Stony Brook University

Verification of Extratropical Cyclones Within NCEP Forecast Models Using an Automated Tracking Algorithm

A Thesis Presented

by

Michael Charles

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Science

in

Marine and Atmospheric Science

Stony Brook University

May 2008

Stony Brook University

The Graduate School

Michael Charles

We, the thesis committee for the above candidate for the
Master of Science degree, hereby recommend
acceptance of this thesis.

Brian A. Colle – Thesis Advisor
Associate Professor
School of Marine and Atmospheric Sciences

Edmund K.M. Chang
Professor
School of Marine and Atmospheric Sciences

Sultan Hameed
Professor
School of Marine and Atmospheric Sciences

This thesis is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Thesis

Verification of Extratropical Cyclones Within NCEP Forecast Models Using an Automated Tracking Algorithm

by

Michael Charles

Master of Science

in

Marine and Atmospheric Science

Stony Brook University

2008

This thesis provides the first comprehensive verification of extratropical cyclones around North America and the adjacent oceans within the operational models at the National Centers for Environmental Prediction (NCEP) in the last several years. The sea-level pressure (SLP) errors were quantified for cyclones within the NCEP Global Forecast System (GFS) and North American Mesoscale (NAM) models for the 2002-2007 cool seasons (October to March), as well as the NCEP Short-Range Ensemble Forecast (SREF) system for the 2004-2007 cool seasons. The GFS analysis was used as truth for cyclone central pressure, since surface observations indicated that the cyclone pressures in the GFS are more accurate than the NAM and NARR on average. The average position of a GFS- and NAM-analyzed cyclone was used as truth for the cyclone position.

The NCEP GFS is more skillful than the NAM over the continental U.S. and adjacent oceans, with the GFS having a much larger skill advantage over the eastern Pacific given the positive bias in the NAM cyclone central pressure over the Pacific. The GFS and NAM cyclone errors over the Pacific are larger than other regions for the short-term (0-60 h) forecasts. However, the errors for relatively deep cyclones over the western Atlantic are larger than the eastern Pacific by hour 102, since the cyclones over the western Atlantic are underpredicted (too weak) on average in the extended range. Many large cyclone errors over the eastern U.S. in the 72-96 h range can be traced to the relatively large errors over the eastern Pacific 2-3 days earlier. For the extended GFS (> 72 h), cyclones over the western Atlantic are displaced too far east on average. In contrast, both the GFS and the NAM have a significant westward displacement bias for cyclones before forecast hour 60 over the western Atlantic, with the NAM errors significantly larger than the GFS. There has been little improvement in the model 0-2 day cyclone forecasts during the past 5 years over the eastern U.S., while there has been a large improvement in the cyclone pressure predictions over the Pacific in the NAM.

The NCEP SREF was also verified for the 2004-2007 cool seasons for hours 0-60. The SREF cyclone pressure and displacement absolute errors are smaller than the operational NAM in many regions, but not the operational GFS. Not all SREF members are equally skillful for the cyclone central pressure and displacement. Decomposition of the Brier score shows that the SREF overforecasts the probability of deep cyclone events in all regions, and has poor resolution for average cyclone events, which factors large (poor) Brier scores. Brier skill scores indicate that a blend of the GFS and NAM has nearly the same amount of predictive skill as the SREF. The SREF suffers from overdispersion and positive biases in many regions early in the forecast, which hurts its probabilistic performance.

Table of Contents

List of Figures

List of Tables

I. INTRODUCTION

a. Background

Extratropical cyclones are responsible for a wide range of hazardous weather, such as flooding, severe convection, strong winds, and heavy snow. For example, the 9-11 September 1978 extratropical cyclone, also known as the Queen Elizabeth II (QEII) storm, explosively deepened nearly 60 mb to 945 mb over a 24 hour period, which resulted in damage to the ocean liner Queen Elizabeth II and the sinking of a trawler near Newfoundland (Gyakum 1983). The "Inauguration Day cyclone" made landfall on the Washington coast on 20 January 1993, with wind gusts reaching as high as 45 m s^{-1} to the Pacific Northwest. This storm resulted in 5 casualties, the destruction of 79 homes, and left 750,000 households without power across the Puget Sound region (Steenburgh and Mass 1996). Meanwhile, during the 12-14 March 1993 superstorm (Kocin et al. 1995), there was snow as far south as the Florida Peninsula and 25-75 cm of snow fell from Alabama to Maine, while winds gusted over hurricane force along parts of the East Coast.

The skill in numerical weather prediction models in forecasting these major storm events has varied. Some storms, such as the Superstorm 1993, were well forecast several days in advance (Uccellini et al. 1995). An accurate 1-2 day mesoscale simulation of the Inauguration Day Storm was also obtained over the Pacific Northwest (Steenburgh and Mass 1996). On the other hand, other cyclones, such as the 25 January 2000 cyclone along the mid-Atlantic coast (Zhang et al. 2002), have been forecast relatively poorly at 1-2 day lead times. There have been few recent studies looking at the predictability of hundreds of extratropical cyclone events across North America and its surrounding oceans, so a long-term verification is needed for model predicted cyclones. This thesis first reviews the climatology and some previous verification studies of extratropical cyclone events as motivation for this research.

b. Previous Cyclone Climatologies

There have been several studies quantifying the frequency and characteristics of extratropical cyclones across the U.S. and surrounding oceans. Colucci (1976) employed a manual method of tracking winter cyclones over the eastern U.S. and western Atlantic between 1964 and 1973 using 3-hourly North American surface charts. This study indicated several favored cyclone tracks over the eastern U.S., such as one from coastal North Carolina east-northeastward along the northern boundary of the Gulf Stream (Miller Type A, Miller 1946), one from Cape Hatteras northeastward across Cape Cod and into the Gulf of Maine (Miller Type B, Miller 1946), and one over the eastern Great Lakes. These tracks agree with those found by Reitan (1974) and Zishka and Smith (1980).

Sanders and Gyakum (1980) defined a "bomb" cyclogenesis event as a deepening cyclone in which the central pressure falls at a rate of 1 Bergeron, which ranges from 28 mb $(24\ h)^{-1}$ at the pole to 12 mb $(24\ h)^{-1}$ at 25 °N. It was found that the majority of bombs were oceanic cyclones, with the largest maxima occurring in the westernmost portions of the Pacific and Atlantic along the largest sea surface temperature (SST) gradients.

Roebber (1984) manually cataloged 1-year of extratropical cyclones (February 1980 through January 1981) using National Meteorological Center (NMC) 12 hourly surface charts. This study expanded upon Sanders and Roebber (1980) by also including non-explosively deepening cyclones. The regions with most prominent cyclogenesis were located off the coast of Japan, in the lee of the Rockies, near the southeast coast of the U.S., and offshore the coast of Newfoundland. Explosive cyclones were mainly confined to the westernmost portions of the Pacific and Atlantic oceans, as in Sanders and Roebber (1980).

Hoskins and Valdes (1990) used a linear, stationary wave baroclinic model, described by Hoskins and Simmons (1975), to investigate storm tracks across the Northern Hemisphere. Cool season extratropical cyclones tend to develop in the highly baroclinic environments associated with the boundary between cold continental offshore flow and the warmer ocean surfaces over the western Pacific and Atlantic Oceans. They found that the major reason for the persistence of the Northern Hemisphere Pacific (Atlantic) storm track is the latent heating from condensation over the relatively warm waters along the east coast of Asia (North America). Chang et al. (1993) showed that energy leading to downstream development of a baroclinic wave initially originates from ageostrophic geopotential fluxes, and later on from conversion of baroclinic potential energy into kinetic energy.

The development of automated cyclone tracking routines made it possible to study cyclone properties over longer time periods. Hirsch et al. (2001) used an automated routine and the National Center for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis dataset (1948, 1951-97) at 2.5°x 2.5° spatial resolution in order to construct a long-term climatology of East coast cyclones. The authors found that East Coast winter storms had an average minimum pressure of ~993 mb, with average maximum winds of 20.5 m s^{-1}. They also found that El Niño increased East coast cyclone activity by 44% over neutral winters. Eichler and Higgins (2006) showed a southward shift in the North Pacific storm track during El Niño, which favors more cyclogenesis along the East Coast. In contrast, La Niña periods are associated with a northward shift of cyclones in the North Pacific and the U.S. east coast storm tracks, as well as more intense Midwest/Great Lakes cyclones.

Hoskins and Hodges (2002) used an automated routine to locate and track 22 years of cyclones using European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis grids from 1979-2000. They found mean storm tracks consistent with Miller (1946), Collucci (1976), and Roebber (1984), thus indicating that an automated routine can accurately track cyclones.

c. Cyclone Predictability

Improving our understanding of extratropical cyclone predictability requires an objective verification of many historic cyclone events. There have been previous studies that have quantified model forecast performance of extratropical cyclones. For example, Silberberg & Bosart (1982) investigated forecast errors in the Limited area Fine Mesh model (LFM) in 1972, which was operational at 190.5 km horizontal grid spacing. They found that the LFM underdeepened Atlantic and Pacific cyclones by 6 to 10 mb at forecast hour 48, and overdeepened cyclones to the east of the Rockies through the

eastern U.S. by up to 4 mb. Mullen & Smith (1990) found that the Nested Grid Model (NGM) during the 1987-88 cool season for forecast hours 24 and 48 also tended to underdeepen oceanic cyclones (4-8 mb in the Pacific, and 2-6 mb in the Atlantic) and cyclones in the lee of the Rockies (up to 2 mb). Smith and Mullen (1993) found that at the NGM, in the 1987-88 and 1989-90 cool seasons at forecast hour 24 and 48, slightly overdeepened weak and continental cyclones (2-4 mb), and underdeepened deep and oceanic cyclones (2-6 mb in the Pacific, and 2-4 mb in the Atlantic). Meanwhile, the medium range Aviation model (AVN) during the same cool seasons underdeepened all cyclones (2-4 mb, except up to 6 mb over the Pacific). The mean displacement error for the AVN at hour 48 was at least 30 km less than the NGM, and at least 60 km less for stronger cyclones (< 980 mb). The NGM had a westward track bias over the Atlantic of ~100 (~150) km at 24 (48) hours, while the track bias for the AVN was not statistically different than zero.

More recently, Froude et al. (2007a) globally tracked extratropical cyclones using both 850 mb vorticity fields from 2 winters and 2 summers of ECMWF 0-7 day forecasts. It was found that the model skill in forecasting cyclone intensity and position was greater in the northern hemisphere (NH) than the southern hemisphere (SH), with about 1 day added skill in the NH. When considering only intense storms (850 mb vorticity >= 8.0×10^{-5} s^{-1}) the NH model skill is reduced by about 1 day for both season, with less of an effect in the SH. They also found a negative (slow) bias of about -0.25 m s^{-1} in the propagation speed of predicted cyclones in the southern hemisphere, and about -0.5 to -0.8 ms^{-1} in the southern hemisphere.

Uccellini (2008) studied the difference in skill of oceanic cyclone predictions by the forecasters at the NOAA/NCEP Ocean Prediction Center (OPC) between the 1992-93 and 2002-05 cool seasons for the Pacific, and for the 2002-2005 cool seasons for the Atlantic. It was found that day 4 forecasts of these cyclones have improved since the early 90s. The mean SLP errors were positive, indicating that the cyclone strength is frequently underforecast. Absolute errors of SLP were larger in the Atlantic than the Pacific by day 4 of the forecast for the 2002-2005 cool seasons, and larger in both regions for more intense cyclones. Pacific position errors have also decreased 15-20% since the early 90s, with position errors generally lower in the Atlantic and for more intense cyclones.

Cyclone forecast errors result from uncertainties in initial conditions and model physics (Hacker et al. 2003). Silberberg and Bosart (1982) examined cyclone forecast errors by the LFM for one cool season and showed the importance of IC errors in deteriorating the skill of 1-2 day forecasts for several cyclone events. For the 24-25 January 2000 “surprise” snowstorm, Langland et al. (2002) attributed most of the forecast error to growth of IC uncertainties 96 hours before the event over the eastern Pacific. In addition, Brennan & Lackmann (2005) found that operational models did not forecast the large area of convection over Alabama and Georgia for the 25 January 2000 event, and therefore missed the low-level PV maximum from the convection. This low-level PV maximum is important in intensifying the low-level circulation associated with a cyclone (Davis & Emanuel 1991). Zhang et al. (2002) found that the Fifth-Generation NCAR / Penn State Mesoscale Model (MM5) better resolved the precipitation patterns for the “surprise” snowstorm with increasing horizontal resolution. They also initialized the

MM5 with several different initial conditions, and found that an IC ensemble significantly improved the forecast.

To estimate the forecast uncertainty in operational models, a number of slightly perturbed IC's can be used to initialize an ensemble of model runs, such as the "breeding" method used at NCEP (Toth and Kalnay 1997). The different ICs can be combined with different physics to diversify the ensemble (Eckel and Mass 2005). Du et al. (1997) studied the impacts of short-range ensemble forecasting on the precipitation forecasts for the 14-16 December 1987 cyclone over the continental U.S., which deepened to 978 mb and produced a foot or more of snow in a band from the Texas Panhandle to Michigan. It was found that simply averaging the precipitation at forecast hour 36 from a 25-member Mesoscale Model 4 (MM4) ensemble reduced the root-mean-squared error (RMSE) of the precipitation forecast by about 50-60%. They showed that an ensemble of PSU-NCAR MM4 members at 80 km grid spacing could match or exceed the skill of a deterministic model at 40-km grid spacing when forecasting 500 mb geopotential height, 850 mb temperature, and surface precipitation.

Stensrud et al. (1999) illustrated the benefits of ensemble forecasting of cyclone positions by comparing the performance of a 29-km Meso Eta model and a 10-member 80-km Eta ensemble from the NCEP Short Range Ensemble Forecast system (SREF) in 1994. It was found that the mean of the lower resolution ensemble Eta members was as accurate as the higher resolution deterministic Eta in forecasting cyclone position on average across the U.S. Ensemble spread was greatly increased by including the RSM members, suggesting that model diversity within an ensemble is important in determining the forecast uncertainty.

Froude et al. (2007b) verified extratropical cyclone tracks in the ECMWF ensemble prediction system (EPS) and the GFS-based NCEP EPS between 6 January 2005 and 5 April 2005. The ECMWF EPS consisted of 50 perturbed members out to 10 days, with a spectral resolution of T25540L, while the 10-member NCEP EPS was run resolution of T126L28 (T62L28) for the first 7.5 (last 8.5) days. They found that the ECMWF ensemble has greater skill than the NCEP ensemble for cyclones in the NH, while in the SH the NCEP ensemble has more skill. The ECMWF ensemble mean had greater skill than the control member, although the ECMWF ensemble was under-dispersed.

d. Motivation

Since the mid-1990's, there has been little objective cyclone verification of the NCEP short-range operational models, such as the North American Mesoscale model (NAM) and Global Forecast System (GFS) models. Several recent studies, such as McMurdie and Mass (2004), Colle et al. (2002), and Uccellini (2008), have studied the performance of NWP models in forecasting cyclones. However, McMurdie and Mass (2004) only focused on a select number of cyclone cases over the Pacific NW from the ETA model, while Uccellini (2008) focused on the skill of forecasters, and not specific model performance. This thesis will address this by completing a multi-year climatology of cyclone errors across the U.S. and surrounding oceans for the 1-5 day forecast period for the GFS and NAM.

While ensembles have been shown to outperform deterministic models for several meteorological variables, and have been shown to improve cyclone forecasts on an individual case study basis (Buizza and Chessa 2002), there has not been much assessment of ensemble predictability of extratropical cyclone strength and position over a few seasons. Froude et al. (2007b) provided a quantitative assessment of the ECMWF and NCEP EPS; however, their data available was limited to one year, and they focused primarily on the extended forecasts. This study will provide a 3-year verification of NCEP's SREF in comparison with the deterministic GFS and NAM.

A long-term verification dataset of cyclones will also help our understanding of the large-scale factors that control the predictability of these events at different forecast periods. For example, it is hypothesized that large cyclone errors in one region may impact errors later in the forecasting at another region downstream to the east.

Overall, this study will address the following motivational questions:

1. *What are the cyclone position and strength errors in NCEP operational models for the 1-5 day forecasts?*
2. *How do the forecast errors vary from the eastern Pacific eastward to the western Atlantic?*
3. *Have cyclone forecasts improved since the previous generation of NCEP models?*
4. *Do cyclone errors in the Pacific have a downstream impact on cyclone forecasts in the Eastern U.S.?*
5. *Is there an advantage to using ensembles to forecast cyclones?*

II. DATA AND METHODS

a. Data

Cyclones were tracked in the Global Forecast System (GFS) and North American Mesoscale (NAM, formally the ETA) models, and one operational ensemble, the Short Range Ensemble Forecast system (SREF). The GFS is a global spectral model run at the National Center for Environmental Prediction (NCEP) 4 times daily. It uses a T254 global Gaussian grid, which is roughly equivalent to a horizontal resolution of 0.5°. Since the GFS became operational in 1985 (then known as the Medium Range Forecast, MRF model), it has undergone many upgrades (Table 2.1). The NAM is a regional mesoscale model run four times daily. Since 2001 the NAM has been run at 12-km grid spacing with 60 vertical layers. The NAM has also undergone many upgrades over the past several years (Table 2.2), including a change to the Weather Research and Forecasting (WRF)-NMM core in 2006 (Du et al. 2006).

Observed and forecast cyclones were tracked for the 5 cool seasons (October through March) from 2002 to 2007. For the first 3 cool seasons (2002-2005), the GFS grids were available 4 times daily out to forecast hour 120 every 6-h at ~95-km grid spacing across all of North America and surrounding oceans (Fig. 2.1). During these same 3 cool seasons, the NAM was available every 6-h on a ~40 km grid for a ~1000 km region surrounding the continental U.S. (CONUS), out to 60 hours at 0000 and 1200 UTC, and 48 hours at 0600 & 1800 UTC. For the 2005-2007 cool seasons, the NAM grids were available four times daily at 12-km grid spacing out 60 hours, while the GFS had a 1-degree resolution at the same temporal resolution. The North American regional

reanalysis (NARR) grids (Mesinger et al. 2006) were also used to track observed cyclones, but these analyzed cyclones were found to be not as accurate as the GFS analyses, as will be discussed in chapter III.a. The NARR grids around North American and adjacent oceans were available at ~32-km resolution every 3 hrs.

The SREF was available for the last 3 years of the study (2004-2007) at ~40-km grid spacing over the CONUS and adjacent oceans (Fig. 2.1). There were two runs daily at 0900 and 2100 UTC, with 3-hourly forecast intervals to 63 hours. The SREF consisted of 15 members for the first 2 years: 5 ETA members using Bett-Miller-Janic convection (EBM), 5 ETA members using Kain-Fritsch convection (EKF), and 5 RSM members using Simple Arakawa-Shubert convection (RSM). During the last (2006-2007) cool season, six additional Weather Research and Forecasting (WRF) members were also available, with three members from the Non-hydrostatic Mesoscale Model (NMM) core, and two others from the Advanced Research WRF (ARW) core (Table 2.3).

The method by which members are perturbed is the *breeding method*, described by Toth and Kalnay (1997). First a small arbitrary perturbation is added to the 3-day old global analysis (for the global ensemble's breeding cycle). This perturbed analysis is then truncated to the SREF domain and used as an initial perturbed field for the SREF breeding cycle. The model is integrated out to 12 hours from both the original initial state and the perturbed initial state. The difference between these forecasts is then scaled down so that the RMS difference between the two 850 mb temperature forecasts is less than 0.5 °C. This scaled perturbation is then used to initialize another 12 hour run, and this process is repeated for consecutive runs up to the current analysis time, in essence *breeding* growing errors (typically associated with baroclinic instability), and eliminating decaying errors. After several days, the errors will reach an asymptotic value. This new perturbation is added to and subtracted from the current SREF analysis (EDAS for ETA members and GDAS for RSM members) to generate positive and negative ensemble members, respectively, for the current SREF cycle.

Since the automated cyclone tracker program utilized model data on a latitude-longitude grid, all grids were interpolated to 0.8° latitude-longitude grids. This resolution was applied to all model grids since it is closest to the GFS data, which has the coarsest resolution of all models.

b. Methods

The cyclone tracking routine for this study was originally developed by Timothy Marchok at NCEP in order to track tropical cyclones (Marchok 1998). For this thesis several changes were made to the system, including the ability to better track extratropical cyclones, and increased precision for cyclone central pressure and position. A cyclone was located and tracked by first locating the lowest sea-level pressure (SLP) point on a latitude-longitude grid. In order to show that this point was associated with a regional area of lower pressure and not just an artifact of a highly variable SLP field, a SLP gradient of 1.5 mb per 1000 km (0.0015 $^{mb}/_{km}$) was required anywhere within 300 km of this lowest pressure point. If this gradient was satisfied, then it was determined whether a 2-mb closed contour surrounded the grid point. If either of these tests failed, the grid point was masked out and the point with the next lowest SLP on the grid was tested. If a 2-mb closed contour was found, then the gridpoint was flagged as a cyclone,

and the cyclone's latitude, longitude, and central pressure were recorded. To prevent locating any additional points with this cyclone in future iterations, all grid points surrounding the cyclone to the location where the SLP gradient reversed were masked out. All the above steps were repeated until either all grid points were masked out, or all remaining SLP values were greater than one-half standard deviation above the domain-wide mean SLP. This tracking routine was run for every analysis and forecast grid in the dataset.

In order to objectively verify the cyclone forecasts, an accurate sea-level pressure analysis was needed for the cyclone central pressure and position. For this study the analyzed central pressure was obtained from the GFS initialization, while the position of the observed cyclone was obtained by averaging the position from the GFS and NAM initial analyses. The reasons for this analysis selection will be discussed in chapter III.a. In order to keep track of the forecast cyclones, a unique identification number (ID) was assigned to each cyclone event for a particular model run. For example, at forecast hour zero (F000), all cyclones were given a unique ID, and an attempt was made to guess the future position of the cyclone at F006 using the mean 850-500 mb wind as the steering flow. At the next time (F006), a search was done around the new guess-position (radius a function of storm speed) to locate the relevant cyclone. If the cyclone was found again, and the average speed the storm would have had to travel to get to this point was less than 30 m s^{-1}, the cyclone was given the same ID as the previous time (F000). If the cyclone was not found, or if its speed exceeded 30 m s^{-1}, then that ID was retired.

To test the skill of the tracking routine, all F048 GFS cyclones during October 2004 were checked manually against the GFS analyses. A "false alarm" was counted if a cyclone was marked by the routine but it did not have at least 1 closed 2mb-contour. Cyclones found manually to have at least one closed 2-mb contour, but not marked by the routine were considered "missed." For this month, 647 cyclones occurred in the GFS between 20 and 70°N. Of the 647 cyclones, 11 were missed by the tracking routine and there were 78 false alarms. The probability of detection (number matched divided by the total number that occurred) was found to be ~98%, and the false alarm rate (number of false alarms divided by the total number that were matched) was found to be ~11%. These numbers suggest that the tracking routine is fairly reliable for model validation. Nearly all of the false alarms occurred over the high terrain of the western U.S., where SLP reduction is difficult and spurious lows are sometimes found by the tracker, given the irregularly-shaped pressure gradients over steep terrain. The majority of missed cyclones had a fairly weak pressure gradient associated with them, which made it difficult for the tracker to find a closed 2-mb contour.

Once all forecast cyclones were located and tracked, each forecast cyclone was paired with an observed cyclone. For each model run, the distance between each forecast and observed cyclones was calculated. These distances were ordered from smallest to largest, and the forecast and observed cyclones were paired by choosing forecast and observed cyclones with the smallest separation (Fig. 2.2). This pairing process continued until either: 1. All forecast cyclones were matched, in which case any remaining observed cyclones were considered "missed" by the model; 2. All observed cyclones were matched, in which case any remaining forecast cyclones were considered false alarms; or 3. The pairing distance exceeded 800 km, in which cases any remaining observed (forecast) cyclones were considered missed (false alarms). This pairing process was also

manually checked for October 2004 during which there were 524 GFS forecast and observed cyclones matched at hour 48. There were only 10 pairs that were considered incorrectly matched, giving the pairing algorithm a ~98% accuracy. For this study, only cyclones that were successfully matched were considered in the verification.

Cyclones were also grouped into 8 distinct geographical regions (Fig. 2.3). The regions were chosen such that errors noted for particular geographic regions could be summarized succinctly and to insure that there were a significant number of cases per region (Fig. 2.4 and 2.5). The regions also correspond to important geographic features, such as the oceans and the Rocky Mountains.

In order to test for statistical significance, a bootstrapping approach was used to resample the data and to obtain proper confidence intervals around the means (Zwiers 1990). To test if two means were significantly different, first each sample (for example, cyclone central pressure errors over 5 years) was resampled by generating a new sample of the same size by randomly selecting from the same original sample, and allowing for repeated selections. A new sample was generated 1000 times, and the 90% confidence intervals around the mean were determined by finding the 5^{th} and 95^{th} percentile of the means of all 1000 resamples. If the confidence intervals of both tested samples did not overlap, then they were significantly different at the 90% level.

Date	Change	Old	New
10/29/02	Resolution	T170L42 (180 hrs) T62L28 (384 hrs)	T254L64 (84 hrs) T170L42 (180 hrs) T126L28 (384 hrs)
08/28/03	Longwave Radiation	Geophysical Fluid Dynamics Laboratory (GFDL) model	Rapid Radiative Transfer Model (RRTM)
02/24/04	Mountain Blocking	---	Parameterization of the separation of airflow in the vertical over mountainous terrain.
05/31/05	Resolution	T254L64 (84 hrs) T170L42 (180 hrs) T126L28 (384 hrs)	T382L64 (180 hrs) T190L64 (384 hrs)

Table 2.1. Major updates to the Global Forecast Systen (GFS) model during the period of this study (2002-2007).

Date	Change	Old	New
03/04	Data Assimilation	---	Daily rain gauge data bias adjusted
05/03/05	Eta 3DVAR	---	Surface temperatures turned on
		---	Use level II.5 WSR-88D radial velocity data
	Precipitation assimilation	---	Improved use of observed precipitation
	LSM changes	---	Improved LSM
	Clouds/radiation	---	Improved cloud cover and radiation scheme
06/20/06	Model replaced	Eta step-mountain coordinate model	WRF Non-hydrostatic Mesoscale (WRF-NMM) model
	Data Assimilation	Eta 3D-Variational Analysis (EDAS)	Gridpoint Statistical Interpolation (GSI)
	Model initialization	---	High resolution snow analysis
	Data assimilation	---	Using WSR-88D Level II radial wind data, GPS-Integrated Precipitable Water (IPW), and NOAA-18 radiances
08/15/06	Model	---	Changes to LSM, diffusion, and SST preprocessing, and added surface temperature data to GSI
12/19/06	Convective parameterization and cloud microphysics	---	Improvements made

Table 2.2. Same as Table 1 except for the NAM model.

Date	Change	Old	New
08/31/05	RSM members		Optimized, domain covers Alaska and Hawaii
	RSM resolution	40 km	45 km
	Forecast length	63 hrs	87 hrs
	Lateral Boundary Conditions	9 hr old GFS ensemble forecasts	3 hr old GFS ensemble forecasts
12/06/05	Added 6 members	---	3 40km/50lev WRF-NMM with EMC physics 3 45km/35lev WRF/ARW with NCAR physics

Table 2.3. Same as Table 1 except for the SREF system.

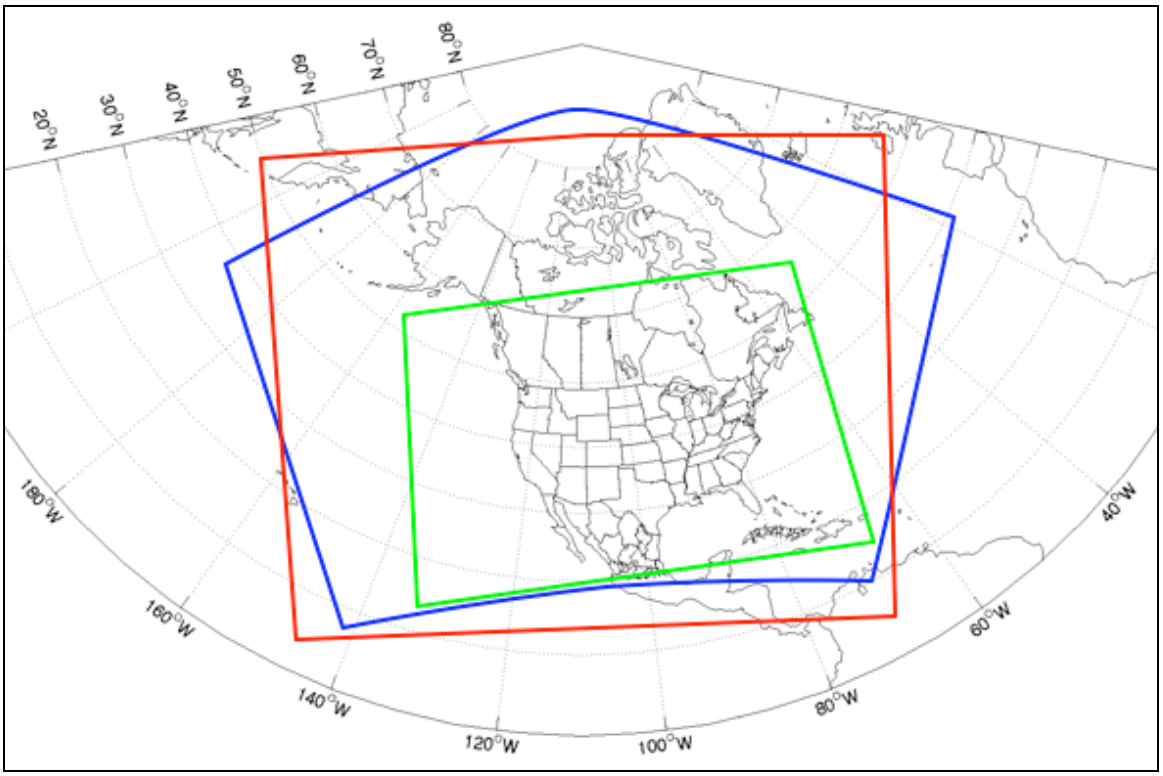

Figure 2.1. Model domain boundaries for the NARR (red), GFS (blue), and the NAM and SREF (green).

