program. *Bull. Amer. Meteor. Soc.*, **66**, 4–13.

Sheets, R. C., 1990: The National Hurricane Center – past, present, and future. *Wea. Forecasting*, **5**, 185–232.

Shuman, F. G., 1989: History of numerical weather prediction at the National Meteorological Center. *Wea. Forecasting*, **4**, 286–296.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105**, 228–229.

Toth, Z., 2004: Personal communication. re: Is vortex relocation used in the perturbed members of the GFS ensemble? An e-mail exchange.

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

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.

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

Tracy, J. D., 1966: Accuracy of Atlantic tropical cyclone forecasts. *Mon. Wea. Rev.*, **94**, 407–418.

Velden, C., 1993: The relationship between tropical cyclone motion, intensity and the vertical extent of the environmental steering layer in the Atlantic basin. Preprints, *20th Conf. on Hurricanes and Tropical Meteorology*, San Antonio, TX, Amer. Meteor. Soc., 30–34.

Vigh, J., 2002: Track forecasting of 2001 Atlantic tropical cyclones using a kilo-member ensemble. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 212–213.

Vigh, J., 2004: Evaluation of a kilo-member ensemble for track forecasting. Preprints, *26th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 160–161.

Vigh, J., S. R. Fulton, M. DeMaria, and W. H. Schubert, 2003: Evaluation of a multigrid barotropic tropical cyclone track model. *Mon. Wea. Rev.*, **131**, 1629–1636.

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

Weber, H. C., 2003: Hurricane track prediction using a statistical ensemble of numerical models. *Mon. Wea. Rev.*, **131**, 749–770.

Wiin-Nielsen, A., 1959: On barotropic and baroclinic models, with special emphasis on ultra-long waves. *Mon. Wea. Rev.*, **87**, 171–183.

Wilks, D. S., 1995: *Statistical Methods in Atmospheric Science*. Academic Press. 467 pp.

Zhang, Z. and T. N. Krishnamurti, 1997: Ensemble forecasting of hurricane tracks. *Bull. Amer. Meteor. Soc.*, **78**, 2785–2795.

Zhang, Z. and T. N. Krishnamurti, 1999: A perturbation method for hurricane ensemble predictions. *Mon. Wea. Rev.*, **127**, 447–469.

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

Zwiers, F. W. and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351.