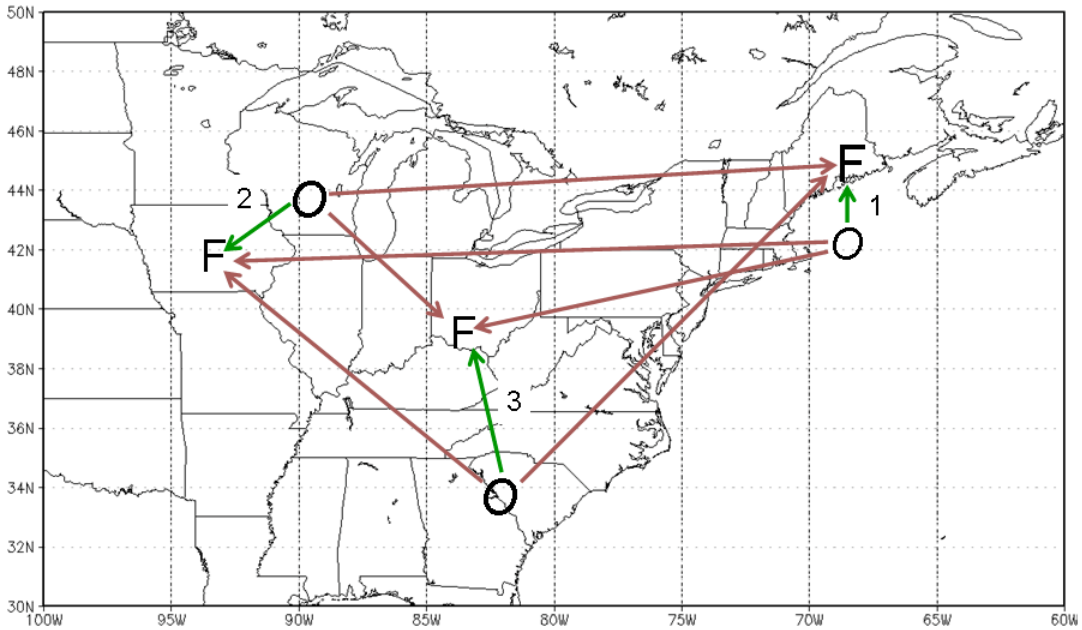

Figure 2.2. Illustration of how the forecast and observed cyclones were paired. Each observed cyclone (O) is paired with all forecast cyclones (F), and the distances are ordered from smallest to largest The smallest distance for each O and F match is used as the pairing for the verification. The green arrows indicate the chosen cyclone pairs using this approach, while the brown arrows indicate cyclone pairs that were not chosen.

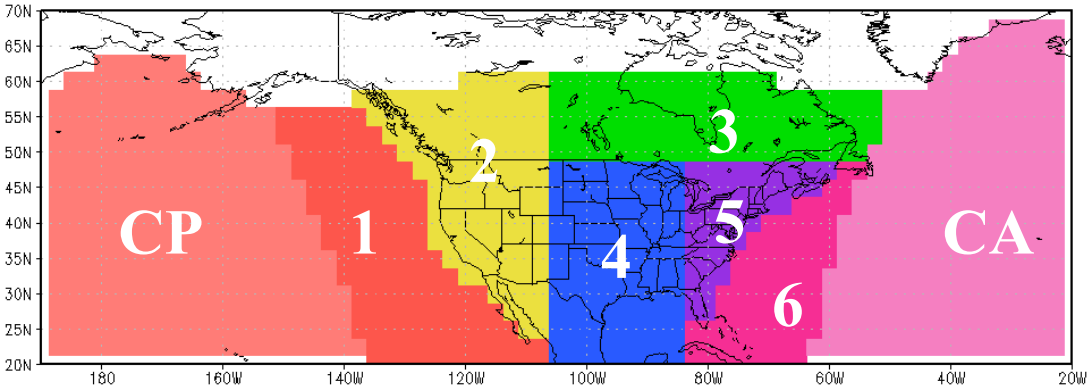

Figure 2.3. Geographical regions used to verify the operational model cyclones. CP is the Central Pacific and CA is the Central Atlantic.

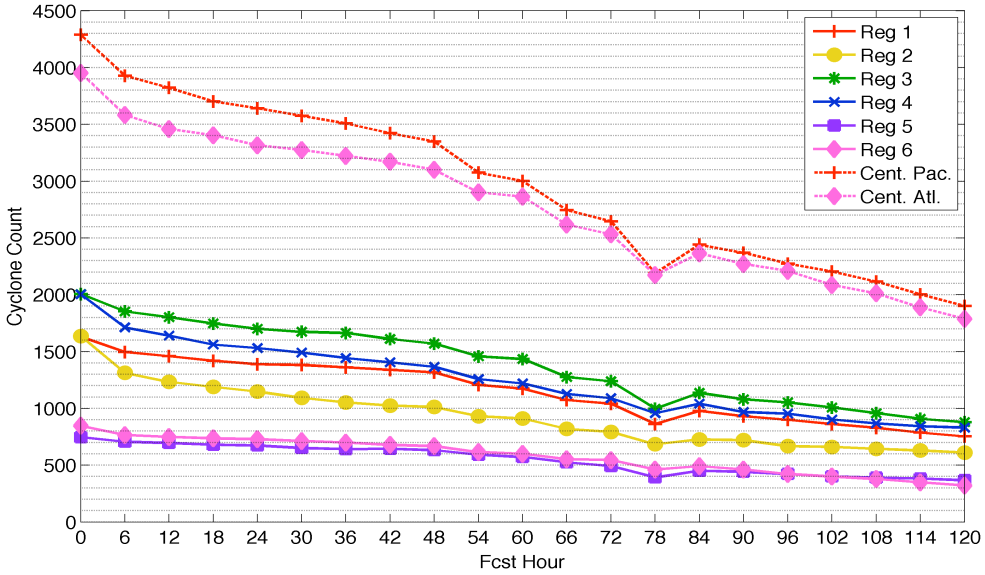

Figure 2.4. Number of GFS cyclones matched at each forecast hour between 0 and 120 h over the 5 cool seasons of 2002-2007.

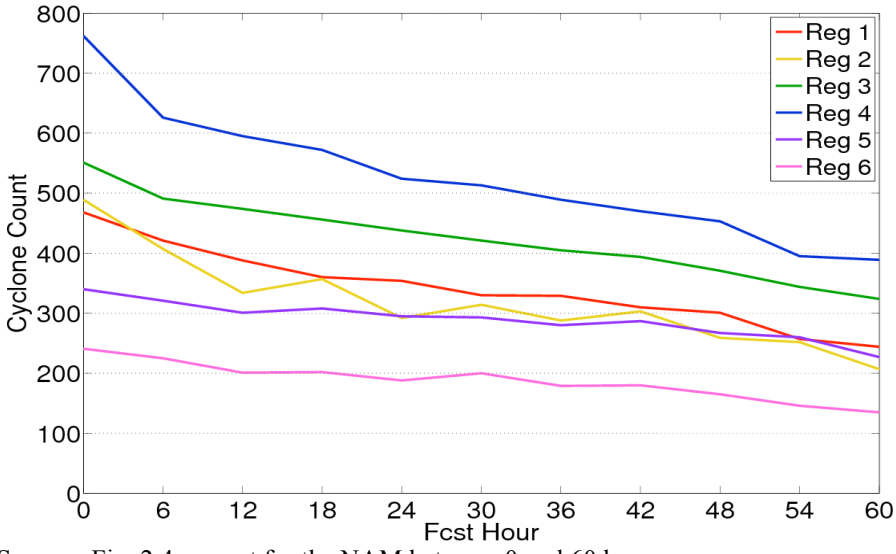

Figure 2.5. Same as Fig. 2.4, except for the NAM between 0 and 60 h.

III. VERIFICATION OF NCEP GFS AND NAM MODELS

a. Cyclone Climatology

After all cyclones were gridded, the spatial distribution of cyclone events during the study period was quantified on a 2.5° grid. Figure 3.1 shows the number of GFS analyzed cyclones identified using the automated approach for each cool season (October - March). For all five years the three largest maxima in cyclone density occurred in the Gulf of Alaska, east of the southern tip of Greenland, and in the lee of the Rockies, which agrees with previous studies (Zishka and Smith 1980, Hoskins and Hodges 2002). It is hypothesized here that the Gulf of Alaska and Greenland have an especially high cyclone count because 1, they are located within major climatological cyclone tracks and 2, they are frequently sites of cyclolysis, during which cyclones occlude, lose speed, and remain in these regions sometimes for days (Hoskins and Hodges 2002). There is large interannual variability in cyclone frequency at particular locations. For example, during the 2002-2003 cool season, there were 3-6 cyclones per 2.5° grid box in the Gulf of Alaska (Fig. 3.1a), and interestingly this cyclone activity has steadily increased over the last 5 years. In the 2006-2007 cool season there was apparently a fairly active Pacific storm track, which led to 20 to more than 35 cyclone events per grid box in the Gulf of Alaska (Fig. 3.1e). Similarly, there were 20-30 cyclone events per grid box east of Greenland in 2002-2003 (Fig. 3.1a), but only 12-14 cyclone events per grid box the following year (Fig. 3.1b).

b. Analysis Error

Before utilizing a particular sea-level pressure analysis for model verification, the quality of the different model pressure analyses needed to be assessed. To date, there has not been a direct comparison between the GFS, NAM, and NARR pressure analyses for cyclones around North America. In order to compare each of these analyses, the closest cyclones separated by less than 800 km between the NARR, NAM, and GFS was identified every 6 hours. The central pressure and displacement differences were calculated between the NARR and the NAM cyclone and the corresponding GFS cyclone. It should be noted that over higher terrain, particularly the Rockies and West Coast, values of SLP are less accurate due to the approximation used to reduce the surface pressure to SLP (Pauley 1998). In addition, the NARR and NAM use a different method of SLP reduction than the GFS (Mesinger and Treadon, 1995), so SLP differences between these analyses over high terrain is less reliable than surrounding regions.

Figures 3.3a,b show the sea-level central pressure difference between the GFS and NARR and the GFS and NAM analyzed cyclones, in which positive values indicate locations where the NARR and NAM cyclone central pressures are greater than the GFS on average. Over the Pacific and Atlantic Oceans, the eastern Rockies, and all across Canada, the analyzed cyclones in the NARR are 1-4 mb weaker than the GFS (Fig. 3.3a). Over the eastern half of the U.S., the NARR analyzed cyclones are ~1 mb or less weaker than the GFS. Over the southwest U.S. and Pacific Northwest, the NARR cyclones are 0.5 to 2 mb deeper than the GFS cyclones. The pressure differences between the NAM

and GFS are similar to the NARR minus GFS (Fig. 3.3b). Over the oceans, the eastern Rockies, and Canada, the NAM cyclones are generally 0.5 to 2 mb weaker than GFS cyclones, while there is less than a 0.5 mb difference over the eastern half of the U.S. The NAM cyclones are 0.5 to 2 mb deeper than the GFS over the southwestern U.S. and Pacific Northwest.

In order to determine the accuracy of the sea-level pressure analyses, 37 surface National Weather Service stations were randomly selected across the U.S., Canada, Eastern Pacific, and Western Atlantic for the SLP verification (Fig. 3.3). For each station, a bilinear interpolation was applied to obtain the analyzed SLP, from each model, at that station for cyclones found within 500 km of the station. Figure 3.4a shows the mean absolute error of the analysis SLP for all stations from west to east, with the shading for the domain-averaged error (horizontal bars) indicating the 99% confidence intervals. The NARR and NAM errors (1.5-2.5 mb) are larger than the GFS (~1 mb) over the eastern Pacific, with the NARR and NAM errors decreasing toward the coast. The NARR and NAM have large SLP errors (1.5-2.5 mb) over the Rockies, and the GFS has about 0.5 mb smaller absolute SLP errors. Over the eastern half of the U.S., the SLP errors are 1 and 1.5 mb for the NAM and NARR, and just over 0.5 mb in the GFS, The eastern U.S. errors are lower than anywhere else in the domain, presumably because of the wealth of data upstream over the U.S. for data assimilation. Over the western Atlantic, the errors are 1.0-2.5 mb in all analyses, peaking at ~2.5 mb (2 mb) in the NARR and NAM (GFS). Across the full domain, the NARR and NAM have a mean absolute error of ~1.5 mb, while the GFS analysis has an error just below 1 mb. Overall, the error in the GFS is significantly (hereafter to mean at least statistically significant at the 90% level) less than the NARR and NAM.

Figure 3.4b shows the bias in SLP for each analysis. Except for station YRV in British Columbia, the NARR has a 0.5-2 mb positive SLP error at every station, with the largest positive bias over the eastern Pacific, western Atlantic, and the Rockies. The NAM has a similar distribution of pressures biases, except for a smaller positive bias over the Rockies. The GFS has a positive bias along the West Coast and the western Rockies, a slight negative bias over the eastern Rockies, and a near zero bias at other regions.

All analyses have a positive bias averaged across the domain. The NARR, NAM, and GFS have a bias of ~1 mb, 0.7 mb, and 0.2 mb, respectively. Therefore, since the GFS analysis has the smallest SLP analysis errors for cyclones as compared to the other models, the GFS was used to obtain the observed cyclone central pressures.

To see whether the model analyses have improved over the last several years, figure 3.5 shows the analysis pressure errors around cyclones for each cool season averaged across the domain. The NARR gradually improves in region 1 (eastern Pacific), 2 (western U.S.), and 3 (eastern Canada) from 2002 to 2006, with no significant change in regions 4-6 (Fig. 3.5a). The absolute SLP errors decrease from ~2.4 mb to ~1.6 mb over the eastern Pacific from 2002 to 2006, while from 2006 to 2007 the errors increase by 0.3-0.5 in all regions. The pressure bias around cyclones in the NARR remains positive for all regions, but fluctuates in magnitude over the 5 cool seasons, especially in regions 1-3 (Fig. 3.5b). The NAM has a slight decrease (increase) in absolute SLP errors in region 1 (2-6) from the 2002-03 cool season to the 2003-04 cool season (Fig. 3.5c). Afterwards, the NAM has little change in absolute SLP errors until the 2006-2007 cool season, when the errors decreased significantly (by 0.3-0.4 mb) since 2005-2006. The

bias in the NAM is fairly steady for regions 2-6, but the NAM bias over the Pacific decreases significantly from ~1.5 to ~0.7 mb over the 2005-2007 cool seasons (Fig. 3.5d). Meanwhile, the skill of the GFS SLP analyses has changed little (Fig. 3.5e), and there has been only a slight increase in the positive bias for regions 2-6 (Fig. 3.5f).

In order to determine whether a bias in cyclone position error contributed to the larger central pressure errors in the NAM and NARR, the SLP difference between the central pressure of analyzed cyclones was plotted only for cases in which NARR (NAM) was within 100 km of the corresponding GFS cyclones (Fig. 3.6). On average, NARR cyclones within 100 km of the same GFS cyclones are analyzed 2-3 mb weaker than the GFS cyclones over the Pacific, eastern Rockies, Canada, and the Atlantic, less than 1 mb weaker than the GFS cyclones over the eastern half of the U.S., and 1-2 mb deeper than the GFS cyclones over the western Rockies (Fig. 3.6a). The distribution of analysis differences for NARR cyclones within 100 km of the GFS cyclones is very similar to that of all cyclones (Fig. 3.3a). NAM cyclones within 100 km of the GFS cyclones are initialized ~1-2 mb weaker than the GFS cyclones over the Pacific, eastern Rockies, Canada, and the Atlantic, within 0.5 mb of the GFS cyclones over the eastern third of the U.S., and ~1-2 mb deeper than the GFS cyclones over the Upper Midwest and the Pacific Northwest (Fig. 3.6b). This distribution of errors closely matches the distribution for all NAM cyclones (Fig. 3.3b). This suggests that most of the larger analysis errors in the NAM/NARR are the result of positive central pressure biases (underdeepened cyclones), rather than large displacement errors.

Since it is impossible using scattered surface observations to determine whether the GFS or NAM had the best analyzed cyclone positions, the mean position of the GFS and NAM analysis cyclone was used as the observed cyclone position for verification. The NARR was not included in the average, since the NARR results are similar to the NAM (Fig. 3.4).

c. Short-term (0-60 h) GFS and NAM Verification

i. Spatial Distribution of Errors

In order to illustrate the spatial distribution of MAEs in cyclone central pressure, the pressure errors were interpolated to a 2.5° latitude-longitude grid and averaged over the 5-cool seasons for particular forecast hours (Fig. 3.7). The hatched boxes represent grid points in which the NAM error was significantly (90% level) different than the GFS error. Early in the forecast (18-36 h), the Pacific absolute errors in the GFS and NAM are 1-3 mb greater than the other regions. The NAM has larger errors (4-6 mb) than the GFS (2-4 mb) over the Pacific Ocean and Rockies (Figs. 3.7a,c), which is significant at the 90% level at many grid points. The pressure MAEs in the NAM are relatively small (1-3 mb) from the Southern Plains northeastward to the Great Lakes, mid-Atlantic, and Hudson Bay. Meanwhile, the MAEs in the GFS are 1-2 mb from the Midwest to the U.S. East Coast.

For the 42-60 h period (Figs. 3.7c,d), the MAEs in the GFS and NAM for all regions are 1-2 mb larger than the 18-36 h period. The NAM errors over the Pacific are 5-8 mb as compared to 4-6 mb in the GFS, although there are fewer cyclone matches for this later forecast period to yield statistical significance at many points. The NAM errors

also exceed the GFS in all regions except the Hudson Bay region. In particular, the NAM errors are 3-5 mb over the northern Plains and southern Canada, eastern Canada, the Northeast U.S., and over the western Atlantic, while the GFS errors are 2-3 mb in these areas.

Figure 3.8 shows the mean error (bias) in cyclone central pressure error. Early in the forecast (18-36 h), the NAM has a 2-4 mb positive bias over the eastern Pacific (cyclones too weak), while the GFS has little bias over this region. In contrast, over the steep terrain of the British Columbia coast and the Pacific Northwest, there is a 1-2 mb negative bias in the NAM (cyclones too deep), and a 1-4 mb negative bias in the GFS, which extends a few hundred kilometers offshore. This suggests that the forecast cyclones do not experience enough cyclolysis in the models as they approach the steep coastal terrain. Meanwhile, the NAM has significant positive errors over the Rockies and Midwest, which are 2-4 mb smaller than the GFS. From central Canada southeast across the Great Lakes, there is a 1-2 mb negative bias in the NAM, while in the GFS has a significant (2-5 mb) positive bias over the Hudson Bay and surrounding regions. The NAM also has a positive bias over the Northeast U.S. and Atlantic Ocean. The GFS has a negative bias along the East Coast, with no significant bias over the western Atlantic.

Later in the forecast period (42-60 h), the mean errors over the eastern Pacific, Pacific Northwest, and Rockies are nearly the same in both models as the earlier 18-36 h forecast period (Figs. 3.8c,d). However, the errors over eastern Canada in the NAM are more negative (by 2-3 mb) during the later period, while errors over the Atlantic are more positive (by 1-3 mb) in both models.

ii. Errors vs. Forecast Hour

1. Cyclone Central Pressure Errors

In order to more easily compare the evolution of forecast errors in each region, figure 3.9 shows the absolute errors of central pressure versus forecast hour averaged for each region. The Pacific (region 1) has 1-2 mb larger mean absolute errors than other regions for the NAM and GFS. The errors for this region increase rapidly to ~5.5 (~4) mb in the NAM (GFS) by F54. Meanwhile, the western Atlantic (region 6) errors in the NAM increase to ~4 mb by F60, and this is the second largest error of all regions. The GFS has a ~3 mb error over the western Atlantic, which is not statistically different than the other regions. The errors in region 4 (over the Plains and Mississippi Valley) in both models are significantly lower than in all other regions by F60. Overall, the NAM has larger mean absolute errors than the GFS, which is statistically significant except for region 2.

The NAM has a large (2-3 mb) positive bias over the Pacific (Fig. 3.9), which reaches ~3.5 mb by F54. The GFS has a small (~1 mb) positive bias, which is only significantly different than zero before F24. The GFS has a ~1 mb positive bias over central and eastern Canada (region 3), which is significant at the 90% level. In contrast, the NAM has a 1-1.5 mb negative bias after F30 in region 3, which is also statistically significant. There is a significant negative bias in the GFS and the NAM of about -0.6 and -0.8 mb, respectively, over the eastern U.S. (region 5). The GFS has significant negative errors (-1 mb) over the West Coast and Rockies (region 2).

It is unclear whether the central pressure errors arise from a few cases with large errors, or from many cases with moderate errors; thus, the distribution of cyclone errors was examined for each region and model. Figure 3.11 shows histograms of cyclone SLP errors for the NAM and GFS for each region. The red lines indicate the mean cyclone SLP error and 1 standard deviation above and below the mean. The error distributions are clearly wider for region 1, especially for the NAM. For example, there are several cyclone events with NAM errors greater than 15 mb in region 1, which suggests there are severe predictability problems over the Pacific at times. In contrast, there are few cyclone events with errors greater than 12 mb for regions 3-6. Overall, both models have a nearly normal distribution of cyclone errors in all regions. There is a tendency for the positive tail of the distribution to be slightly larger than the negative tail, indicating that when a large error event occurs, it tends to involve an underforecast cyclone. The fact that most of these error distributions are nearly normal suggests that the mean SLP errors result from persistent biases and not just a select number of very large error events.

Figure 3.12 shows the annual variation of cyclone central pressure errors at F48 for the NAM and GFS for each region. For the NAM, only region 1 has a statistically significant trend of decreasing errors of ~0.5 mb per year, which is likely the result of improvements in data assimilation in this data sparse region (Fig. 3.5). The GFS also has a decreasing error trend in region 1 (since the 2003-2004 cool season) of ~0.5 mb per year. In the NAM, region 4 has a slight increase in errors over the 5-year period, while regions 2, 3, 5, and 6 have no significant error trend . The GFS errors in region 2 have decreased by about 1 mb since the 2003-2004 cool season. Region 3 has a slight decrease in error for the GFS, while regions 4, 5, and 6 (eastern U.S. and western Atlantic) have no change in errors in the GFS. Interestingly, the 2003-2004 cool season has the largest errors in region 1 for both models, while the 2006-2007 cool season has the smallest errors in region 1.

There are also interesting annual trends in the bias of cyclone central pressures (Fig. 3.12). Central pressure errors in the GFS have changed little over the period, except for region 1, where cyclones are too deep in 2004-2007. However, the positive bias in the NAM central pressures errors have decreased in all regions, particularly in the last 2 cool seasons (2005-2007). This trend is very evident in region 1, where errors have decreased from almost 5 mb in 2002-2003 to nearly 0 mb in 2006-2007.

It is important to know whether using the NAM analysis as truth will change the results. Figure 3.10 shows cyclone central pressure errors versus forecast hour using the NAM as truth. The NAM has smaller cyclone SLP absolute errors to hour 12, since the NAM forecast is still relatively close to its initial state. From F12 to F60, the NAM errors are larger than the GFS for all regions, with Pacific (region 1) MAEs in the NAM peaking at just over 5 mb. This is consistent with the above results using GFS as truth (Fig. 3.9). The mean error does decrease in a few regions by ~1 mb using the NAM as truth, but this is a result of the positive bias in the NAM analyses as shown in section III.b.

2. Cyclone Displacement Errors

The average cyclone displacement errors were quantified for each region (Fig. 3.9). In the GFS, region 2 (western U.S.) has the largest cyclone displacement errors that

are statistically significant for nearly the entire forecast of 60 hours, and the errors peak at about 290 km at F60. Meanwhile, the largest displacement errors in the NAM are in region 2, exceeding 250 km by F42 when region 1 (eastern Pacific) errors become more comparable. The region 1 displacement errors in the GFS peak at about 260 km at F060, and are significantly larger than regions 3-6 by F48. The other regions have more similar errors with each other, peaking at 270-290 km (240-260 km) in the NAM (GFS), and all regions in the NAM have larger displacement errors than the same regions in the GFS (statistically significant except for region 3).

Cyclone displacement errors have also varied substantially over the 2002-2007 cool seasons (Fig. 3.12). For the NAM, region 6 has a significant increase in cyclone displacement of ~25 km per year since the 2003-2004 cool season. Region 4 also has a notible increase in the displacement errors over the period. Meanwhile, regions 1, 2, 3, and 5 do not have an increasing error trend over the 5-year period, but regions 1 and 5 have the largest displacement errors in the 2006-2007 cool season.

In the GFS, the most distinct trend in displacement errors was for region 3 (Fig. 3.12), where the displacement errors have decreased 15-20 km per year on average. There is also a decrease in the errors for region 1 of 10-15 km per year since the 2003-2004 cool season. Regions 2, 4, and 6 have a slight decreasing trend of displacement errors, while region 5 does not have a significant trend.

Figure 3.13 shows the histogram distributions of cyclone displacement errors. There are no well-defined large tails in the errors, with the mode in most regions of 100-200 km. The largest tail of larger errors are found in region 2 in both models, indicating the difficulty in both forecasting the exact cyclone position over the high terrain, as well as tracking and SLP reduction difficulties over terrain. Overall, these error distributions suggest that the mean displacement errors discussed above are a result of many cases with similar errors, and not from a large set of anomalous forecast bust cases.

3. Direction of Displacement Error

It is also important to identify any cyclone position biases in operational models. Figure 3.14 shows position histograms for the NAM and GFS for the 18-36h and 42-60h forecast periods for regions 1, 4, and 6. For F18-F36, the cyclones for region 1 in the GFS and NAM are generally displaced to the right (east) of the observed. For the NAM, this displacement is significant at the 95% level[1] to the south, and significant east and southeast in the GFS. For the F42-60 period, the majority of region 1 cyclones are displaced in directions other than to the west of the observed, with the southward displacement (ME ~40 km) having a statistical significance. In region 4, cyclones in the NAM (GFS) are significantly displaced to the north by 30-40 km (40-50 km) at all forecast hours. The NAM in region 6 has a significant westward bias during the entire forecast. The GFS has less of a bias, though there is still a significant west and southwest displacement.

For the other regions, region 2 cyclones are significantly displaced (ME 40-50 km) to the northwest in both models at all forecast hours (not shown). In region 3 at F18-

[1]For a given direction of displacement to have statistical significance, the number of cyclones falling within that directional bin is significantly different at the 95% level than the average number of cyclones in each bin.

36 (not shown), most cyclones are displaced from the southwest, west, to the north sectors in the NAM (significant to the southwest at 40-50 km), and south and southwest in the GFS (both significant). Most cyclones are displaced similarly at F42-F60, with southwest being the only significant direction. Region 5 cyclone displacements (not shown) are very similar to region 6 displacements.

iii. Intraseasonal Cyclone Errors

There are also important variations in cyclone errors within each cool season as illustrated using a time series of cyclone central pressure errors at forecast hour 48 (F48) during the 5 cool seasons. In the eastern Pacific (region 1) (Fig. 3.17) both the NAM and GFS cyclone SLP errors vary greatly over the 5 cool seasons. There are some interesting periods to be noted. For example, nearly the entire first 2 cool seasons (2002-03 and 2003-04) are characterized by underforecast cyclones in the NAM over the eastern Pacific (Fig. 3.17a, box 1), which is not the case for the GFS (Fig. 3.17b). Another period of underforecast cyclones in the NAM extends from the end of January 2005 to the end of March 2005 (Fig. 3.17a, box 2). This period matches a period of overforecast GFS cyclones during most of January 2005 (Fig. 3.17, box 1), and then a short period of underforecast GFS cyclones (Fig. 3.17b, box 2). Finally, during most of November and December 2006, the GFS overforecasts the strength of many cyclone events (Fig. 3.17, box 3). Finally, this time series suggests that it is just as likely to get a large cyclone event in the past year or two as back in 2002-2003, so there is still room for improvement in operational cyclone forecasts.

In the central U.S. (region 4) there are also some periods that have some systematic patterns. For example, in February 2005, the NAM underforcasts almost all cyclone events, yet during the next month the NAM overforecasts most cyclone events (Fig. 3.18a, box 1). In 2006, the reverse occurs, with the NAM overforecasting nearly all cyclone events for October and November, but significantly underforecasting cyclone events in December 2006 (Fig. 3.18a, box 2). The GFS appears to have fewer of these error oscillations, but it does show a few periods of consistent underforecasting followed by overforecasting, such as January 2004 and March 2005 (Fig. 3.18, box 1 and 2).

In the Atlantic (region 6), for the same 1-month period in late January and early February 2003, both the NAM and GFS significantly underforecast a majority of cyclones (Figs. 3.19a,b, box 1). In the second half of the 2003-2004 cool season, the NAM showed a cycle from overforecast cyclones to underforecasted cyclones (Fig. 3.19b, box 2). In the second half of the 2004-2005 cool season, the GFS shifted from underforecasting many cyclones to overforecasting most cyclones (Fig. 3.19b, box 2). Future work outside of this thesis will need to take a closer look at the reasons for these error transitions.

iv. Comparison with Previous Studies

To compare these results to cyclone errors in prvious studies, Fig. 3.27 compares cyclone displacement errors for various models from two previous cyclone verification studies to the cyclone displacement errors shown above. The LFM-II displacement errors were very high for the 1978/79 cool season, ranging from ~300 km to ~440 km from hour

24 to 48. By the late 1980s, cyclone position forecasts over the Atlantic improved by ~30% over cyclone forecasts over the U.S. and oceans in the late 1970s. In the 1987/88 and 1989/90 cool seasons, the AVN had smaller displacement errors than the NGM (~190km to ~300 km versus ~220 km to ~350 km). The displacement errors in the NAM and GFS for the 2002-2007 cool seasons are smaller than all previous studies (significant for the GFS). This comparison shows that cyclone position forecasts have improved since earlier studies.

d. 0600 and 1800 UTC Model Run Cycles

Forecasters are interested in whether the 0600 and 1800 UTC model cycles add any skill to the previous 0000 and 1200 UTC runs, respectively. Although more recent observations are available in the 0600 (1800) UTC cycle as compared to the 0000 (1200) UTC run, there are less upper air observations (rawinsondes) available (Zapotocny, 2005). In order to address this question, the errors for all cyclones in the 00 (12) UTC simulations were compared with the same cyclones in the 06 (18) UTC runs. The two time series of errors for the 0000/1200 UTC (solid) versus 06/18 (dashed) UTC were also matched at the same verifying times in Fig. 3.20

For the eastern Pacific and the central U.S. (regions 1 and 4) in the NAM, there is a statistically significant (90% level) skill improvement for cyclone central pressure of at least 12 hours for the 0600 (1800) UTC forecast runs as compared to the 0000 (1200) UTC (Fig. 3.20). For the western Atlantic (region 6), there is only a significant improvement by the 0600 and 1800 UTC runs after F30. The 1800 UTC run adds more skill to the 1200 UTC run than the 0600 UTC compared to the 0000 UTC. For all regions in the GFS, there is a statistically significant skill improvement of 6-12 hours by the 0600 (1800) UTC runs as compared to the 0000 (1200) UTC runs. Like the NAM, there is a greater improvement from the 1800 UTC run in the GFS than for the 0600 UTC run.

For cyclone position in the NAM (Fig. 3.20), all regions show an improvement in skill with the 0600 and 1800 UTC runs when compared to the 0000 and 1200 UTC runs. The improvement (6-12 hours) is statistically significant at the 90% level, except before F36 over the western Atlantic (region 6) in the 0000 versus 0600 UTC runs. For the GFS, all regions indicate an improvement for the 0600 and 1800 UTC runs as compared to the 0000 and 1200 UTC cycles. These GFS improvements of 6-12 hours are statistically significant, except only before F24 in the Atlantic for the 0000 versus 0600 UTC cycles.

These results indicate that although there is less rawinsonde data available for the 0600 and 1800 UTC runs, both models still show a statistically significant improvement over the preceding 0000 and 1200 UTC run cycles.

e. Verification of Deep Cyclones

Cyclone verification was also completed for relatively deep cyclone events, in which the central pressure (forecast *or* observed) was more than 1.5 standard deviations below the mean cyclone central pressure for each region (Table 3.1). Figure 3.21 shows the central pressure MAEs for these deeper cyclone events in each region. The largest errors for central pressure are over the Pacific (region 1) in both the NAM and the GFS, where the NAM (GFS) errors exceed 7 (6.5) mb by F60. In the GFS, the western U.S.

(region 2) has the second largest errors, peaking at ~4.5 mb by F48. The Atlantic (region 6) also has fairly large errors for both the NAM (7 mb) and GFS (4.5 mb). The errors in the other regions are between 3-4.5 mb (2.5-4 mb) in the NAM (GFS). Overall, except for eastern Canada (region 3), where GFS errors are slightly larger than NAM errors, the NAM errors are generally 1-3 mb greater than the GFS in each region.

The central pressure bias (mean error) also varies by region for these deeper cyclone events (Fig. 3.21). For the NAM, all regions except region 3 have an increasing positive bias after F42. For example, the NAM has a large positive bias in region 1 of ~5 mb by F54. By F48, regions 3-6 have a positive bias in the GFS. Region 6 has the largest error growth of any region after F36, with mean errors of ~5 mb by F60. Except for region 3, by F48 the NAM has a larger (positive) bias in each region than the GFS.

For the NAM, the region 3 displacement errors are significantly less than other regions after F30, with the region 3 errors increasing from ~120 km at F30 to ~240 km by F60 (Fig. 3.21). The eastern U.S. and western Atlantic (regions 5 and 6) have significantly smaller displacement errors (~250 km by F60) in the NAM than from the eastern Pacific to the Midwest (regions 1, 2, and 4) (~320 km by F60). For the GFS, region 2 has the largest cyclone displacement errors, ranging from ~220 km at F36 to ~300 km by F60. Region 1 has the second largest displacement errors, peaking at ~250 km by F60. All other regions have similar errors by F60, with errors 200-220 km.

For region 1 the deep cyclones are displaced to the east through south on average in the NAM as compared to the analyzed (observed) positions at F18-36 (Fig. 3.22), which is significant at the 90% level. By F42-60, the displacement bias is more to the south and southeast, and it is still significant with a mean displacement vector of ~50 km. The GFS displaces these deeper cyclones to the northeast through southeast (significant at 90%), with a mean displacement vector of ~30 km at F18-36, The GFS error is east through south by F42-60, with a mean displacement vector of ~15 km (not significant).

In region 4 in the NAM, most of the deep cyclones are displaced to the southwest at F18-36, while the majority of deep cyclones are displaced to the northeast by F42-F60. For the GFS in this region, deep cyclones are significantly displaced to the northeast to east over the 60-hour forecast, with a mean displacement vector of 40 (90) km at F18-36 (F42-60). The NAM has a significant displacement to the west and southwest in region 6 throughout the forecast period. Meanwhile, for the GFS, the cyclones are displaced to the southwest at F18-36, and to the south and southwest by F42-60, with a mean displacement vector of ~40 km.

When compared to all cyclones, absolute SLP errors in both models are 30-60% higher in all regions when forecasting deep cyclones. For the NAM, central pressure errors are shifted in the positive direction for deep cyclones when compared to all cyclones. Except for region 3, all regions have positive errors at nearly all forecast hours. For the GFS, central pressure errors in all regions, except for region 2, are shifted in the positive direction. However, errors in region 2 actually become more negative. For the NAM, deep cyclones have larger displacement than all cyclones for the central U.S., and smaller displacement for central Canada, the eastern U.S. and the Atlantic. For the GFS, deep cyclones are displaced less in all regions than all cyclones.

f. Extended GFS Verification

In order to understand how cyclone forecast errors evolve in the medium range forecast period (days 3-5), the GFS was also verified to 120 h. This analysis included the Central Atlantic (CA) and Central Pacific (CP) regions (cf. Fig. 2.3). Figures 3.23a,b show the spatial distribution of central pressure absolute errors. Cyclone central pressure absolute errors at F72-90 are generally 5-8 mb over the CA and western Canada, and 6-9 mb over the CP and Alaska. Some large errors (3-6 mb) are also found near Hudson Bay, extending southward to the interior eastern one-third of the U.S. (Fig. 3.23a). At F102-120 (Fig. 3.23b), the errors over the oceans increase by ~30-40%, particularly over the CA, where the errors become comparable to the CP errors (8-10 mb) by this hour. There is also a 20-30% increase in the errors across eastern Canada between F72-F90 and F102-F120.

At F72-90, there are large negative mean errors over Alaska, western Canada, and the Gulf of Alaska (-2 to –6 mb) in the GFS, while there are large positive errors (2-6 mb) over the western Atlantic (region 6) and Eastern Canada (region 3) (Fig. 3.23c). At F102-120, the spatial distribution of central pressure bias is very similar to that at F72-90.

The errors were also combined for each region (Fig. 3.24). The CA (CP) was not included because results here were very similar to the western Atlantic (eastern Pacific). Region 2 was also not included, since problems with SLP reduction made results here unreliable. Through about F090, the western Pacific (region 1) has the largest absolute errors of ~5.5 mb (statistically significant) in cyclone central pressure at ~F054 (Fig. 3.24a). However, after F072, the central pressure error growth in the western Atlantic (region 6) increases, making the Atlantic errors comparable with the Pacific errors by F96. In contrast, region 4 has the smallest absolute errors through the period, which peak at ~4 mb by F120.

Figure 3.24b shows the central pressure mean errors. After F24, the eastern Pacific (region 1) has a negative bias until F108, peaking at about -0.75 mb at F066-F084. All other regions have a positive bias after about F054 of 0.5-1 mb (statistically different than zero after F84). The western Atlantic (region 6) has the largest positive bias by F120 (~1.5 mb).

The cyclone displacement errors increase by 50-70% during the F60-F120 extended GFS forecast (Fig. 3.24c). Region 4 has the smallest displacement errors of ~370 km by F120. All other regions have similar errors throughout the forecast period.

Cyclone errors versus forecast hour for deep cyclones (central pressure [forecast or observed] more than 1.5 standard deviations below the mean) are shown in figure 3.25. The Pacific (region 1) has the largest MAEs through F102, but the western Atlantic (region 6) errors are greater than the Pacific by F108. The central U.S. (region 4) has the smallest MAEs of ~6 mb, which is significantly smaller than the other regions. The MAEs for regions 1, 3, and 5 are clustered around 10-11 mb by F120.

The GFS has an increasing negative bias in the Pacific (region 1) through F108 (~ -2.2 mb) (Fig. 3.25b). The bias over the Atlantic is near 0 through most of the forecast period, until F120 when the bias peaks at more than 2.5 mb. Regions 3-5 have large positive biases after F48, especially the Eastern U.S., which peaks at ~5.5 mb by F120.

Figure 3.25c shows the displacement errors for the deep cyclones. By F108 the largest displacement errors of > 450 km are over the western Atlantic (region 6) and the

errors are statistically significantly greater than the other regions after F102. The smallest displacement errors (~375 km) are over the central U.S. (region 4). Displacement errors are fairly evenly distributed across all regions by F120, with the errors ranging from 375 km to 475 km.

From personal experience and discussion with forecasters, medium range (F072-F120) forecast cyclones over the Eastern U.S. and western Atlantic are known to 'drift' westward with each successive run of the GFS for large East Coast and western Atlantic cyclones events. Also, GFS cyclones tend to be forecast too slow as they travel up the coast. To quantify these assessments, figure 3.26 shows angle histograms (like those shown in III.c.ii.3) of GFS cyclone displacement, over the western Atlantic, in order of decreasing lead-time (a-f). At long lead times (> hour 90, Fig. 3.26a,b), the majority of western Atlantic cyclones are forecast to the east of the observed cyclone (significant), with a mean displacement of about 40 km. As lead-time decreases (Fig. 3.26c, d) the western bias disappears, and cyclones then tend to be forecast too far to the south, by 40-70 km. For short lead times (Fig. 3.26e, f) cyclones are generally forecast too far southwest, suggesting that cyclones moving northeast along the coast are forecast too slow. These results confirm qualitative assessments of forecasters, and suggest that there are systematic errors in the GFS whose source may be detectable.

Region	*1*	*2*	*3*	*4*	*5*	*6*
SLP	977.3	989.4	982.6	995.3	986.4	989.6

Table 3.1. Sea level pressure thresholds (in mb) for the "deep cyclones" For each region. Thresholds were calculated for each region as the mean central pressure of all cyclones minus the 1.5 standard deviation value.

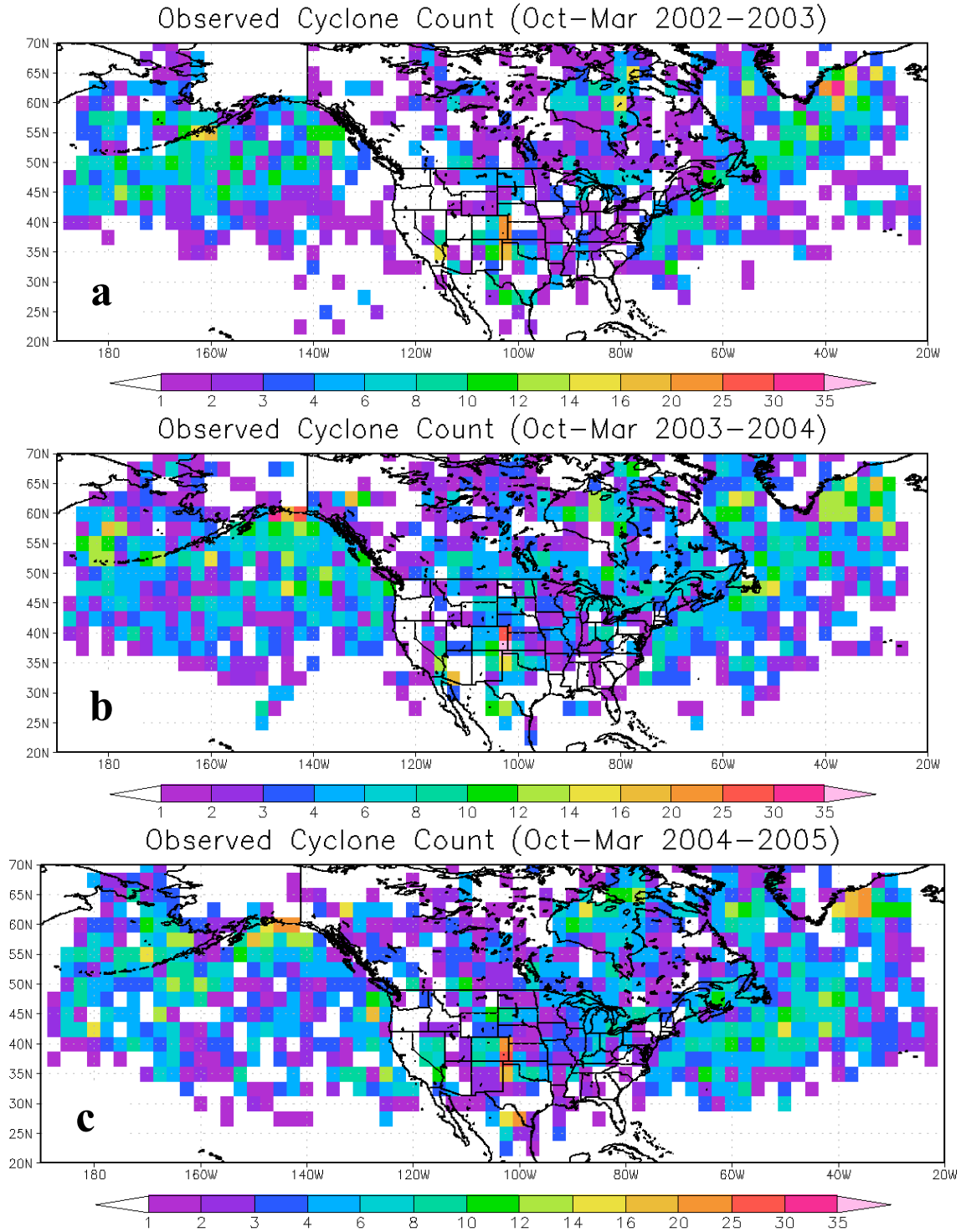

Figure 3.1. Observed cyclones per 2.5° grid box during the (a) 2002-2003, (b) 2003-2004, and (c) 2004-2005 cool seasons.

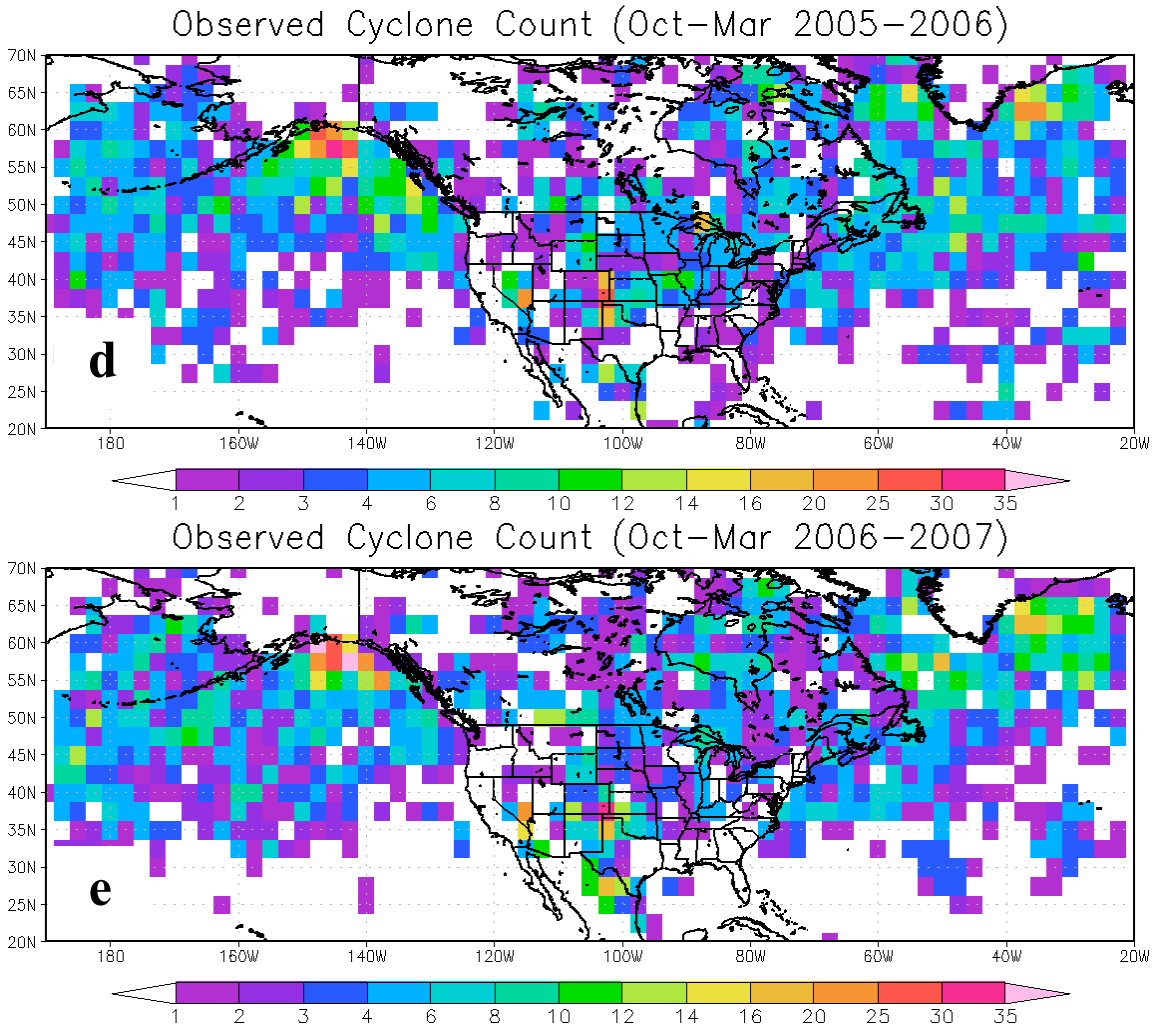

Figure 3.1 continued. Observed cyclones per 2.5° grid box during the cool season (d) 2005-2006 and (e) 2006-2007 cool seasons.

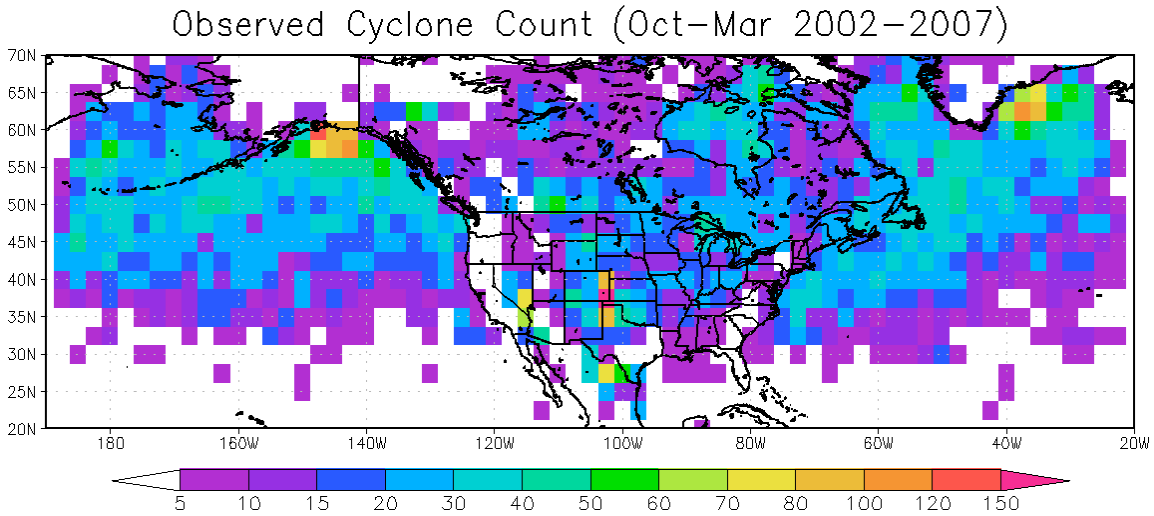

Figure 3.2. Total observed cyclones per 2.5° grid box during the cool seasons (2002 to 2007).

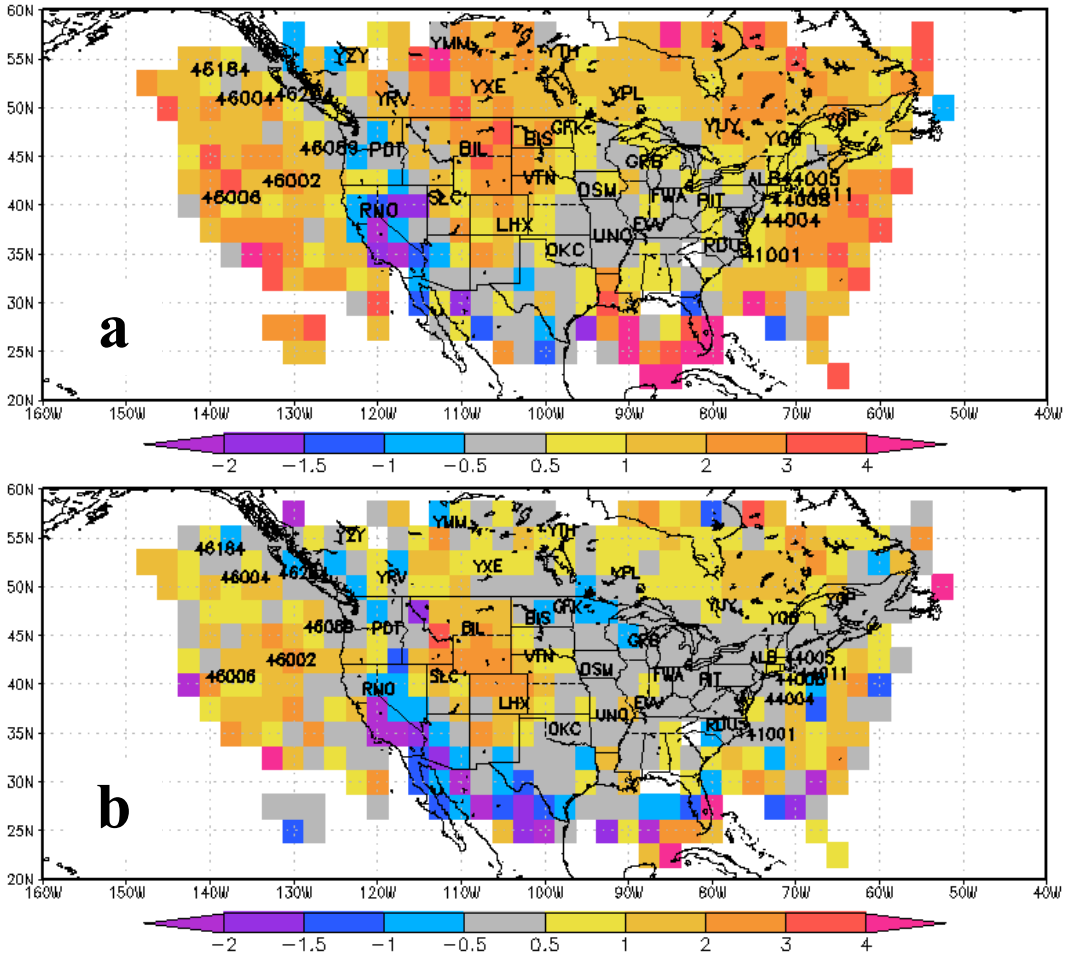

Figure 3.3. Difference in cyclone central pressure (color filled in mb) between the (a) NARR – GFS and (b) NAM – GFS. The surface stations used to verify pressure analyses are also shown.

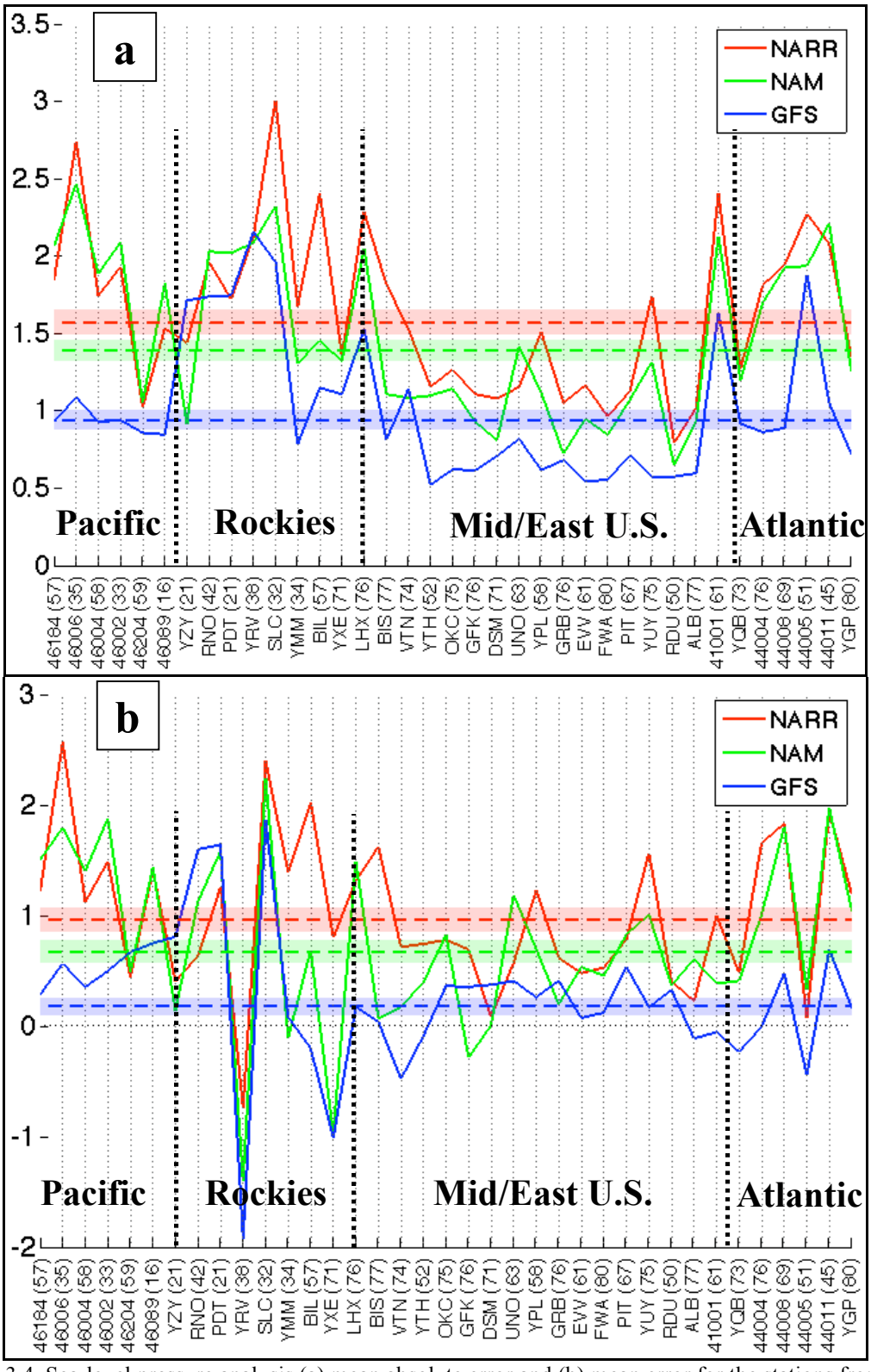

Figure 3.4. Sea-level pressure analysis (a) mean absolute error and (b) mean error for the stations from west to east in Fig. 3.3. The number of cyclones verified are shown in the parentheses..

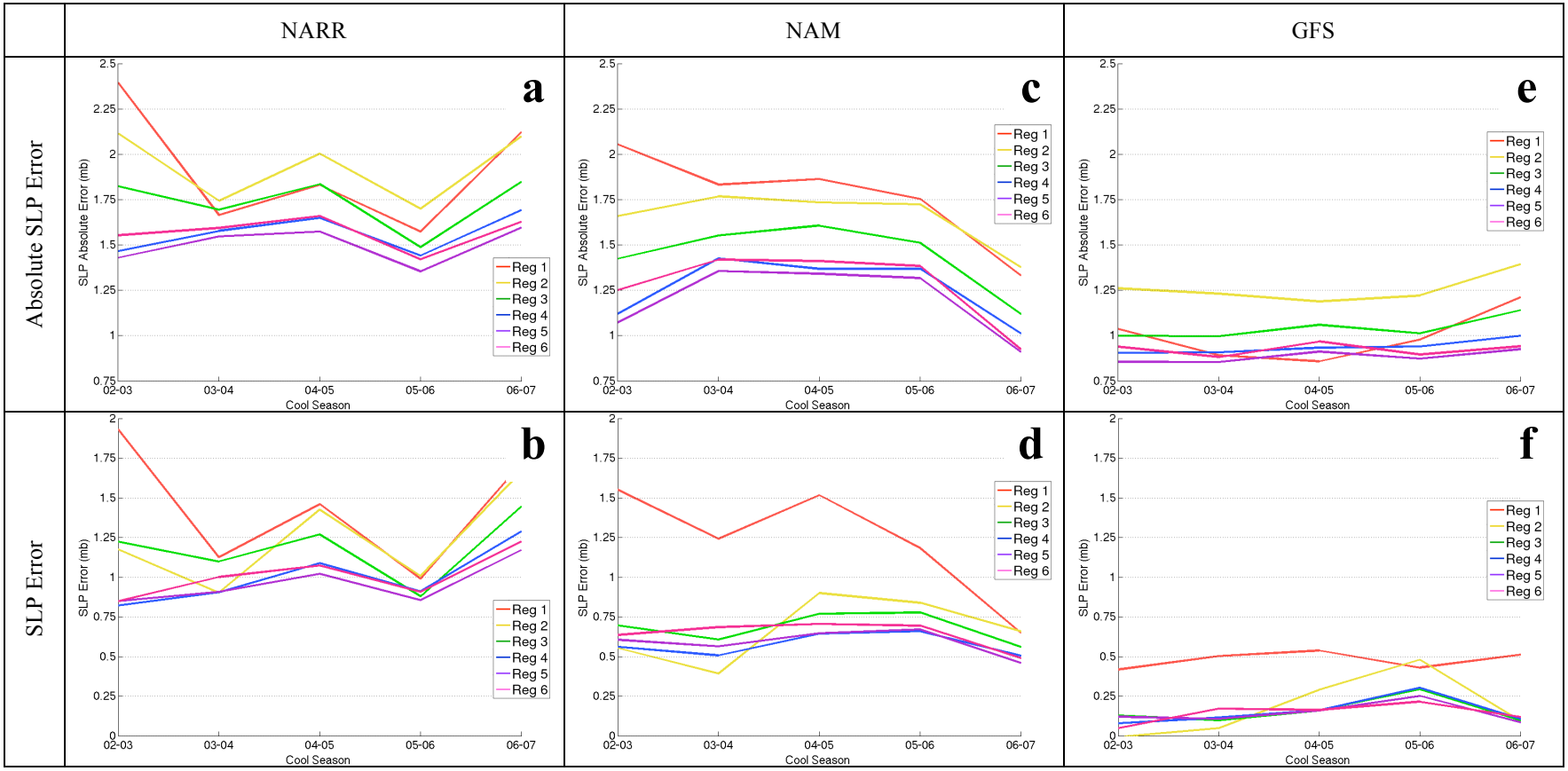

Figure 3.5. Analysis error of sea-level pressure (in mb) for the each of the 2002-2007 cool seasons and region for the NARR, NAM, and GFS.

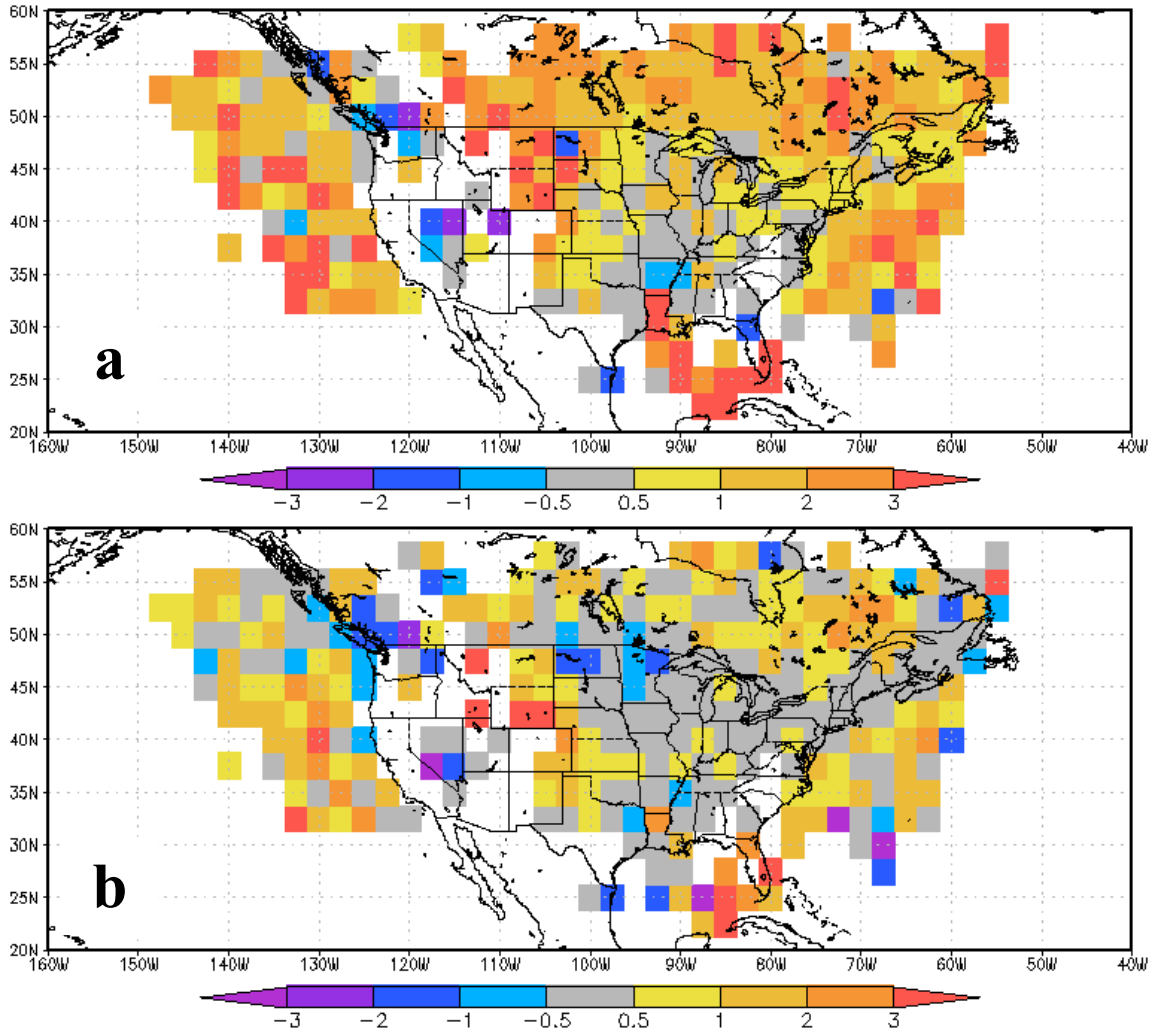

Figure 3.6. Same as Fig. 3.3 except for the (a) NARR and (b) NAM cyclones within 100 km of the corresponding GFS cyclone.

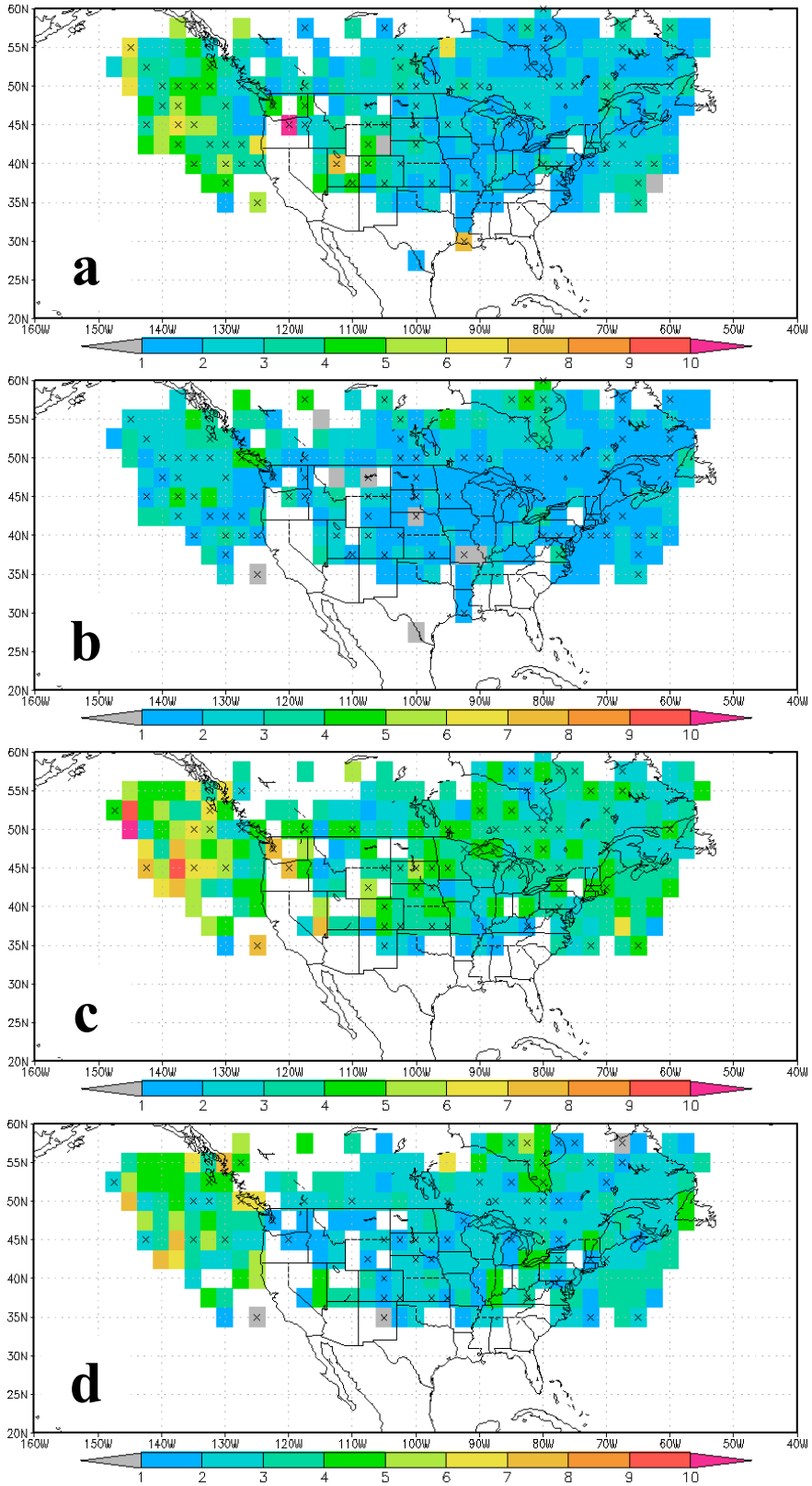

Figure 3.7. Spatial distribution of mean absolute errors of cyclone central pressures (filled in mb) for the (a) NAM and (b) GFS for 18-36 h. (c) and (d) Same as (a) and (b) except for 42-60 h. The X locations mark points where the GFS pressure is significantly different than the NAM at the 90 % level.

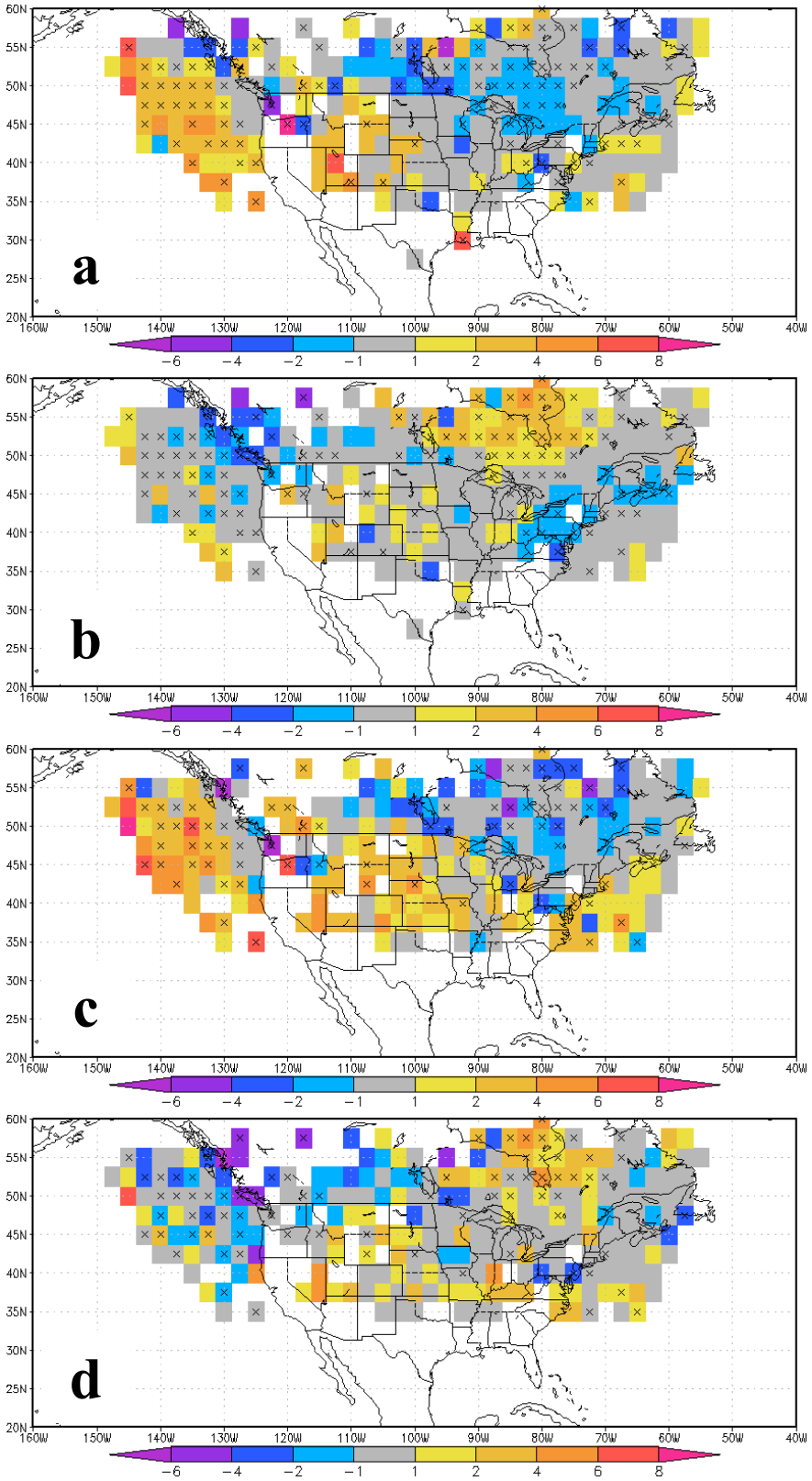

Figure 3.8. Same as in Fig. 3.7 except for mean error (bias) of cyclone central pressure (filled in mb).

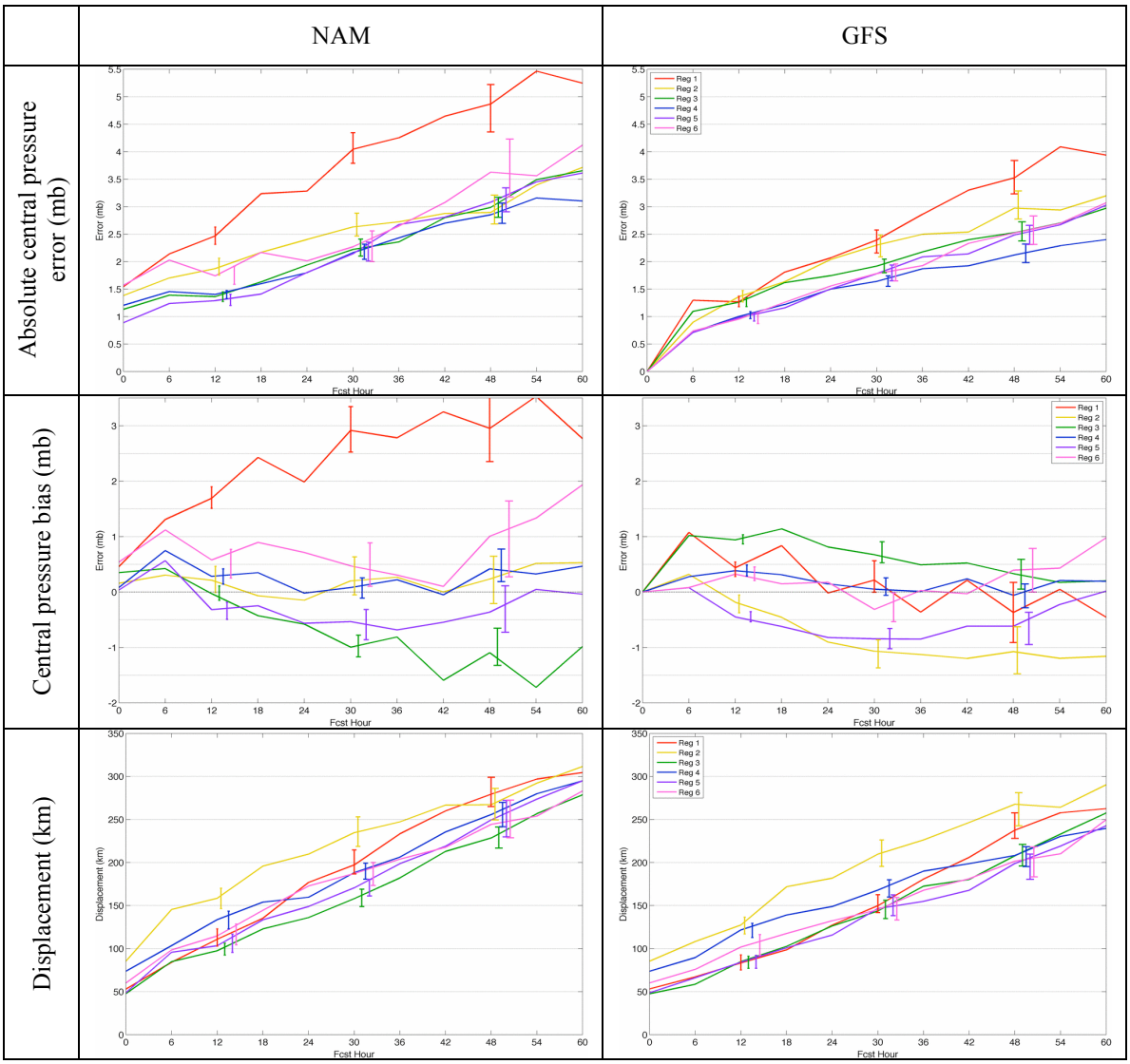

Figure 3.9. Cyclone mean absolute and mean errors of cyclone central pressure as well as displacement errors versus forecast hour for the 2002-2007 cool seasons for the NAM and GFS using regions 1-6 specified in Fig. 2.3. Confidence intervals at the 90% significance level are given by the vertical bars.

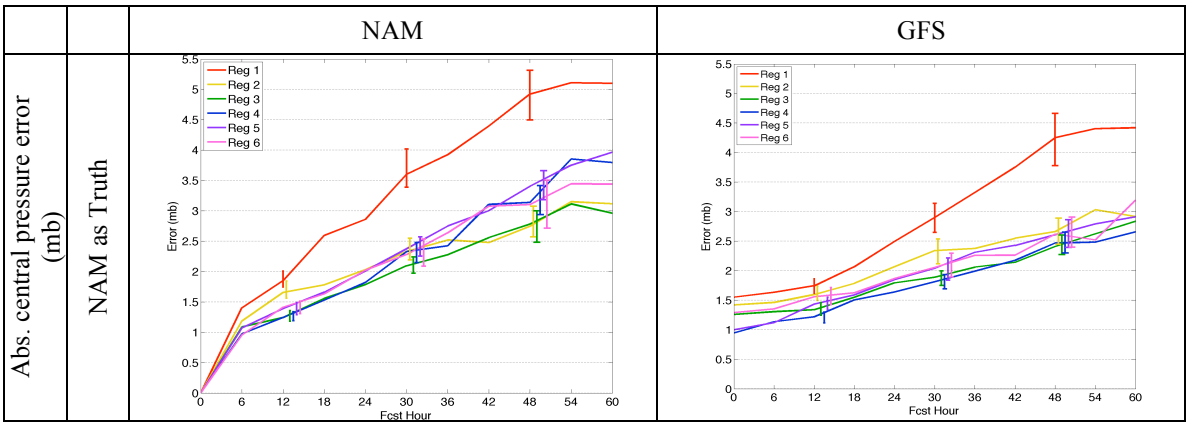

Figure 3.10. Cyclone mean absolute errors of cyclones central pressure versus forecast hour for the 2002-2007 cool seasons for the NAM and GFS using regions 1-6 specified in Fig. 2.3, using the NAM as truth.

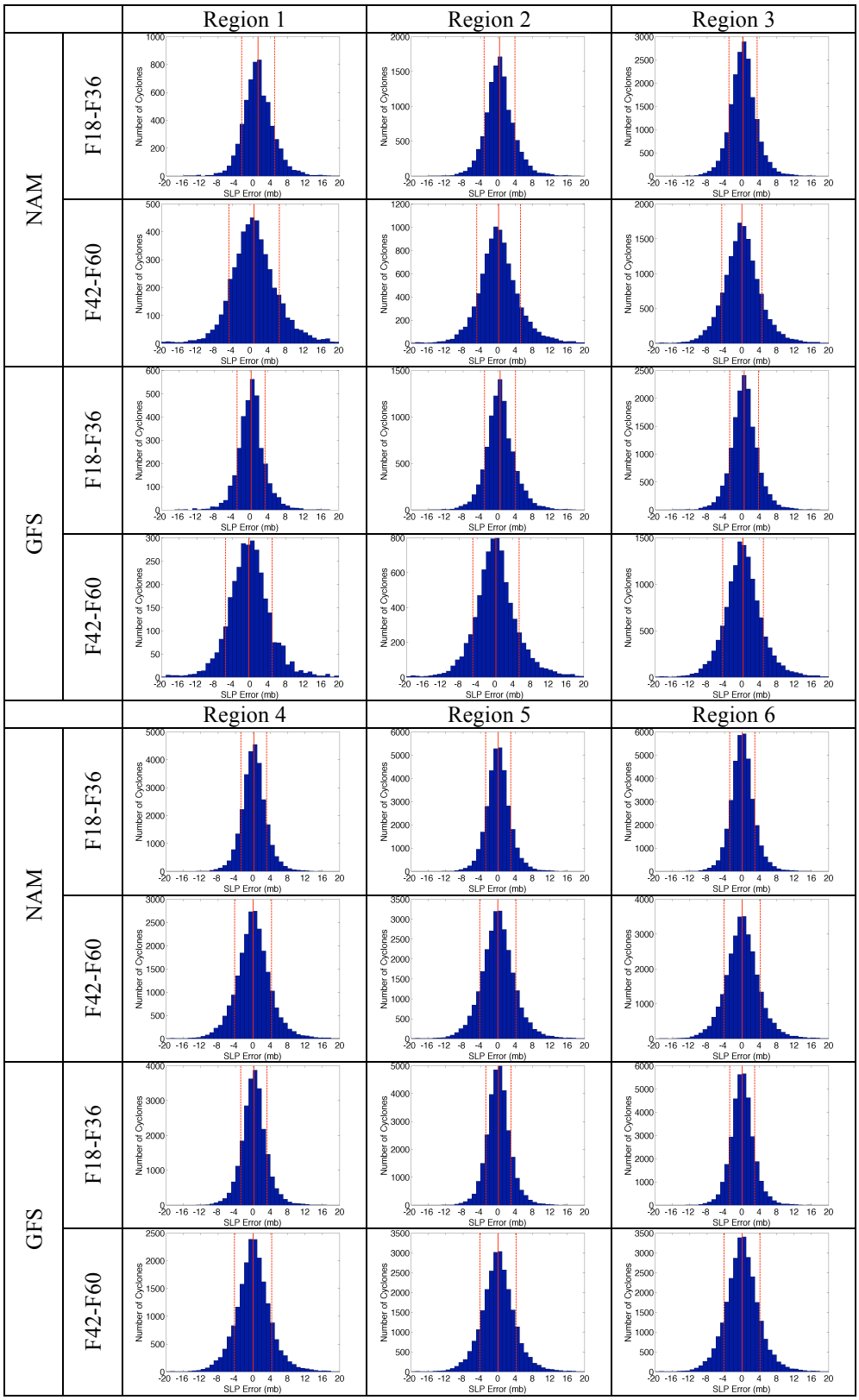

Figure 3.11. Histograms of cyclone central pressure errors for regions 1-6 and forecast periods 18-36h and 42-60h in the GFS and NAM.

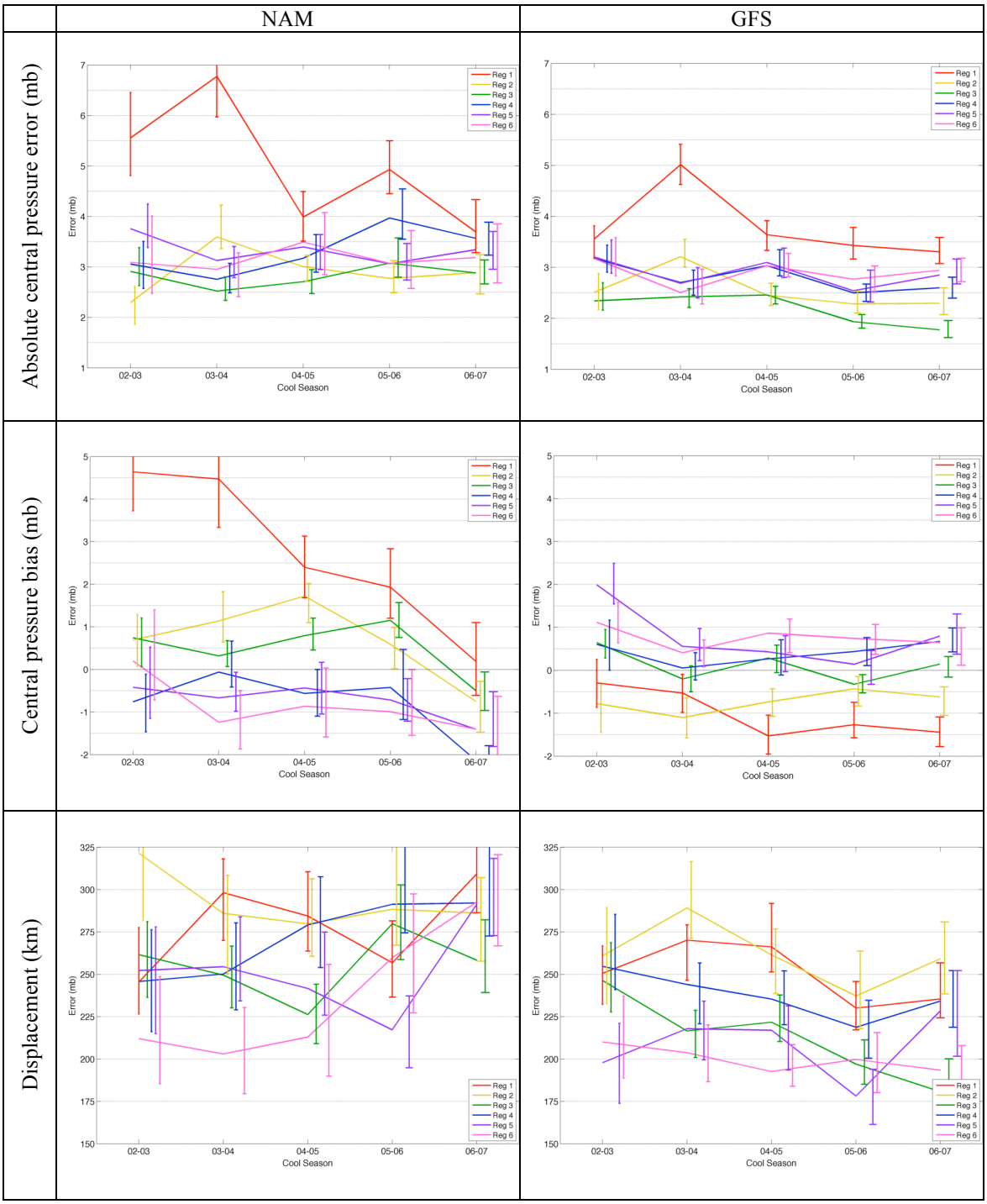

Figure 3.12. Cyclone mean absolute errors (top panels), mean errors (middle panels), and displacement errors in km (bottom panels) at hour 48 for each cool season for the NAM (left) and GFS (right).

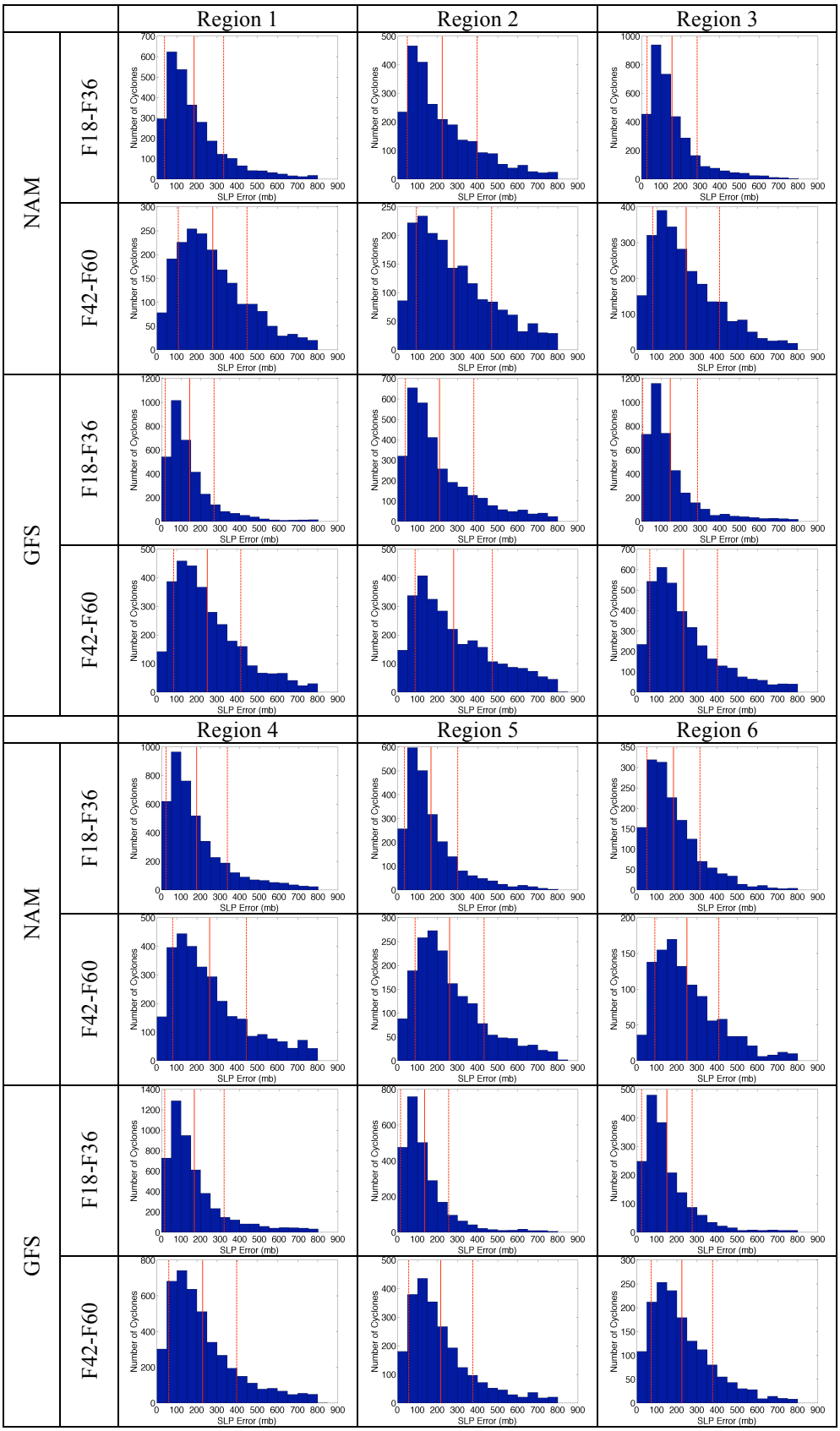

Figure 3.13. Same as figure 3.11 except for cyclone displacement errors.

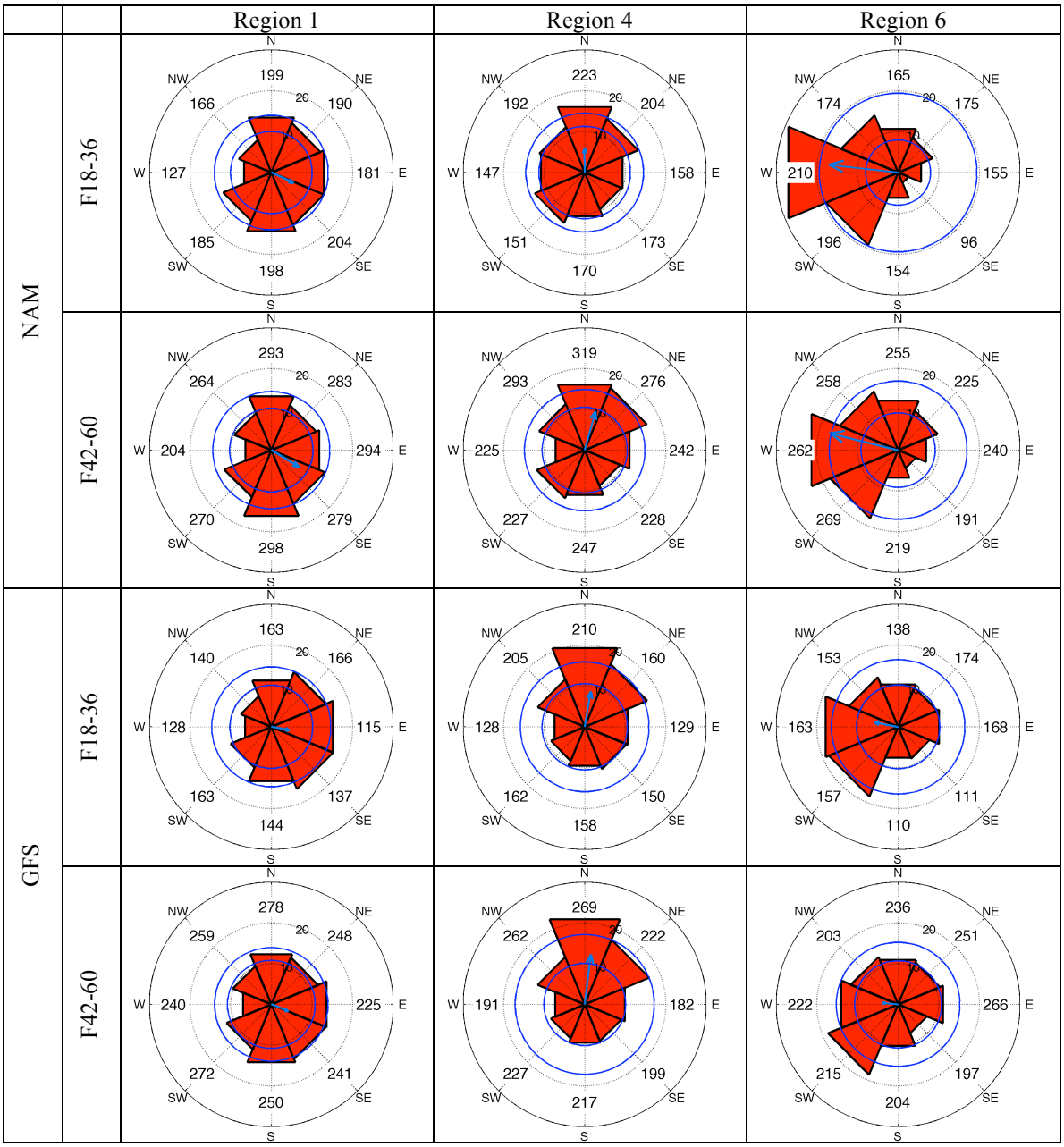

Figure 3.14. Histograms showing the frequency of model cyclone positions relative to the observed position (center point) for 45° bins centered on N, NE, E, etc… in the NAM and GFS for hours 18-36 h and 42-60 h for regions 1,4, and 6. The black range circles are every 10%, from 0 to 30. The blue range circles represent the 90% confidence intervals below and above the range of the bins. The numbers for each radial represent the average displacement error (in km) within that directional bin. The blue vectors indicate the mean displacement vector, with each gray range ring every 50 km.

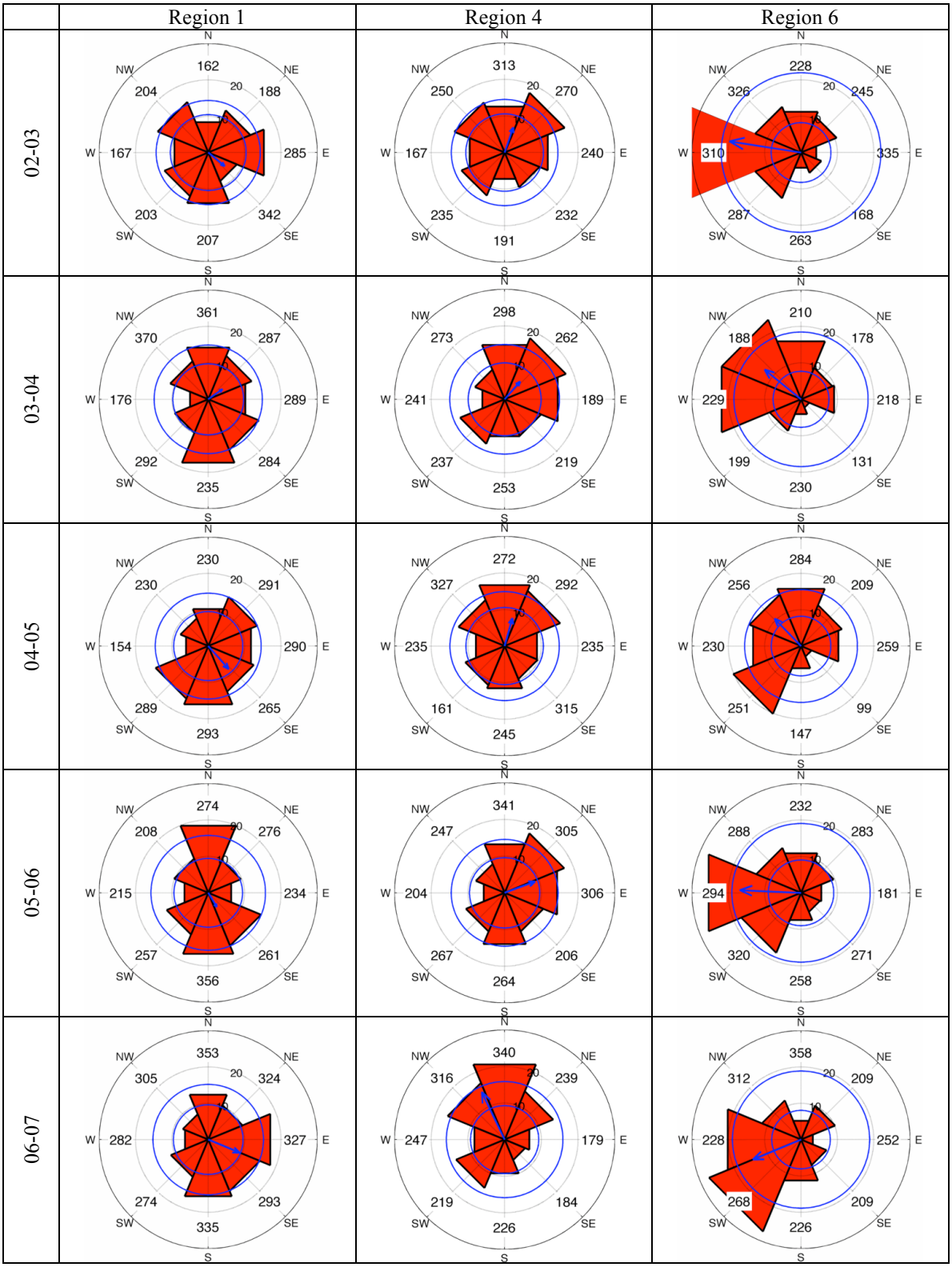

Figure 3.15. Same as Fig. 3.14, except only for the NAM at F42-60 for each 2002-2007 cool season.

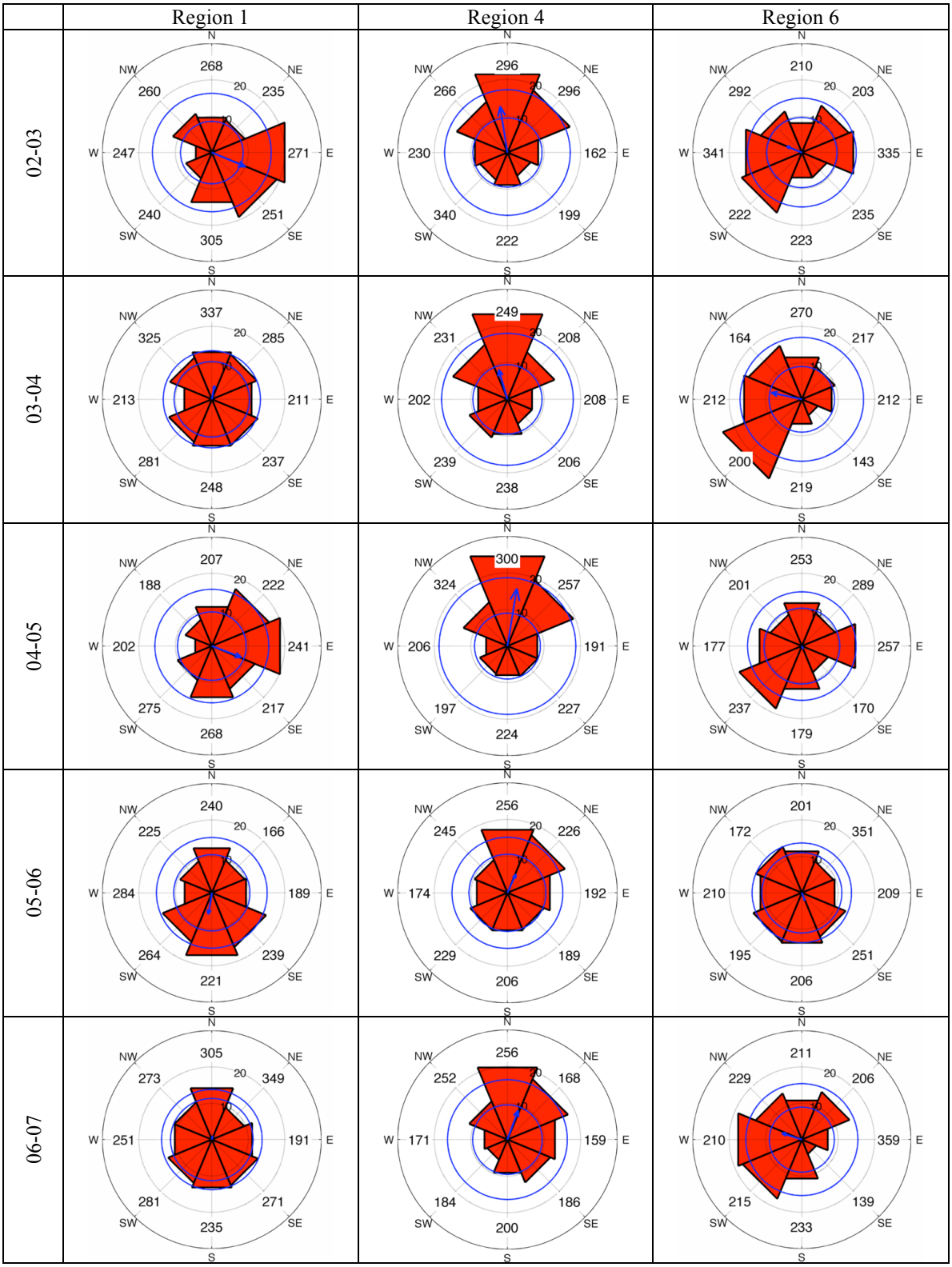

Figure 3.16. Same as Fig. 3.15 except for the GFS.

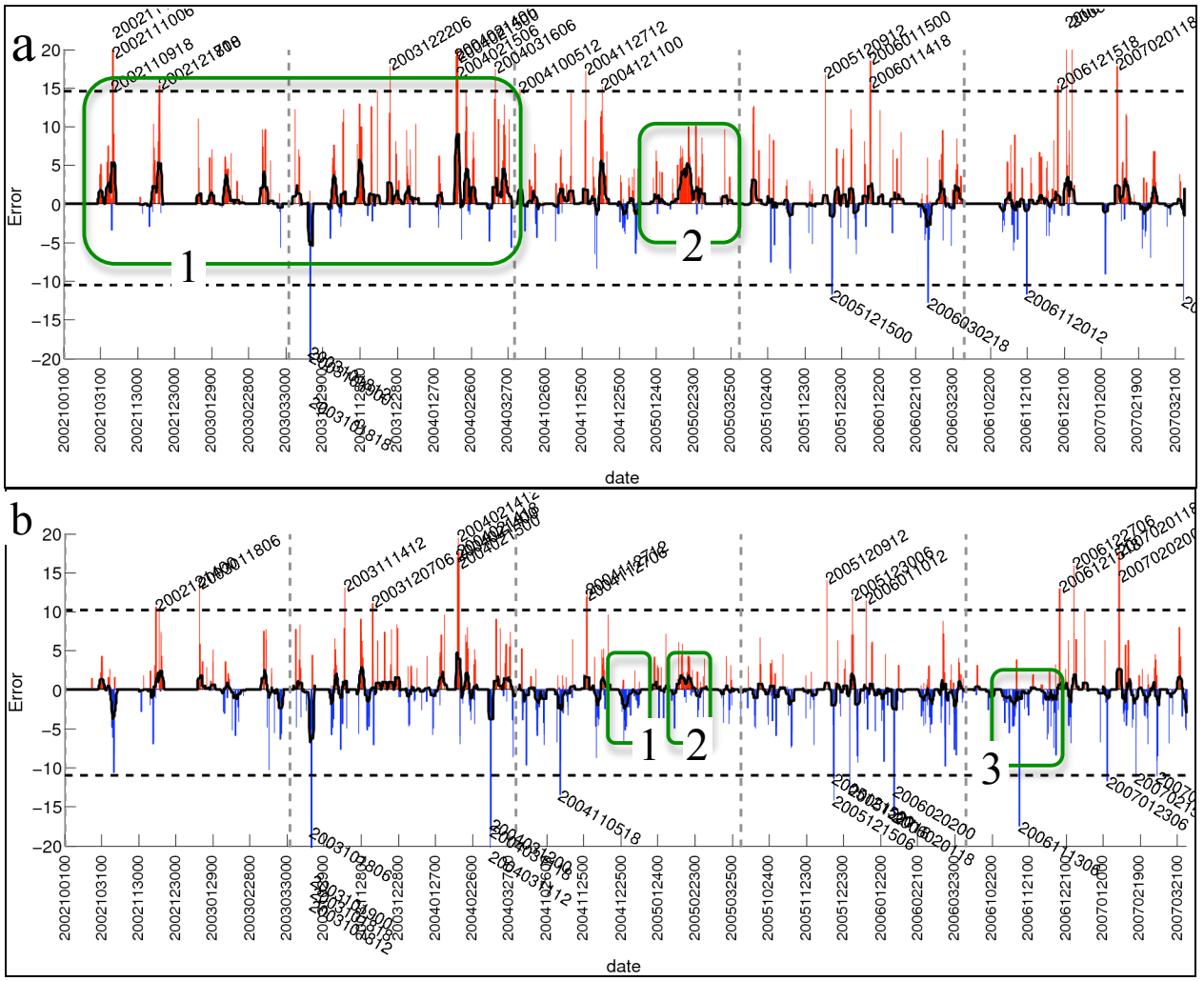

Figure 3.17. Time series of cyclone central pressure errors at hour 48 for region 1 during the 5 cool seasons (separated by gray vertical lines) for the (a) NAM and (b) GFS. Horizontal black dashed lines denote 2 standard deviations above and below the mean central pressure error for region 1, with model forecasts exceeding this threshold highlighted by the model initialization date. The solid black line is the 30-day running mean of cyclone central pressure errors. Interesting periods for text discussion are highlighted with green boxes and numbered.

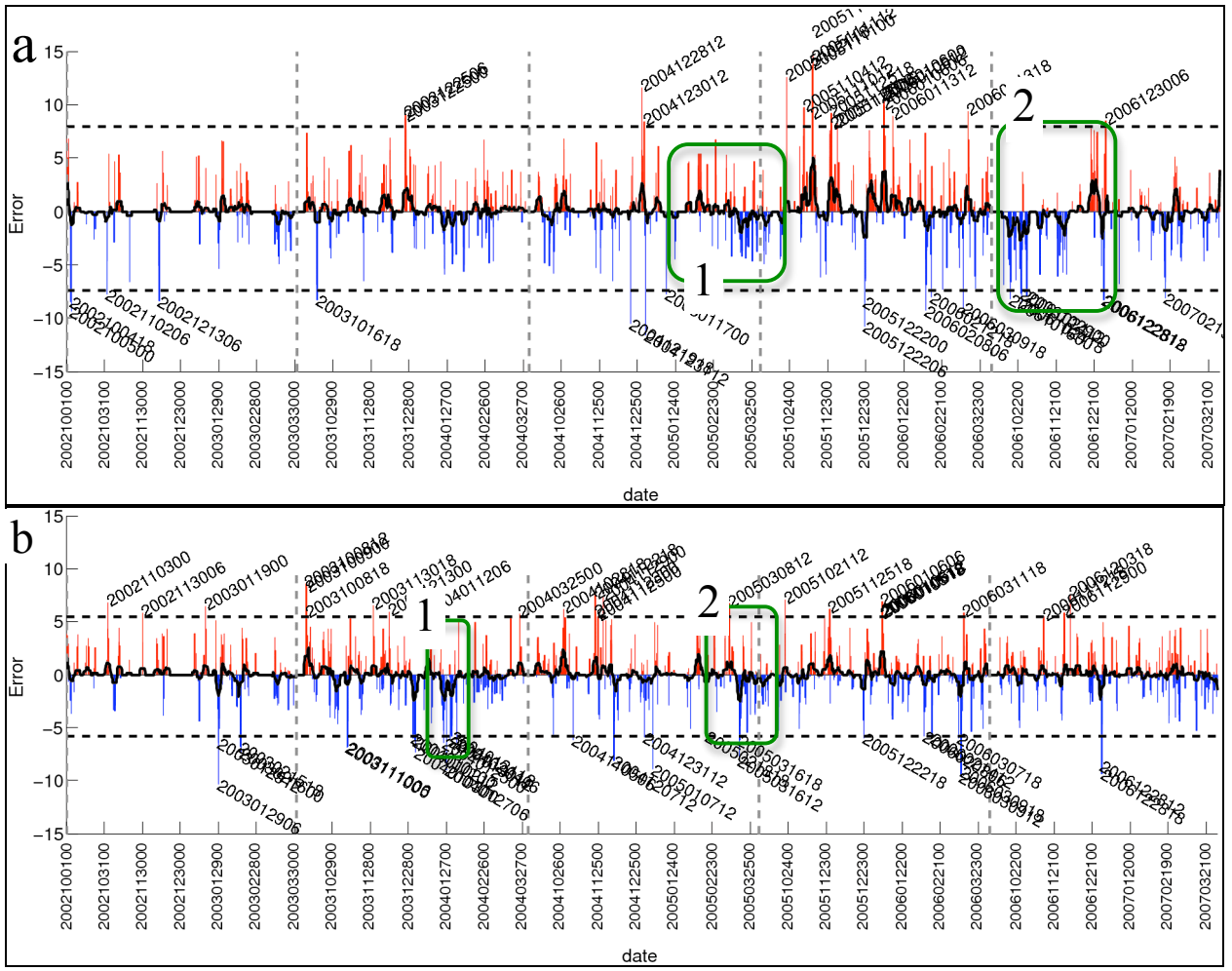

Figure 3.18. Same as Fig. 3.17 except for region 4.

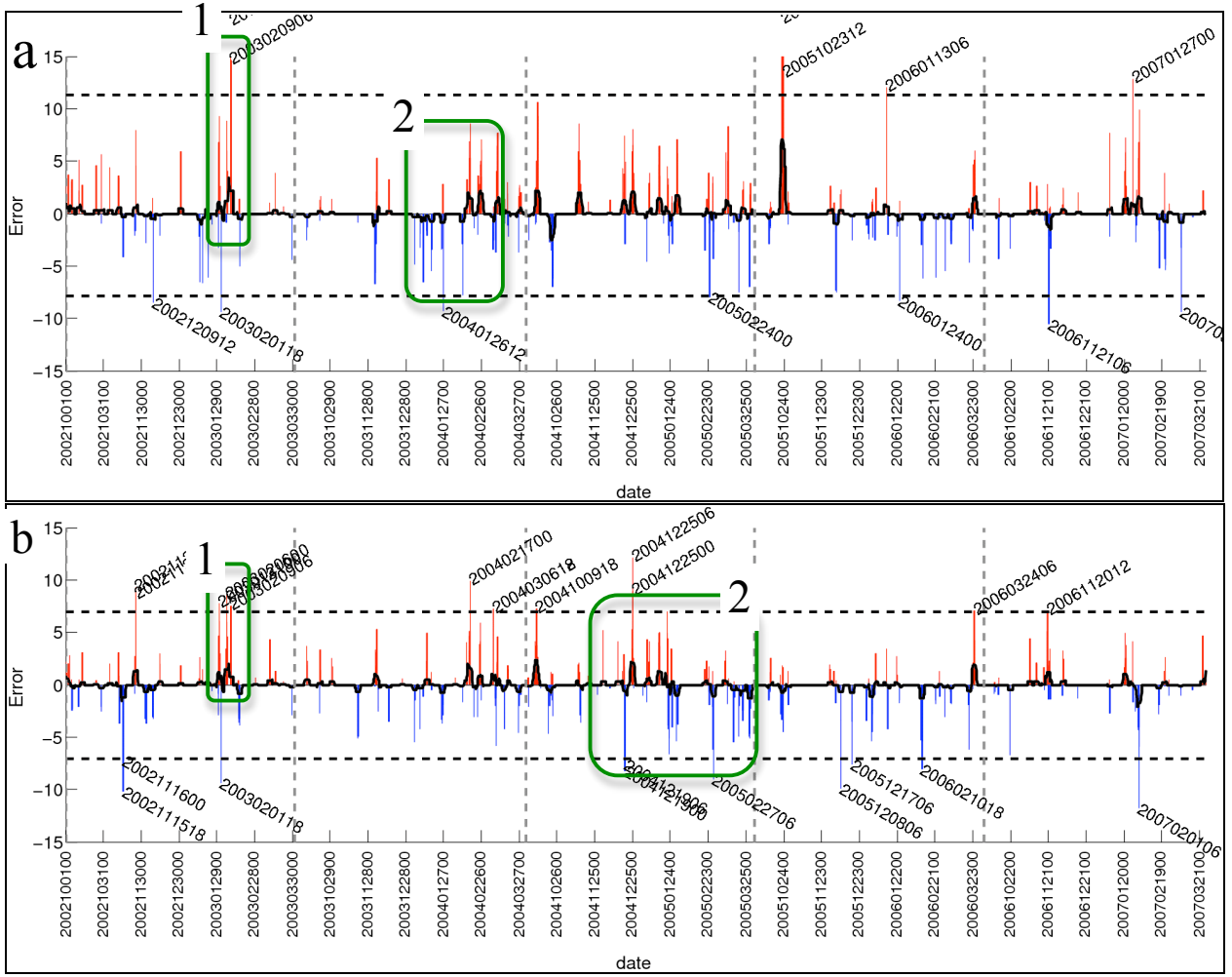

Figure 3.19. Same as Fig. 3.17 except for region 6.

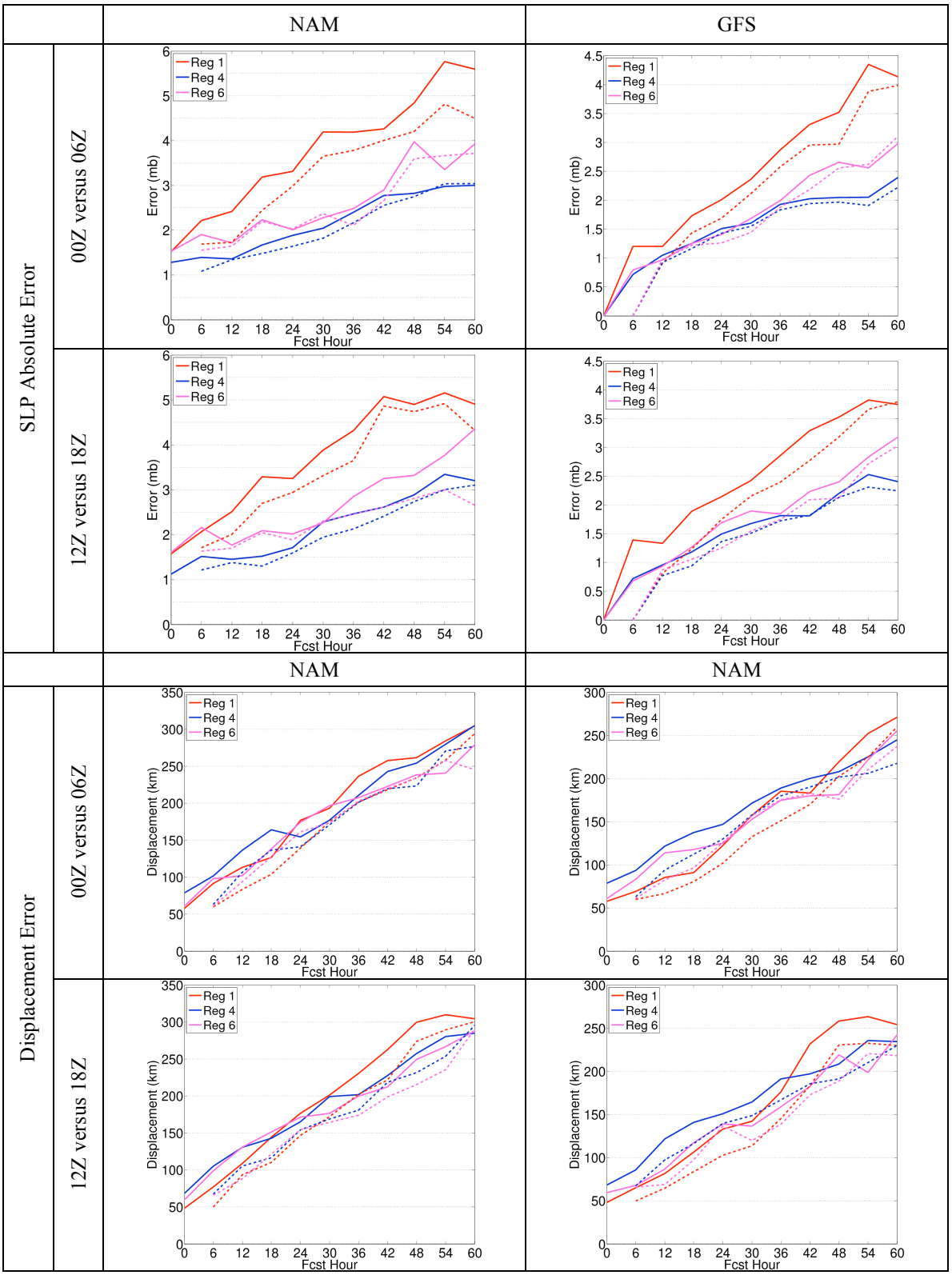

Figure 3.20. Cyclone mean absolute pressure (in mb) and displacement (km) errors versus forecast hour for the 0000 UTC and 0600 UTC runs (valid at the same time) as well as the 1200 and 1800 UTC cycles for the NAM and GFS. The 0000 and 1200 UTC errors are solid, and the 0600 and 1800 UTC errors are dashed.

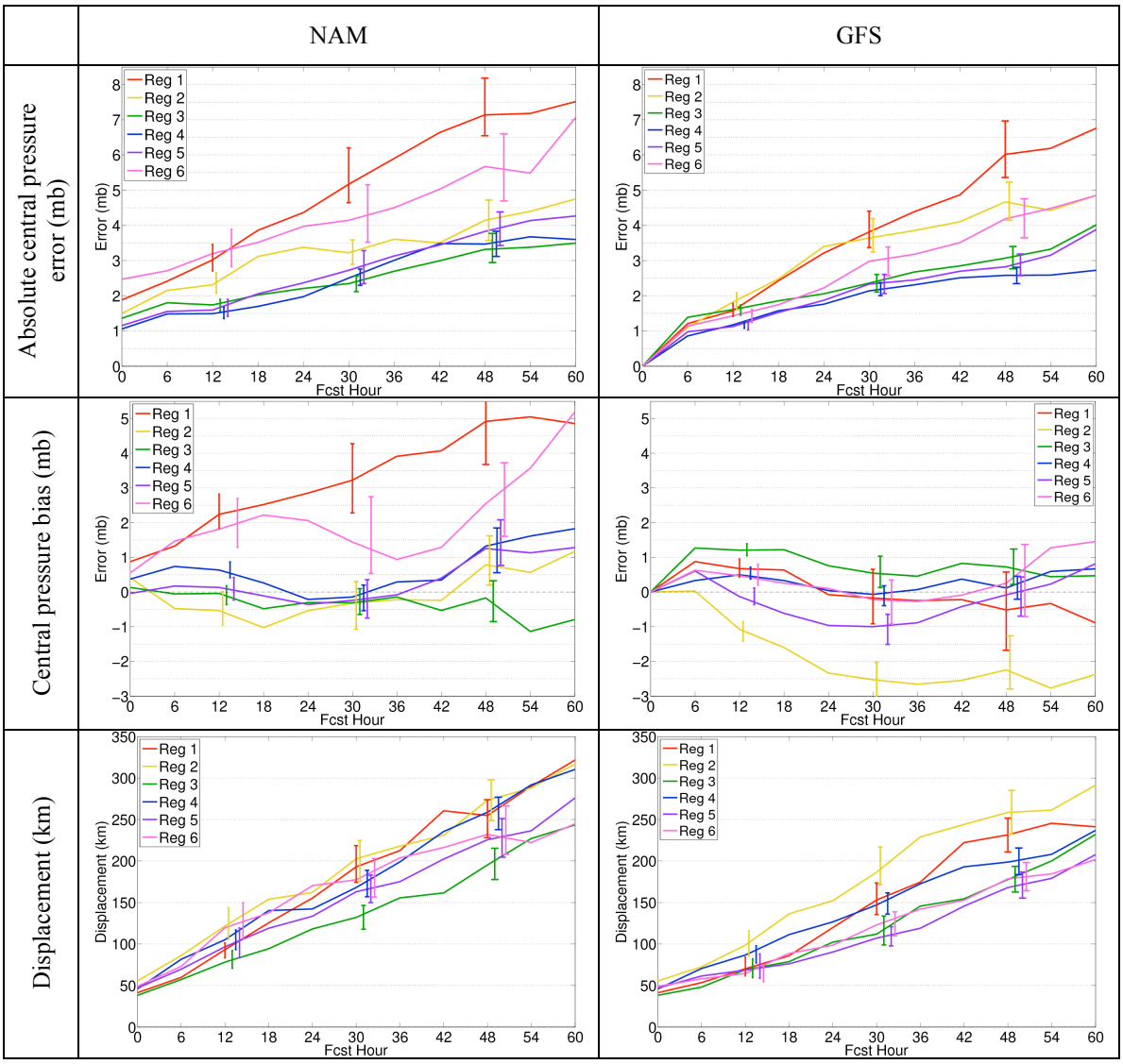

Figure 3.21. Same as Fig. 3.9 except for deep cyclones (more than 1.5 standard deviations below mean).

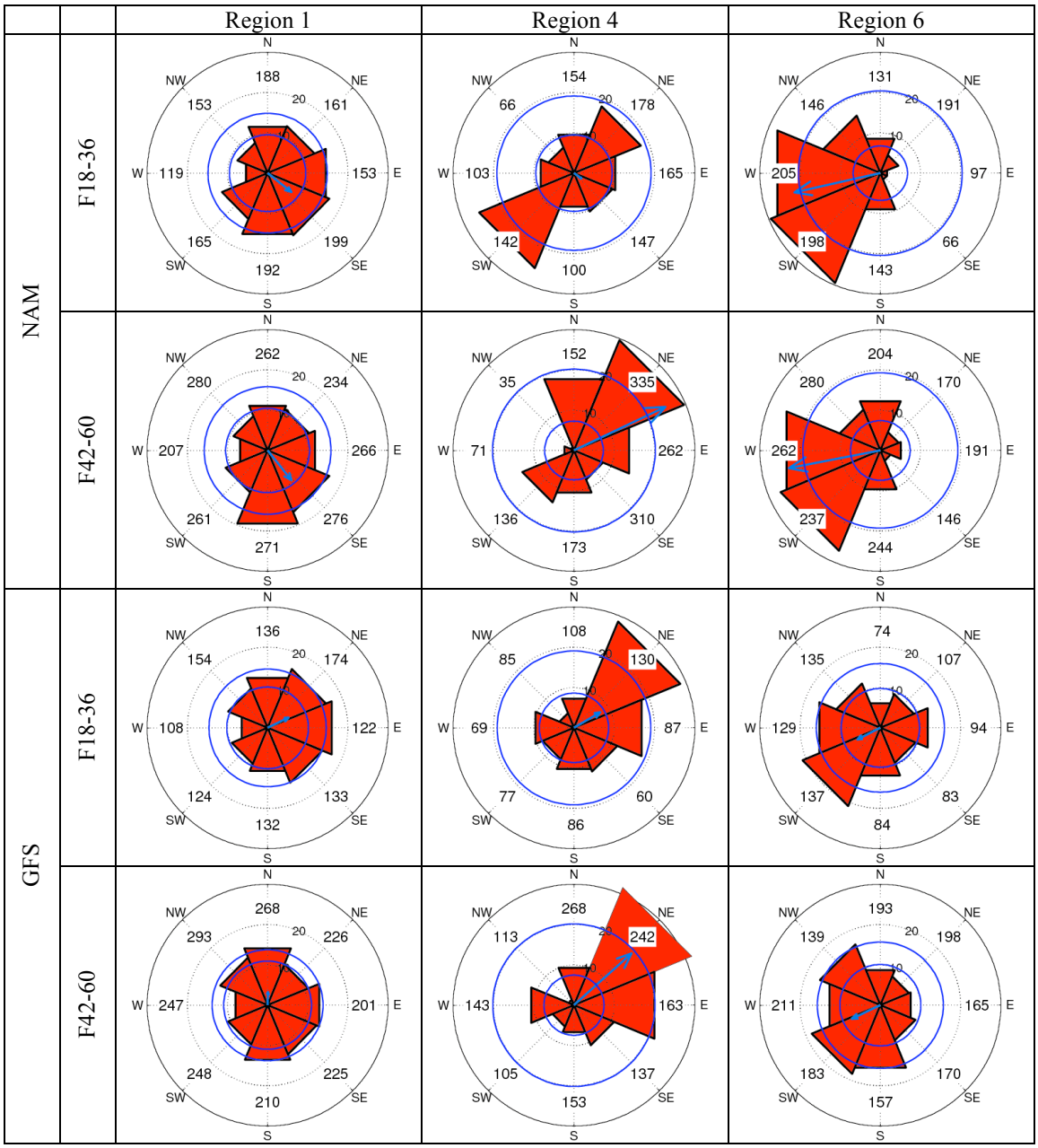

Figure 3.22. Same as Fig. 3.14 except for deep cyclones.

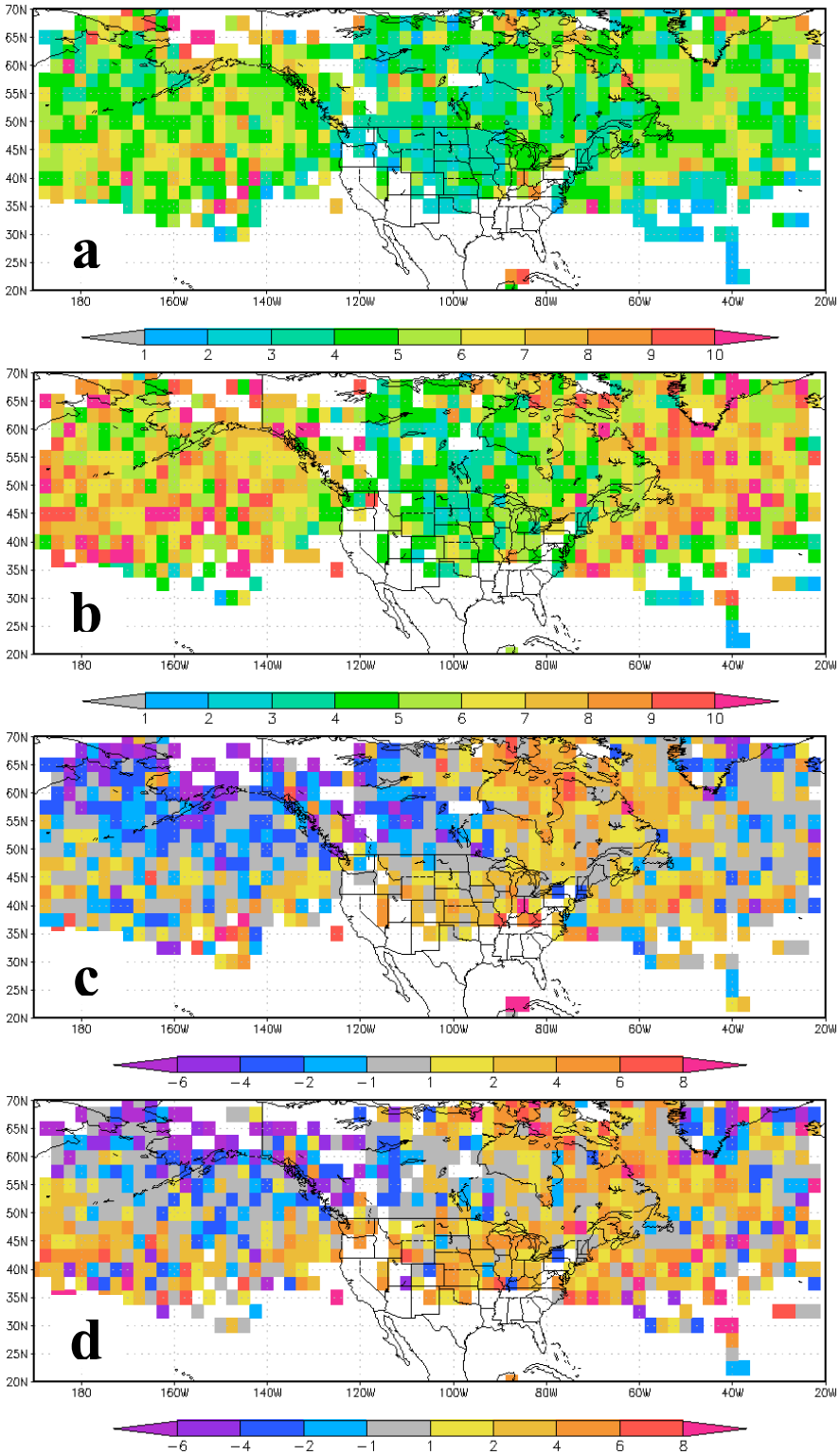

Figure 3.23. Spatial distribution of GFS central pressure absolute errors (in mb) for hours (a) 72-90 and (b) 102-120. (c) and (d) Same as (a) and (b) except for the mean error (in mb).

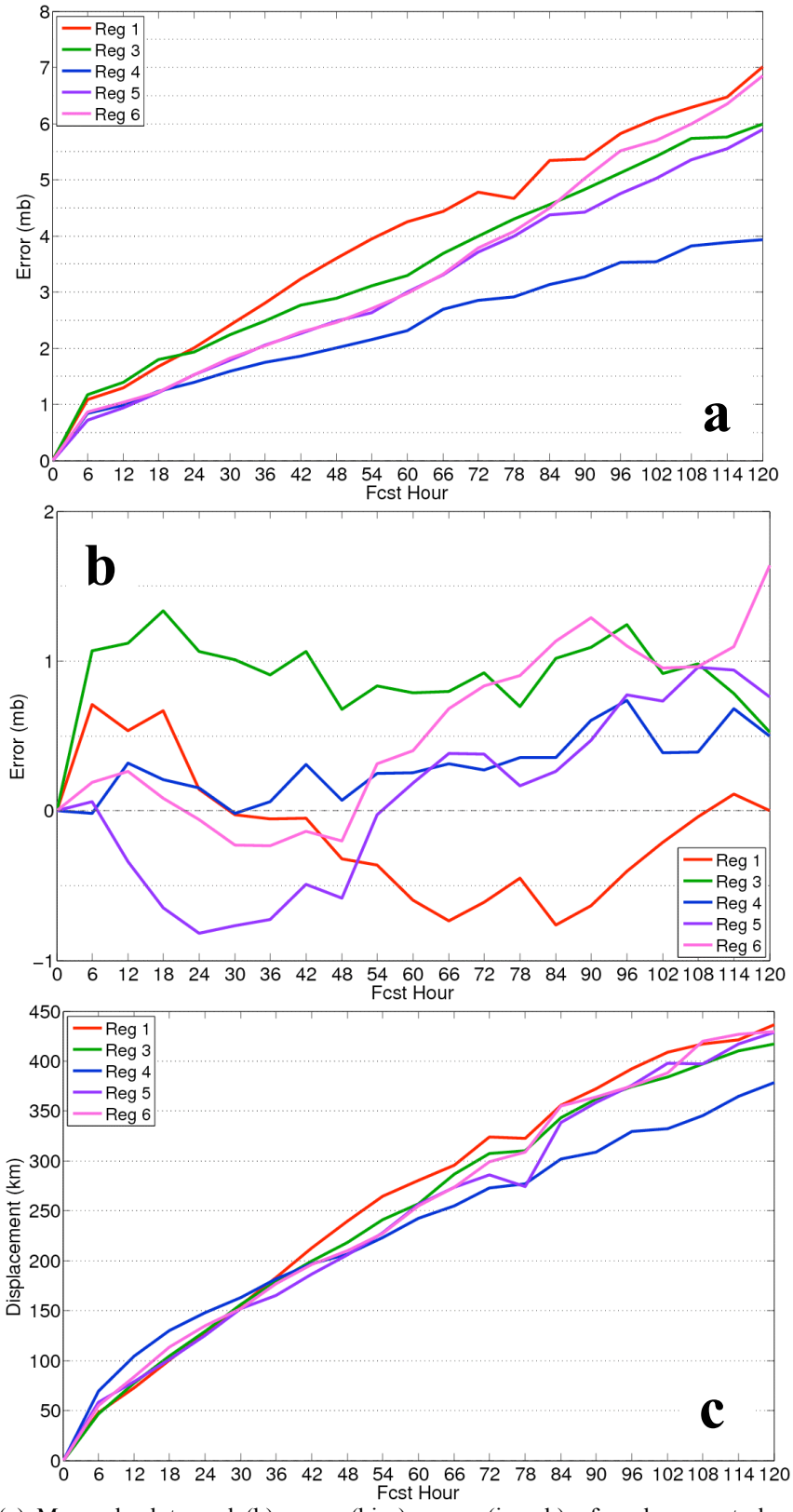

Figure 3.24. (a) Mean absolute and (b) mean (bias) errors (in mb) of cyclone central pressure versus forecast hour for the GFS during the 2002-2007 cool seasons for the various regions specified in Fig. 2.3. (c) Same as (a) except for displacement errors (in km).

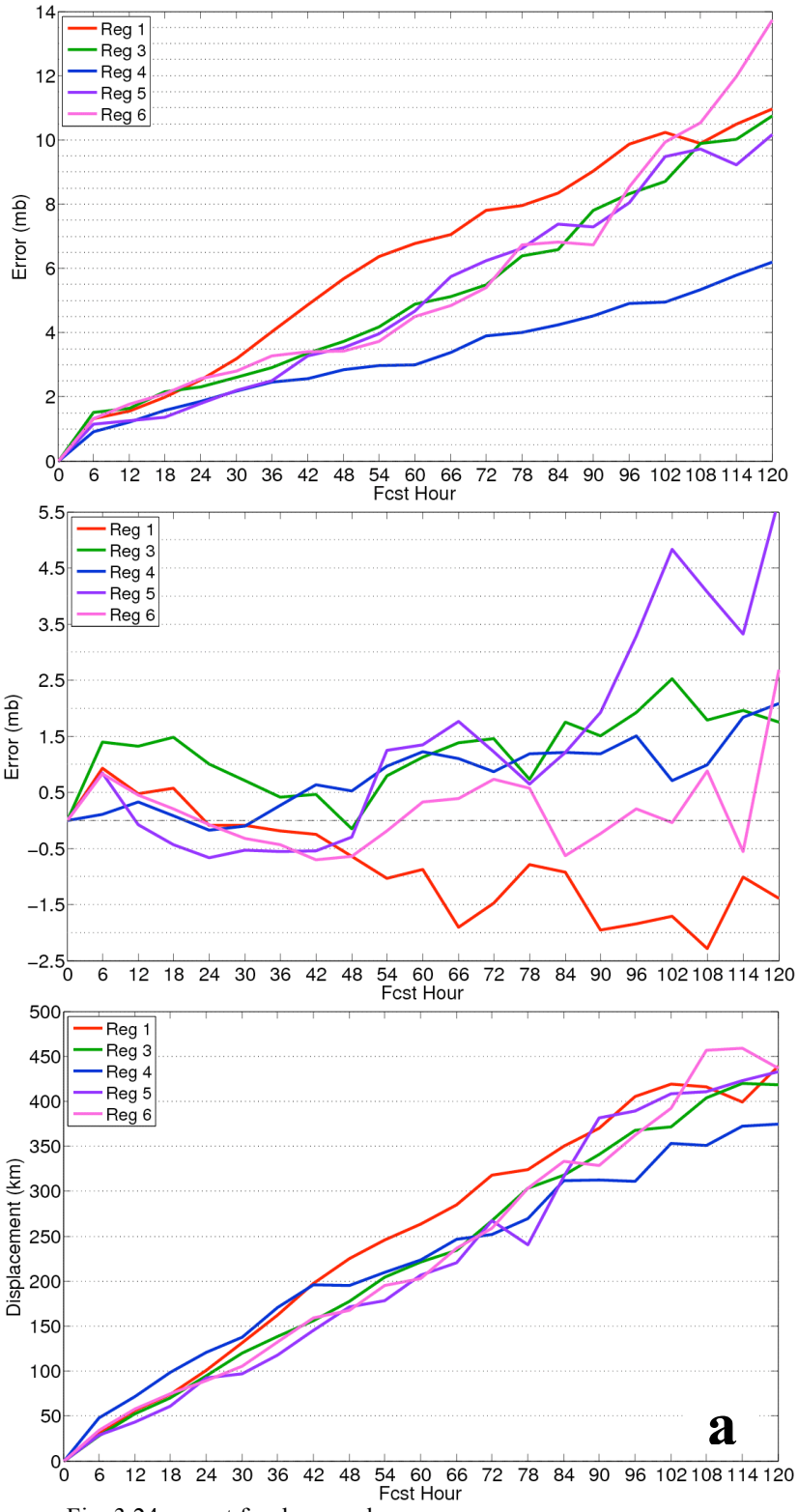

Figure 3.25. Same as Fig. 3.24 except for deep cyclones.

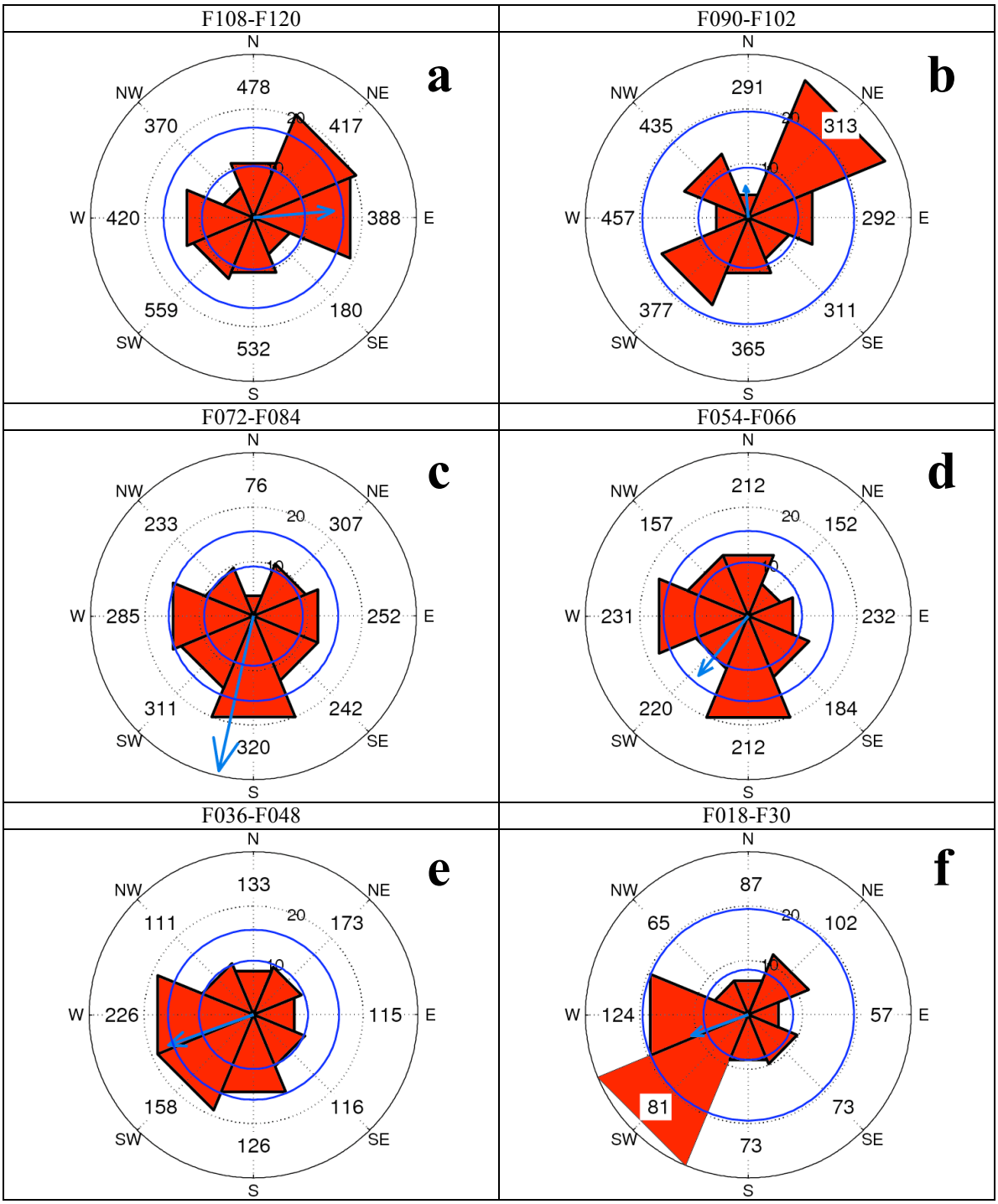

Figure 3.26. Histograms showing the frequency of model cyclone positions relative to the observed position (center point) for 45° bins centered on N, NE, E, etc… in the GFS for decreasing lead time for the western Atlantic. The black range circles are every 10%, from 0 to 30. The blue range circles represent the 90% confidence intervals below and above the range of the bins. The numbers for each radial represent the average displacement error (in km) within that directional bin. The blue vectors indicate the mean displacement vector, with each gray range ring every 25 km.

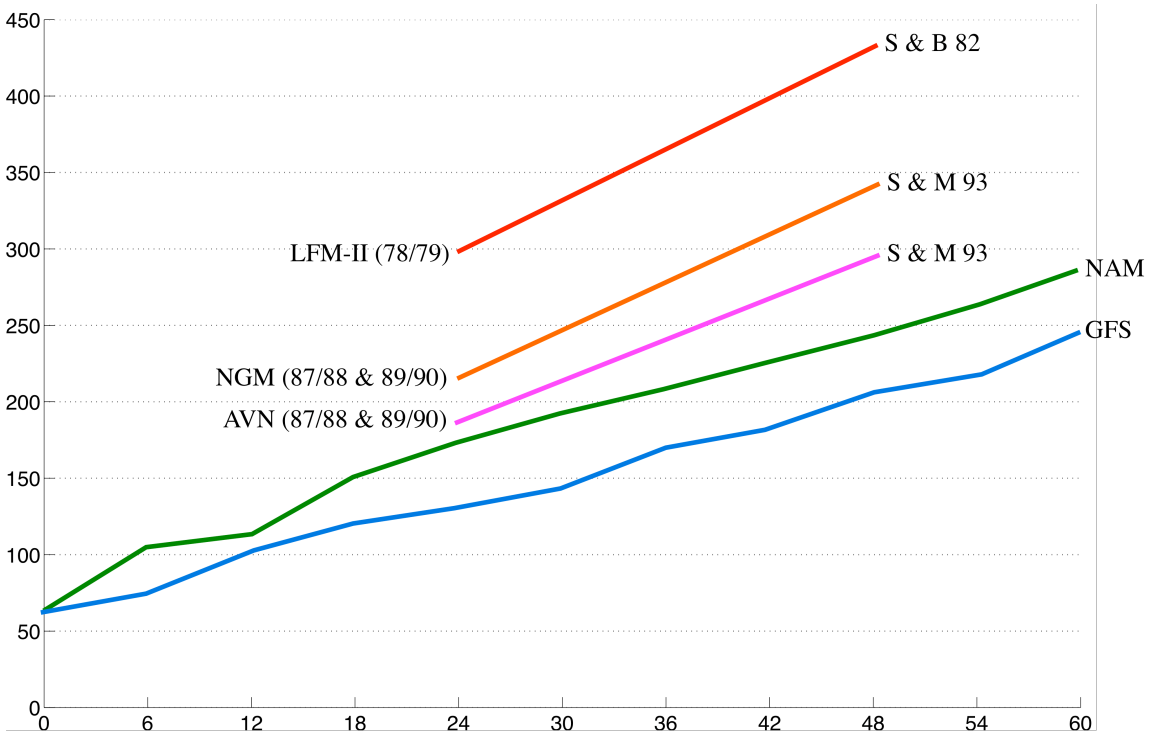

Figure 3.27. Extratropical cyclone displacement errors (in km) versus forecast hour for the LFM-II (Silberberg and Bosart 1982) (1978/79 cool season) (CONUS and oceans), the NGM and AVN (Smith and Mullen 1993) (1987/88 and 1989/90 cool seasons) (Atlantic), and the NAM and GFS (2002-2007 cool seasons) (Atlantic).

IV. SOURCE OF MODEL ERROR

a. Impact of Large West Coast Errors

Another goal of this research was to obtain some insight into the potential source locations of cyclone forecast errors. This was accomplished by analyzing the average large-scale flow patterns associated with certain cyclone errors. Several previous studies have shown that model forecast errors can propagate eastward from the Pacific and impact the eastern U.S. (Langland et al. 2002; Elmore et al. 2006). For example, Elmore et al. (2006) found that the initialization errors of upper-level short-wave troughs along the Pacific Coast persist as the short-wave crosses the U.S. Langland et al. (2002) showed that the poor numerical forecast of the 24-25 January 2000 "surprise" snowstorm may have resulted from errors in the initial conditions over the eastern Pacific on the 21 January.

In order to illustrate whether errors associated with cyclones over the eastern Pacific propagate downstream and impact the cyclone events in the eastern U.S., figure 4.1 shows the evolution of the normalized mean absolute SLP error, averaged for all 28 central Pacific (CP) cyclone cases with F006 cyclone central pressure absolute errors more than 1.5 standard deviations greater than the mean for this region (~4.2 mb). The SLP error at each point was normalized by calculating the mean absolute SLP error field, during all 5 cool seasons at each forecast hour, and then subtracting it from the composite absolute SLP errors at the same forecast hours. For the large F006 central Pacific cyclone error events (Fig. 4.1), there are relatively large mean absolute SLP errors (~2 mb) near the southeast Alaskan coast in association with a cyclone over the northeast Pacific by F030 (circled). Throughout the 48-66 h forecast these errors spread north and east over Alaska and into central Canada, with part of the errors moving more quickly out ahead of the main errors. Through F102, these large errors spread east and south as a surface trough develops over the Great Lakes.

To put these large error cases in perspective with the climatological errors in this region, Figure 4.2 represents the climatological evolution of SLP errors for 28 randomly selected days during the 5 cool seasons. At F030, there is a cyclone in the Gulf of Alaska associated with 0.5-1.5 mb SLP errors. During the next 36 hours of the forecast, relatively large errors (>2 mb) move onshore over the Pacific northwest, but unlike in the large errors events, these errors never make it to the eastern U.S. (region 5). Relatively large errors develop in place over the eastern U.S. by F066. Therefore, it appears that errors still develop late in the forecast, regardless of Pacific errors earlier in the forecast.

In order to verify if large SLP errors in the eastern U.S. always originate over the Pacific, the same composites were made for eastern U.S. (region 5) cyclone events (25) with F96 cyclone central pressure absolute errors more than 1.5 standard deviations greater than the mean for this region (~4.7 mb) (Fig. 4.3). Looking back in time at F030, there are small errors over the Pacific with a cyclone in the Gulf of Alaska. By F054-F078, a cyclone develops and moves east across the Great Lakes and develops an average SLP error of ~1-2 mb. As it crosses the Northeast U.S. by F78, the errors grow rapidly, with >>2 mb errors by F102 over entire Northeast U.S. and eastern Canada. Since cases with large errors over the eastern U.S. are not necessarily associated with large errors

over the Pacific a few days earlier, this suggests that large eastern U.S. cyclone errors can also grow more locally.

b. Eastern U.S. Error Evolution – Cyclone Tracks

To focus on the evolution of large cyclone error events (>1.5 standard deviations) over the western Atlantic, tracks were plotted of all GFS cyclone events with large positive (negative) central pressure error above (below) the mean positive (negative) error at F96 (Fig. 4.5). The red line segments indicate 6-hour periods when the forecast cyclone is filling faster than observed, or is deepening slower than observed, while the blue lines indicates 6-hour periods when the forecast cyclone is deepening faster than observed, or is filling slower than observed.

For the large positive cyclone errors at hour 96, the majority of cyclones track east and northeast along the eastern edge of the Gulf Stream. (Fig. 4.5a). The positive errors for most cyclones begin growing as the cyclone approaches the immediate coast, and continues over the Atlantic. There is also evidence of some of the cyclones originating from Canada, traveling southeast across the Great Lakes. In contrast, most negative cyclone events track toward the northeast through east, closer to the coast than positive events (Fig. 4.5b), with negative errors usually growing from hour 0 on. It also appears that the negative errors are reduced as cyclones move north into Canada.

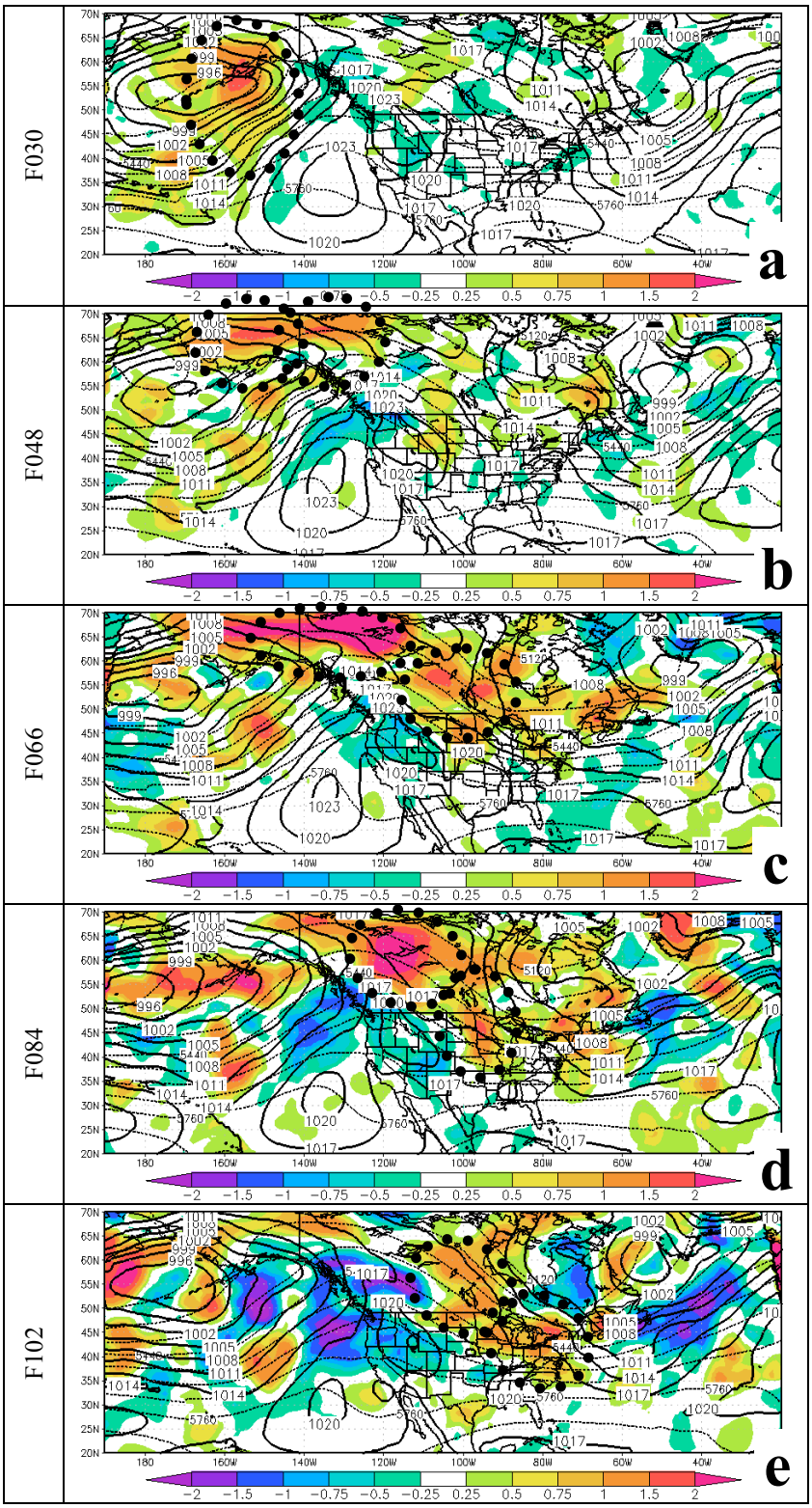

Figure 4.1. GFS mean sea level pressure error (shaded every 1.5 mb), mean observed 500 mb heights (dashed contours every 60 m), and mean observed sea level pressures (solid every 2 mb), averaged at each grid point for all eastern Pacific cyclones with F006 central pressure absolute errors in the GFS more than 3 standard deviations greater than the mean over the central Pacific for cyclones (28 cases) for hours (a) 6, (b) 30, (c) 54, (d) 78, and (e) 102.

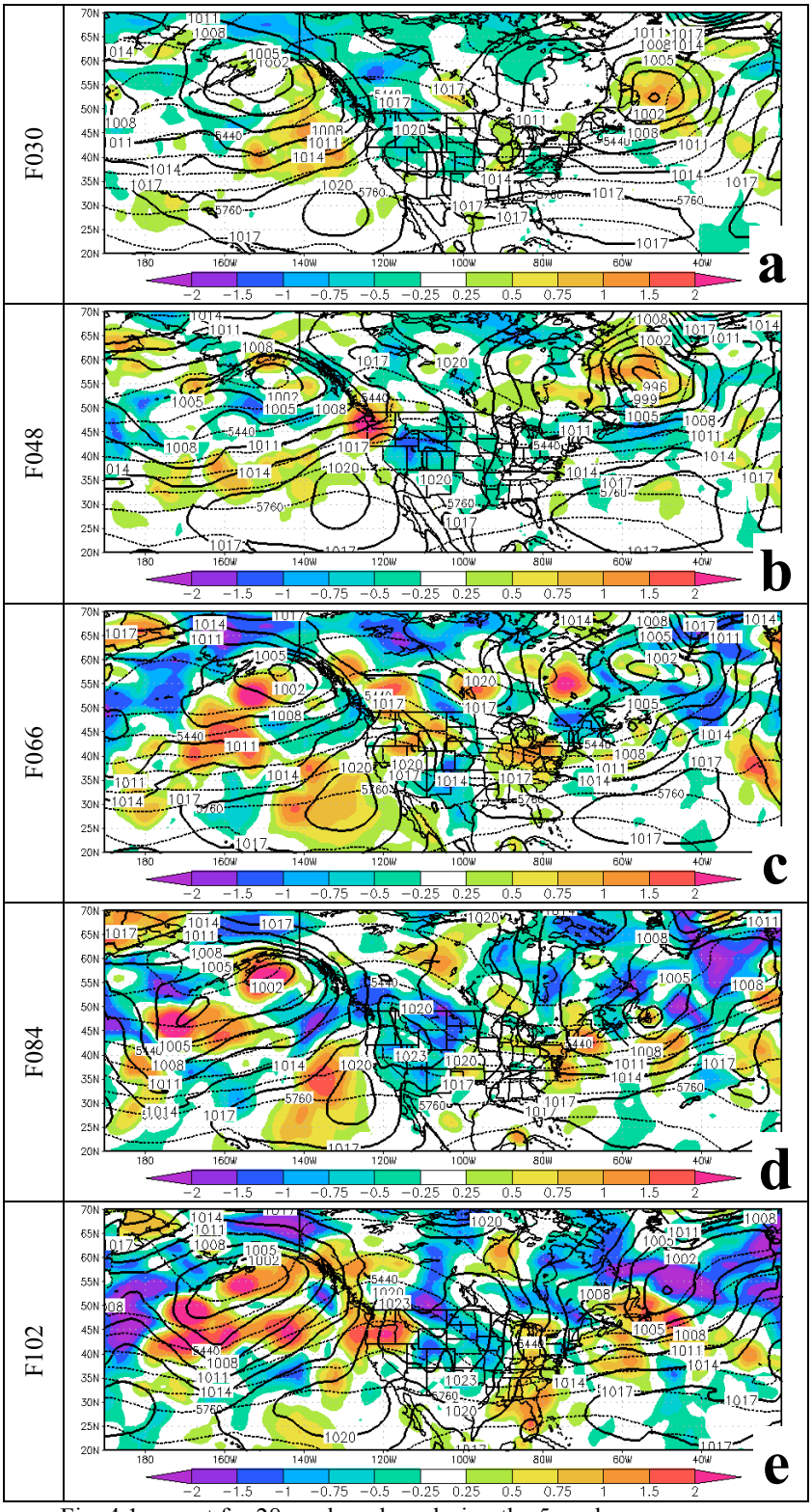

Figure 4.2. Same as Fig. 4.1 except for 28 random days during the 5 cool seasons.

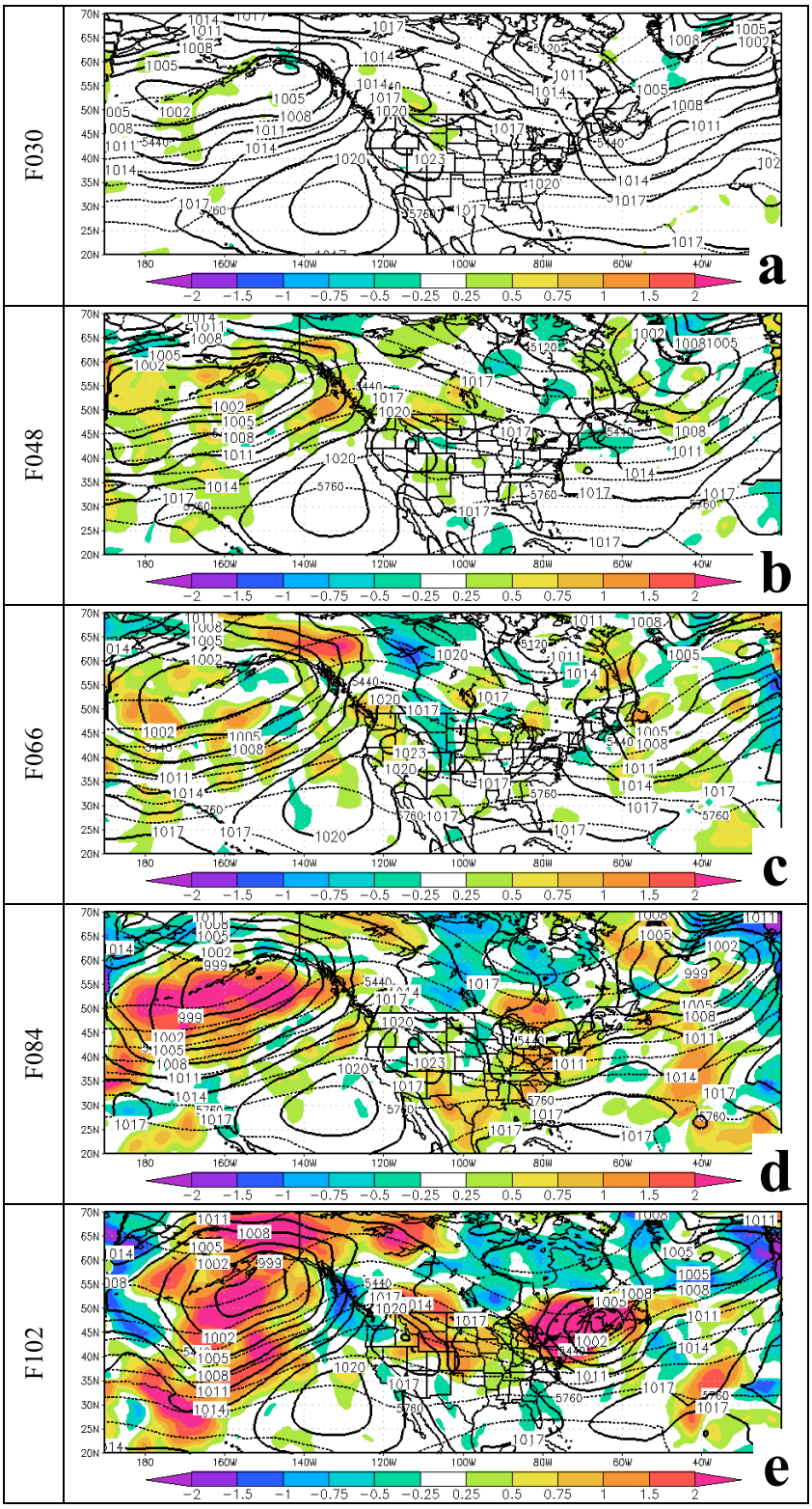

Figure 4.3. Same as Fig. 4.1 except for cyclones with large (> 1.5 standard deviations) errors at F96 in the eastern U.S.

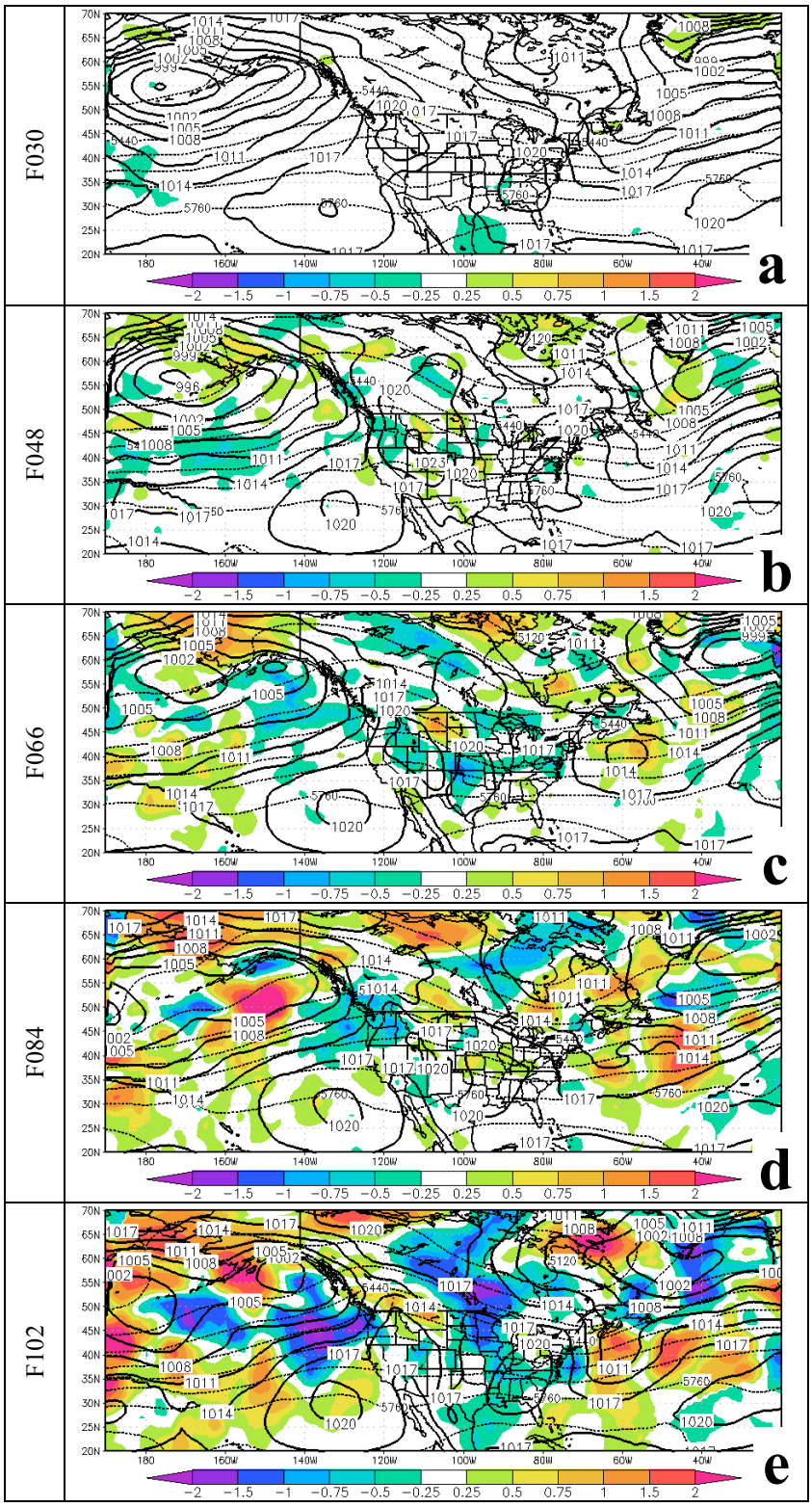

Figure 4.4. Same as Fig. 4.1 except for all eastern Pacific cyclone cases with cyclone central pressure absolute errors less than ¼ standard deviations.

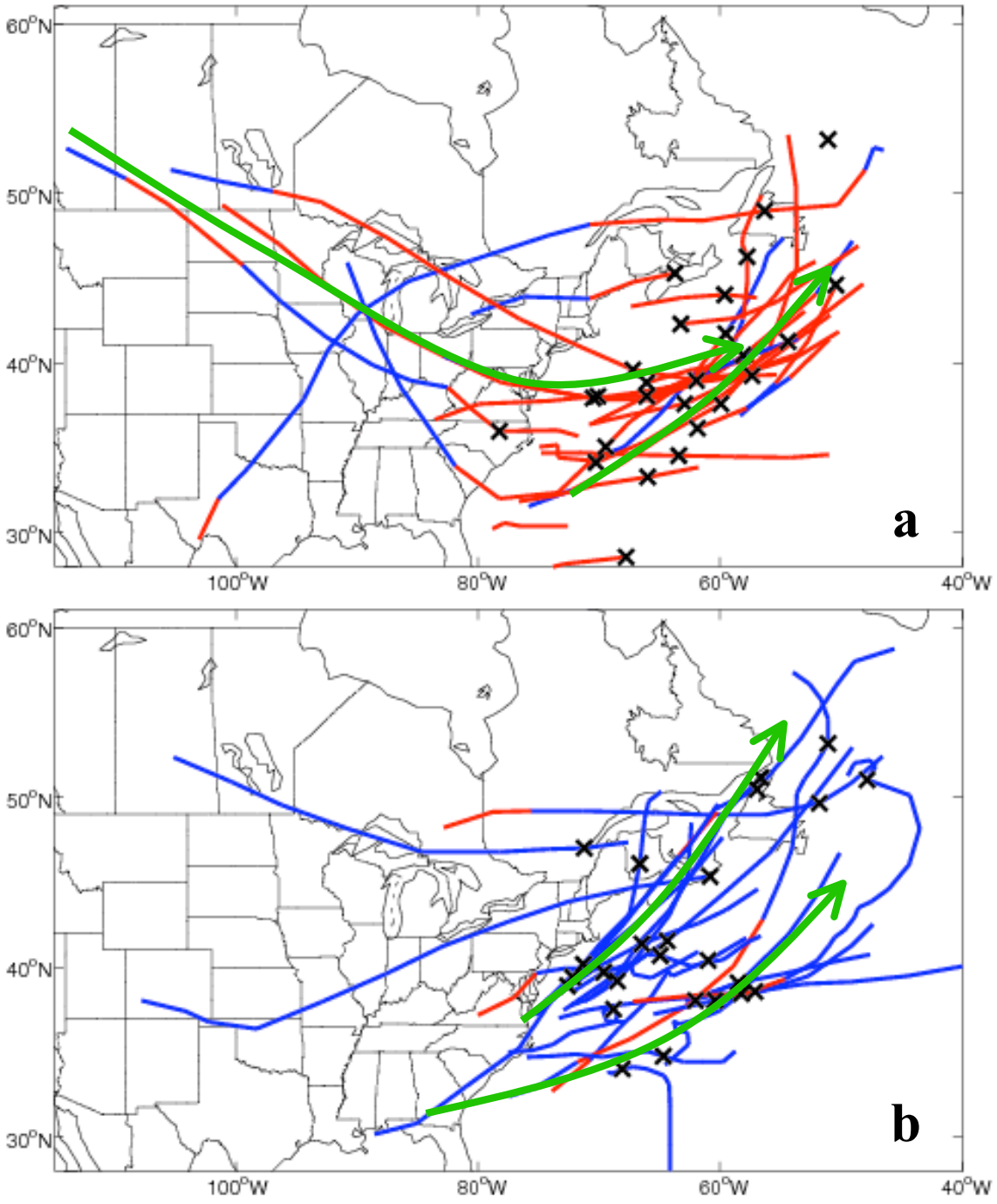

Figure 4.5. Tracks of western Atlantic (region 6) GFS cyclones with large (> 1.5 standard deviation) central pressure errors for cases with (a) positive and (b) negative central pressure errors at hour 96 that are more than 1.5 standard deviations above (below) the mean error of all cyclones with positive (negative) error. The red line segments indicate 6-hour periods when the forecast cyclone growth rate or weakening is slower or faster than observed, respectively. Meanwhile, blue lines indicate 6-hour periods when the forecast cyclone is deepening (or filling) faster (slower) than observed. The black x's indicate the location of the cyclones at F96.

V. SHORT RANGE ENSEMBLE VERIFICATION

Recently, there has been a growing interest in using ensembles in operational forecasting. Ensembles can be composed of different models, physics, and/or initial conditions. The short-range ensemble forecast (SREF) system at NCEP is a combination of all three approaches (Du et al. 2003). Because deterministic model forecasts do not provide any forecast uncertainty, this can result in overconfidence in the forecast. A distribution of forecasts from an ensemble can provide additional probabilistic information to the public.

This chapter describes the cyclone forecast errors in the NCEP SREF system for the cool seasons of 2004-2007, and how well the SREF performance compares with the NAM and GFS for the same cyclone events. To be included in the verification dataset, each cyclone had to be objectively identified by the tracker in 3 out of the 5 members of each SREF subgroup (EKF, EBM, RSM, see chapter II.a) (McQueen et al. 2005).

a. Error vs. Forecast Hour

i. Cyclone Central Pressure Absolute Error

It has been shown that the mean of an ensemble forecast has more skill on average than any single deterministic model or individual ensemble member for many different atmospheric parameters (Stensrud et al. 1999). To see how the SREF skill compares to the deterministic NAM and GFS, the errors in each subgroup of SREF (EKF, EBM, and RSM), as well as the full 15-member mean, were averaged over each region in figure 2.3, at each forecast hour (F03-F63). Figure 5.1 shows the cyclone central pressure absolute errors for the GFS, NAM, and SREF in all regions. For all 6 regions, the GFS has the lowest mean absolute errors of all the models and the various SREF sub-ensemble groups. This result is statistically significant at the 90% level before F51 in regions 1 and 5, before F27 in regions 2 and 6, and during the entire forecast period in region 4. For regions 2 through 5, the RSM group has the largest absolute errors than any other model or group, especially before F45 (significant at 90% level in regions 2, 4, and 5). In all regions, the SREF mean has higher or comparable skill to the mean of each ensemble subgroup. In regions 2 through 5, the SREF mean has less skill than the NAM before F33, but SREF has greater skill after F33. Over the ocean areas (regions 1 and 6) the SREF mean has more skill than the NAM at all forecast hours.

ii. Cyclone Central Pressure Error

Figure 5.2 shows the mean error of cyclone central pressure for the SREF, NAM, and GFS. In all regions except region 1, the cyclone SLP error is negative at F03 for all members and the SREF mean, and (except for regions 1 and 3) cyclone SLP errors become slightly positive by hours 33-45. This trend is especially true for the Atlantic (region 6), which has SREF errors ranging from ~0.5 (RSM members) to ~2 (EBM members) by F63. SREF errors in region 3 are negative at F03, but become even more negative throughout the forecast, peaking between -1 and -1.5 mb by F45. Region 1 has positive errors at F03 in the ETA members, and negative errors to begin with in the RSM

members. The RSM members continue with a large negative bias (-1 to -1.5), while the ETA members and the SREF 15-member mean have a positive bias peaking at 1.5-3.5 mb by F39. Overall, these results suggest that the SREF members cluster by model type, with similar biases for each subgroup.

iii. Cyclone Displacement

Figure 5.3 shows cyclone displacement errors for the GFS, NAM, and SREF vs. forecast hour for all regions. The SREF mean cyclone displacement was calculated by first calculating the mean position of the forecast cyclone. This was done by averaging the latitudes and longitudes separately for all ensemble members. Then the SREF mean cyclone displacement was calculated as the distance between this SREF forecast cyclone and the observed cyclone. In all regions except region 2, the GFS has the smallest displacement error (130-160 km at F45) of all models and SREF subgroups. These errors are significantly (at the 90% level) less than the other models early in the forecast (< F33) for regions 1, 3, and 5, and at F27-F45 for region 6. For all regions, the SREF mean has smaller displacement errors (160-200 km) than the NAM and all SREF groups. The RSM group has significantly larger (by up to 60 km in region 3 and up to 100 km in region 4) displacement errors than all other models and subgroups in regions 3 and 4.

b. Impact of Including WRF Members

During the last cool season of this study (2006-2007), 6 WRF members were added to the SREF (Du et al. 2006). An important question is whether the WRF ensemble members can help increase the skill of the ensemble. Figure 5.4 shows the mean absolute errors of cyclone central pressure versus forecast hour for the 2006-2007 cool season for the GFS, NAM, and SREF (including the 6 WRF members). Except for over the Pacific (region 1) and Atlantic (region 6), the WRF errors exceed the other SREF groups, the GFS, and the NAM. Errors are particularly high over the western U.S. (region 2) and eastern Canada (region 3), with errors peaking at ~6 mb at F33 (F57) in region 2 (3). Except for over the oceans, the SREF 21-member ensemble error (including the WRF members, black dashed) are larger than the SREF 15-member mean without the WRF members. Over the Pacific, the WRF members have more skill than the EBM and EKF groups, while the WRF is comparible to the NAM; therefore, the mean absolute error of the 21-member SREF is lower than the 15-member ensemble. Over the Atlantic, the additional WRF members have very little impact on the skill.

Figure 5.5 shows the cyclone central pressure mean error (bias) including the WRF members. Except in the central U.S. (region 4), the WRF group overdeepens cyclones. Cyclone SLP errors are largely negative in all regions, especially over the Rockies and eastern Canada (regions 2 and 3), where cyclone SLP errors peak at -4 to nearly -6 mb between F33 and F57. In the other regions, the WRF members have an average cyclone SLP error of -1 to -2 mb. Since the SREF mean cyclone SLP errors over the Pacific are largely positive without WRF , the addition of WRF members with a large negative bias helps decrease the mean positive bias of the SREF system. This leads to a larger overall skill including the WRF members than without (Fig. 5.4.1 and 5.5.1). In

regions 2-5, adding WRF members with a large negative bias further increases the overall negative bias of the SREF system.

The inclusion of WRF members also results in little improvement in forecasting of cyclone position in many regions (Fig. 5.6). The Pacific and western U.S. (regions 1 and 2) have a slight decrease in cyclone displacement, but the central U.S. (region 4) has larger displacement errors with the WRF. Eastern Canada, the eastern U.S., and the Atlantic (regions 3, 5, and 6) have little difference in cyclone displacement errors with the WRF. The WRF displacement errors are consistently large relative to the other models for regions 3-6.

Overall, the 6 additional WRF members did little to improve SREF cyclone forecasts. The exception is over the eastern Pacific, where the large negative bias in WRF helps to offset the large positive errors seen in the other SREF members. The WRF members increased cyclone SLP errors in the other regions, and cyclone position forecasts benefited little from the WRF members as well. Therefore, for the remainder of this chapter, the SREF analysis will include the 15 members without WRF between the cool seasons of 2004-2007.

c. Timeseries of SREF and GFS Pressure Errors

As shown in section III.c.iii for the GFS and NAM, the SREF also has intraseasonal variability in skill of the cyclone forecasts. To compare the SREF with the GFS over each cool season, Figure 5.7 shows a timeseries of the SREF mean minus the GFS central pressure error over the 2004-2007 cool seasons at forecast hour 45. For the eastern Pacific (region 1), a large majority (~65%) of cyclone events are better forecast by the GFS than the SREF. For much of February 2005, the SREF cyclone SLP errors are consistently greater than the GFS (Fig. 5.7a, box 1). Also, for nearly all of the 2006-2007 cool season, the SREF cyclone SLP errors are larger than GFS errors, especially for December through February (Fig. 5.7a, box 2). For the central U.S. (region 4), the GFS errors are not always better than the SREF mean. For example, at the end of the 2004-2005 cool season (Fig. 5.7b, box 1) and the beginning of the 2006-2007 cool season (Fig. 5.7b, box 3), there are periods when the SREF pressure errors are less than the GFS. In contrast, for most of November 2005 through January 2006, the mean SREF has larger errors than the GFS (Fig. 5.7b, box 2). In the western Atlantic, from the end of December 2005 through January 2006, most cyclone events are better forecast by the GFS than the SREF (Fig. 5.7c, box1).

d. Ensemble Skill Measures

i. Rank Histograms

In addition to verifying an ensemble's mean absolute error and bias, it is also important to verify the ensemble probabilistically. One way of measuring the skill of an ensemble is through a rank histogram (Talagrand diagram). Rank histograms show how uniform the skill is for all of the members and how representative the ensemble spread is relative to the true uncertainty.

Rank histograms of central pressure were constructed by first taking the 12 SREF members[2] with the smallest central pressure error at a particular forecast run and time period, and ordering the member central pressures from lowest to highest. Then, the observed central pressure was placed in the appropriate pressure bin. An observed central pressure is placed in the left-most bin if the central pressure is less than that of the member with the smallest pressure, and is placed in the right-most bin if the central pressure is greater than that of the member with the largest pressure. The number of cyclones in each pressure bin was summed over each forecast containing at least 12 ensemble members with a matched cyclone (Fig. 5.8). In region 1, there is a persistent underdeepening of SREF cyclones at all forecast hours, since a majority of observed cyclones have a central pressure near the deeper SREF members (left side of histogram). The SREF is also overdispersed (inverted U-shape) in region 1 through F39, which suggests that the SREF central pressure spread is too great. After F39, the SREF is slightly underdispersed given the more U-shape histogram, thus suggesting too little spread.

In region 2, cyclones are consistently forecast too deep for all forecast hours. In region 3, there is some overdispersion early in the forecast, which tends to decrease later in the forecast. Meanwhile, some overdeepening tends to develop in region 3 later in the forecast by F39. Region 4 experiences overdeepening through about F39, while afterwards, the SREF tends to be more underdispersed. There is also overdeepening in region 5, with overdispersion between F03 and F39. The SREF becomes underdispersed after F39. Meanwhile, there is overdispersion in region 6 throughout the entire forecast. Overall, the overdispersion tends to be over the storm track oceanic regions, while some of the largest underdispersion and overdeepening biases occur over the central U.S.

ii. Best Member Diagrams

Another quality of a good ensemble is that every member should have an equal chance of being correct. If the ensemble probability distribution is correctly reproducing natural uncertainty, each member should have nearly the same probability of being the best (Hamill and Colucci, 1997). This can be quantified using a best-member histogram, which was constructed by incrementing a counter for the member with the smallest absolute error for all cyclones in which at least 3 out of 5 members of each group forecasted a cyclone. The total number of times each member was best was then divided by the total number of cyclones forecast, to get the percentage of time each member is best.

Each member in SREF was not equally skillful. In region 1 (Figs. 5.9a,b,c), the best SREF member on average (11-15%) is the RSM control member at all forecast hours. In contrast, the EKF subgroup is least often the best member (~5% per member) for region 1, for all forecast hours. For region 2 (Figs. 5.9d,e,f), the EBM subgroup is consistently (8-10%) the best group. In region 3 (Fig. 5.10a,b,c), the EBM and RSM subgroups are best more often than the EFK subgroup between F03 and F15, but after

[2] A rank histogram has a fixed number of bins in which to place the observed values from each forecast. In order to increase the sample size for the rank histogram, the number of bins used was decreased to 12 to allow more cases, since not all members had a cyclone identified by the tracker. When there were more than 12 members with a cyclone, the best 12 were selected.

F15 all subgroups are similar. For the EBM and RSM in region 3, the control member is most frequently the best member for all forecast hours. For region 4 (Figs. 5.10d,e,f), the EBM and RSM subgroups are best more often than the EKF group. All groups in region 5 have similar skill (Figs. 5.11a,b,c). Between F03 and F39 in region 5, the control members are the best members in the EBM and RSM groups, while after F63 the control members are the best in the EKF and RSM subgroups. In region 6 (Figs. 5.11d,e,f), the RSM control member is consistently the best member between F03 and F15 (~9-10%). Between F27 and F39, the control members are best for each group in region 6. After F39, the EBM control member is the best on average.

Best-member diagrams were also created to show the distribution of skill across each member with respect to cyclone position. In the eastern Pacific (region 1), the control RSM member is always the best member (13-18%) (Fig. 5.12a,b,c). For the western U.S. (region 2) (Fig. 5.12d,e,f), the groups are fairly even through F39, and the RSM group is best at hours 51-63 (8-10% vs. 4-7% for the other groups). In eastern Canada (region 3) (Fig. 5.13a,b,c) the RSM control member is best through F39, especially F03-15 (~20%). The control members are the best for each group through F39 as well. In the central U.S. (Fig. 5.13d,e,f), the RSM control member is best at F03-F15 (~12%). After F15, the EKF group is best slightly less (4-6%) than the other groups (6-8%). In the eastern U.S. (region 5) (Fig. 5.14a,b,c), the control members are always the best members in each group, with the RSM control member being the best member out of all 15 members (~12-16%). For the Atlantic (region 6) (Fig. 5.14d,e,f), the RSM control member is always best, especially at F03-F15 (~20%). For the most part, the groups perform fairly similarly. Since skill is not evenly distributed across all 15 ensemble members, the ensemble distribution may capture the true uncertainty, and the SREF mean forecast may suffer.

iii. Brier Score Decomposition

There are several approaches to quantify the probabilistic skill of an ensemble. One widely used skill score for categorical forecasts is the Brier score (BS), defined as (Wilks 1995):

$$BS = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2 \quad , \tag{5.1}$$

where n is the number of cases, p is the forecast probability, and o is the observed probability (1 if an event occurs, 0 if it doesn't). An ensemble with perfect probabilistic predictive skill would have a BS of 0, and one with no predictive skill would have a BS of 1. The BS is generally used to show how well an ensemble estimates the probability of a given variable (e.g., SLP) that falls below or above a certain threshold/category. For the purpose of this study, a slightly different approach was used to calculate BS. Instead of defining an event based on whether or not the cyclone will fall below a threshold SLP value, an event was defined based on whether or not the predicted cyclone central pressure is *within* a certain range X on either side of a selected SLP value, where X=1.5*E (E is the GFS MAE for SLP at that forecast hour and region). This approach was used since it is important to understand whether SREF can provide probabilistic skill for a SLP range within a representative error bar in the GFS model predictions.

The Brier score can also be decomposed into several components (Murphy 1973),

$$BS = REL - RES + UNC, \tag{5.2}$$

where,

$$REL = \frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2, \tag{5.3}$$

$$RES = \frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2, \tag{5.4}$$

$$UNC = \bar{o}(1 - \bar{o}). \tag{5.5}$$

The REL term summarizes the calibration of the ensemble, describing the mean squared difference between the ensemble's probability forecast, and the actual probability of that event occurring (Eq. 5.1). For a well-calibrated ensemble, the events with a given forecast probability of the event occurring (p_i) will be close to the actual percentage of those events that actually occur (o_i). A perfectly reliable ensemble will have a REL of 0, therefore contributing to a lower Brier score. The RES term summarizes the ensemble's ability to distinguish between events. It measures (Eq. 5.3) the average of the squared differences between the sampled climatologies of each discrete forecast probability the ensemble can produce ($\bar{o}_i$) and the overall climatology of the event ($\bar{o}$). In this manner, if the forecast probabilities sort the events so that each sample's observed frequency differs considerably from the overall climatology of the event, then the RES term will be large, again contributing to lowering the Brier score. Finally, then UNC term (Eq. 4) is a function of only the observations, with a minimum of 0 when the climatological probability [$\bar{o}$] is 0 or 1, and a maximum of 0.25 when $\bar{o}$ is 0.5. When an event is very likely, or very unlikely (eg. the tails of a distribution of events), there is little uncertainty in how the event will unfold. However, when the probability of an event occurring is ~0.5, then it is less certain how the event will unfold, and uncertainty will be higher. A smaller UNC leads to a smaller Brier score.

In order to test the improvement of the ensemble over another ensemble, deterministic model, or climatology, the Brier skill score (BSS) is defined as:

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

and

$$BSS_{c\,\mathrm{lim}} = 1 - \frac{BS}{UNC}, \tag{5.7}$$

where BS_{ref} is the reference BS, and BSS_{clim} is the BSS with respect to climatology. The reference BS is usually based on climatology, but it can be based on another ensemble or deterministic model as well (Weigel et al. 2007). A BSS of 1 would indicate that the ensemble gives a perfect forecast. A BSS of 0 would indicate that the ensemble shows no improvement over the reference score (deterministic model or climatology), while negative infinity shows that the ensemble has absolutely no skill.

While the BS is used mainly to evaluate ensembles, it was also applied to the NCEP NAM and GFS models, since the deterministic forecast distributions are essentially binary yes/no forecasts (1 or 0). Even though the BS score is designed to test how well an ensemble PDF resembles the true probability of an event occurring, the score can still be used to test the skill of deterministic models. By using the NAM and GFS as references in calculating the BSS, one can determine whether SREF has an advantage over the NAM and GFS for probabilistic forecasting.

Figure 5.15 shows the Brier score for cyclone central pressure, calculated every 5 mb and SLP range for each threshold (Table 5.2), for the SREF, GFS, NAM, and a GFS/NAM blend, for hours 27 through 63. The regions were combined (1 and 2, 3 and 4, and 5 and 6) in order to increase the dataset, and the range of the plots was limited to 980-1010 mb, since SLP thresholds outside of this range had less than 100 cases. For the GFS/NAM blend, p_k was calculated by treating the GFS and NAM together as a two-member ensemble. The relative predictive skill of the various models is similar for all regions. The Brier scores are nearly perfect (near 0) for relatively strong (~970 mb) and weak (~1015 mb) cyclones for reasons that will be discussed below. Regions 5 and 6 have the smallest Brier scores for all models, suggesting that predictability is higher in the eastern 3rd of the domain for the SREF, while regions 1 and 2 (eastern Pacific and western U.S.) have the least predictive skill. The NAM (green) has the largest Brier score for most thresholds in all regions, which suggests that it has the least skill. The GFS (blue) always has a lower Brier score than the NAM, but still larger than the SREF (red), which has the greatest predictive skill. However, the GFS/NAM blend (cyan) has a Brier score very close to the SREF's, suggesting that the GFS/NAM mini-ensemble can predict cyclone events nearly as well as the SREF. This can be quantified better using the BSS.

Figure 5.16 shows the BSS (Eq. 5.5) calculated using the GFS, NAM, the GFS/NAM blend, and climatology as a reference. Each line shows the BSS of the SREF using a different model as a reference. One result is how the SREF shows nearly the same level of improvement over the NAM as it does over climatology. The SREF shows less improvement over the GFS than over the NAM and climatology. Finally, it's clear that the BSS of the SREF is lowest when using the GFS/NAM blend as a reference. This BSS is nearly zero in all regions, confirming that the GFS/NAM blend has almost the same probabilistic skill as the SREF. In order to see what properties of the ensemble are contributing to the overall predictive skill shown by the BS, the components of the BS (Eq. 5.2-5.5) were calculated.

Figure 5.17 shows the SREF BS and its components (REL, RES, and UNC) calculated over a range of thresholds for hours 27 through 63. In all regions, REL is very low, suggesting that the SREF forecasts events with probabilities close to the true climatological probability of the events. Because UNC is solely a function of observations, the only other property of the SREF that can be adjusted to improve the Brier score is the RES term. According to Eq. 5.2, RES is inversely proportional to BS, which suggests that a low SREF RES is contributing to an increased BS. This would imply that when events are binned according to the SREF's forecast probability, each subsample's observed frequency is close to the overall climatological frequency of the event.

One problem with the BS as calculated above is that for relatively high (> 1005) and relatively low (< 990) thresholds, a large percentage of the events will fall outside the

1.5 MAE range, likely resulting in p_i and o_i both being 0%. In Eq. 5.1 this equates to a perfect sample of BS, and when iterating BS over all the events, the resulting score is very low, falsely indicating very good performance by the SREF. Operationally, a forecaster is interested in the SREF skill when an event is likely to occur, hence with p_i and/or o_i greater than 0. In order to better capture the SREF's ability to predict events that are more likely to occur, BS was recalculated using only events in which p_i and/or o_i were greater than 0. Figure 5.18 shows the recalculated BS for the SREF, GFS, NAM, and the GFS/NAM blend. With this new method, the BS is much more uniform across the range of cyclone depths. The SREF BS is very similar across all regions, perhaps slightly higher in regions 3 and 4 than other regions. Nevertheless, the relative scores of all models are the same as with the original calculation of BS. The NAM still has the lowest skill, followed by the GFS, in all regions. Again, the skill of the GFS/NAM blend is very close to the skill of the SREF. With many more members than the GFS/NAM blend, the SREF should have a lower Brier score than the GFS/NAM blend, but is likely suffering from poor calibration and/or resolution of events.

In order to visually summarize how each component is aiding or harming the SREF, reliability diagrams were created (Fig. 5.19), which show the observed frequency of events as a function of forecast probability. First, for each region, only events in which at least 12 out of 15 members had a cyclone were kept. For each of these events, only the top 12 members were kept. Then, all events were binned into 7 discrete forecast probabilities from the SREF (0, 1/12 or 2/12, 3/12 or 4/12…11/12 or 1), instead of 13 (0, 1/12, 2/12, 3/12….1), to increase the sample size of each bin. Finally, the frequency of events observed within each bin was plotted as a function of that bin's forecast probability. The solid 1:1 line in figure 5.19 indicates perfect reliability, the sloped dashed line indicates no skill, and the horizontal dashed line indicates no resolution (climatology). Probabilities are defined as whether or not a cyclone central pressure is/will be within 1.5*X mb of either an average or deep cyclone (Table 5.2) for each region, where X is the GFS MAE for each region. A deep cyclone has a central pressure more than 1.5 standard deviations below the mean central pressure for each region.

In regions 1 and 2, the SREF for average cyclones is somewhat overconfident (Fig. 5.19a) at higher probabilities (> 0.7), since the reliability has a smaller slope than 1:1. For example, in Fig. 5.19a, during events for which the SREF gives a 100% chance of the cyclone being of average intensity, only ~80% verify. The same holds true for regions 3-6 (Fig. 5.19c,e), where the SREF shows the same signs of overconfidence when forecasting average intensity cyclones. Also, for regions 5 and 6, since the line is positioned nearly entirely on one side, there is a bias toward higher forecast probabilities (overprediction of cyclones). This means that when the SREF gives a probability of a cyclone being of average intensity, the true observed frequency is less. When the SREF predicts a relatively deep cyclone, there is an unconditional bias toward higher forecast probabilities (overforecasting) in all regions (Fig. 5.19b,d,f). In regions 1 and 2, the SREF has less reliability for forecasting deep cyclones than average cyclones, but in regions 3-6, the SREF actually shows more reliability for deeper cyclones.

iv. CRPS

In recent years, the use of the continuous ranked probability score (CRPS) to verify the probabilistic skill of ensembles has become increasingly popular. The CRPS is basically an integration of BS over all possible thresholds, yielding the equation:

$$CRPS = \int_{-\infty}^{\infty} BS(x_t)dx_t, \tag{5.8}$$

where x_t is the threshold used to calculate BS (Eq. 5.1). An ensemble with perfect probabilistic predictive skill would have a CRPS score of 0, and one with no predictive skill would have a CRPS of 1. The benefit of using CRPS is that it doesn't require an arbitrary threshold to be chosen to calculate the score, as it is integrated over the entire distribution of x.

Figure 5.20 Shows CRPS versus forecast hour for the SREF, GFS, NAM, and the GFS/NAM blend. By F63, the CRPS is highest in the eastern U.S. and western Atlantic (regions 5 and 6) (lowest skill), and lowest in the central U.S. and central and eastern Canada (regions 3 and 4) (highest skill), though these differences across regions are not statistically significant at the 90% level (not shown). Like the BS, the CRPS shows that the NAM has the least skill in all regions, at all forecast hours. The GFS is significantly more skillful than the NAM at all forecast hours, in all regions. The SREF has the most skill, but again the GFS/NAM blend is nearly as skillful, especially before F33 in all regions. These results confirm that while SREF has more skill than the GFS and especially the NAM, the simple GFS/NAM blend proves to be a useful alternative to the SREF with almost no loss in skill.

v. Case Studies

In order to understand how the average SREF errors can get so large for cyclone events, a representative cyclone case was analyzed in more detail. The cyclone formed in southeast Texas late on 01 February 2006 (not shown), and the SREF run investigated was initialized on 09 UTC 02 February 2006. The corresponding Hydrometeorological Prediction Center's (HPC's) surface analysis at 09 UTC illustrates a weak cyclone over western Louisiana of about 999 mb (Fig. 5.21), as a shortwave trough moved eastward across the southeast U.S. (Fig. 5.22). Over the next 24 hours the cyclone quickly moved northeast across the Mississippi and Ohio Value, and was located over Lake Erie at 1200 UTC 3 February (Fig. 5.23 and 5.24).

Figure 5.25 shows the SREF forecast of the cyclone track, including EBM members (red), EKF members (purple), RSM members (yellow), and the GFS (blue) and NAM (green). The dashed black line represents the observed (blend of NAM and GFS analysis) cyclone track. The 15-member SREF mean cyclone track is relatively close to the observed cyclone track. The mean absolute displacement of the SREF mean is 196 km, while the mean absolute displacement of the GFS and NAM tracks are 211 km and 260 km, respectively. Even though the SREF mean provides a good forecast of the cyclone track, some of the tracks have larger errors and are clustered for a particular modeling system. For example, 4 out of 5 RSM members have a consistent large eastward displacement error of about 200 km.

Even though the cyclone track is well forecast by the SREF system, the cyclone central pressures are poorly forecast (Fig. 5.26). The cyclone was initialized at 09 UTC 2 February at ~1000 mb for all members as compared to 999 mb in the HPC analysis (Fig. 5.21). At 12 UTC 2 February, the GFS and NAM are initialized at 999.5 and 998, respectively. From F00 to F03 (09 UTC to 12 UTC) the SREF deepens the cyclone too quickly (1 to 7 mb in 3 hours) compared to the actual cyclone central pressure change (~1 mb in 3 hours), which suggests a SREF spin-up problem for some members. For example, the RSM members only deepen the cyclone by 1-2 mb during the first 3 hours, while the EBM and EKF members deepened the cyclone by up to 7 mb in 3 hours. In contrast, the operational GFS run at 1200 UTC deepens the cyclone in the first 6 hours by only 0.2 mb in 3 hours, while the NAM weakens the cyclone by 1.5 mb over 3 hours. The operational GFS and NAM produce a more accurate cyclone intensification early in the forecast.

The poor forecast of central pressure is evident in figure 5.26, which shows the forecast central pressure in the GFS, NAM, and SREF during the cyclone event, and table 5.3 and 5.4, which show forecast and observed cyclone central pressure, deepening rate, and forecast errors for this case. The cyclone was initialized at 09 UTC at ~1000 mb in all members (column 2 of table 5.3), which was generally within 1-2 mb of the observed cyclone (999 mb). However, in the first 3 hours of the forecast, all members deepened the cyclone (some by more than 7 mb in 3 hours), even though the HPC analyses indicated no deepening over that 3-hour period (columns 3 and 4 of table 5.3). This resulted in errors approaching -7 mb at F03 (column 6 of table 5.3). Clearly, the SREF cyclone deepening rate is much too high here. In fact, with very few exceptions, the observed central pressure lies completely outside of the forecast distribution throughout the entire run (Fig. 5.26). Table 5.4 shows the same as table 5.3, except for the GFS and NAM runs at 06 UTC. Over the same 3-hour period, the SLP trend forecast by the GFS and NAM was -1.1 and -1.3 mb per 3 hours, respectively, with an SLP error of 0.5 and 0.1 mb, respectively, at F06. The GFS and NAM were much better at forecasting the trend in cyclone central pressure than the SREF early in the run, which explains why they remained much closer to the observed central pressure throughout the forecast (Fig. 5.26). The mean central pressure error for the 15-member SREF mean is -4.3 mb, while the error for the NAM and GFS are -1.3 mb and -3.5, respectively. For this cyclone event, the large growth of central pressure errors in the SREF early in the forecast gave the SREF an early disadvantage, which continued throughout the forecast.

Another cyclone event involving large errors in the SREF impacted the East Coast on 15-17 December 2005. The surface cyclone formed along the Gulf Coast near 09 UTC 15 December. An HPC surface analysis at 12 UTC the 15 December shows a developing cyclone over the southeast U.S., with a central pressure of about 1003 mb (Fig. 5.27). The cyclone developed in response to a short-wave trough moving over the southeast U.S. as a broad closed longwave trough progressed eastward toward the Great Lakes (Fig. 5.28). During the next 24 hours, the cyclone moved quickly up the coast as a coastal front set up in the Carolinas. At 12 UTC 16 December, the cyclone was centered over New York City with a central pressure of ~999 mb (Fig. 5.30).

Figure 5.31 shows the observed and SREF forecast tracks (see Fig. 5.25) for the SREF run initialized on 09 UTC the 15th. There is a westward bias for all SREF members, particularly the RSM members, which are displaced to the west by more than

250 km at times. The ETA members are displaced less than the RSM members by 100-200 km, but the ETA members still have a strong tendency to remain west of the observed track. The 15-member mean forecast track is ~100km west of the observed track. In this particular event, there was freezing rain over much of North and South Carolina, so a 100 km track error had a significant impact on the forecast for some regions. It is hypothesized that a poorly positioned cyclone at F00 may have led to larger displacement errors in the SREF (see F03 position over Alabama in Fig. 5.27). The NAM and GFS both performed better than the SREF in this case, with ~70 km displacement on average.

The cyclone central pressure was also poorly forecast for this December 2005 event (Fig. 5.32). While at F03 the cyclone central pressure falls within the SREF central pressure forecast envelope of all members, all members quickly overdeepen the cyclone, such that by F15 all SREF members are too deep. From F12-F24, the cyclone deepened rapidly from ~1006 mb to ~1000 mb. The EBM members (red lines) underdeepen the cyclone by nearly 3 mb, while the RSM members overdeepen the cyclone by up to 9 mb. The cyclone began to fill after F24, and all members until F48 are too deep. By F54, nearly all members are too weak. Through F48, the 15-member SREF mean overdeepens the cyclone by an average of 3 mb.

Overall, these cases show that initial errors in the cyclone position and central pressure can lead to large errors in the forecast. Also, during periods of intense cyclogenesis or cyclolysis, errors can grow very rapidly. Lastly, many times an entire group of members will tend to have the same bias in position and/or central pressure, making it less likely that the observed cyclone position or central pressure will fall within the range of SREF member forecasts.

Fhr	F03-F15						F27-F39						F51-F63					
Reg.	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
EBM	30	47	39	40	36	34	26	41	35	37	38	34	29	40	35	35	32	32
EKF	23	27	24	24	31	25	22	32	29	29	31	31	27	31	33	25	30	35
RSM	47	26	37	36	33	41	52	27	36	34	31	35	44	29	32	40	38	33

Table 5.1. Percent of time each of the three SREF groups (EBM, EKF, and RSM) is best for each region and forecast periods 3-15 h, 27-39 h, and 51-63 h.

	F27	*F33*	*F39*	*F45*	*F51*	*F57*	*F63*
Reg. 1-2	2.7	3.2	3.7	3.9	4.5	4.8	5.3
Reg. 3-4	2.3	2.5	2.8	3.1	3.4	3.8	4.0
Reg. 5-6	2.2	2.7	2.9	3.2	3.6	4.1	4.4

Table 5.2. Window sizes (in mb) used to calculate BS, REL, RES, and UNC. To calculate these scores, it was determined whether a cyclone fell within (above or below) X mb of each SLP value, where X is the SLP value for a particular region and forecast hour shown here (See section 5.d.ii).

(Init. 09 Z)	**F00 SLP (mb)**	**F03 SLP (mb)**	**SLP Trend (mb/3hr)**	**F00 Error (mb)**	**F03 Error (mb)**
EBM.CTL1	1000.7	995.5	-5.2	1.7	-3.5
EBM.N1	1000.5	996.9	-3.6	1.5	-2.1
EBM.N3	1000.1	996.9	-3.2	1.1	-2.1
EBM.P1	999.9	992.7	-7.2	0.9	-6.3
EBM.P3	1000	992.8	-7.2	1	-6.2
EKF.CTL2	1000.7	995.5	-5.2	1.7	-3.5
EKF.N2	999.5	992.3	-7.2	0.5	-6.7
EKF.N4	999.5	992.4	-7.1	0.5	-6.6
EKF.P2	1001.5	997.3	-4.2	2.5	-1.7
EKF.P4	1001.5	997.4	-4.1	2.5	-1.6
RSM.CTL1	999.9	998.6	-1.3	0.9	-0.4
RSM.N1	1000.8	998.9	-1.9	1.8	-0.1
RSM.N2	1000.6	998.8	-1.8	1.6	-0.2
RSM.P1	998.8	997.3	-1.5	-0.2	-1.7
RSM.P2	998.7	995.9	-2.8	-0.3	-3.1
OBS	**09 UTC SLP (mb)**	**12 UTC SLP (mb)**	**Trend (mb/3hr)**		
	999	999	0		

Table 5.3. Shows the 09 UTC SREF run F00 and F03 cyclone central pressure, F00 and F03 cyclone central pressure error (with HPC surface analysis as observed), and the SLP trend during first 3 hours.

(Init. 06 Z)	**F00 SLP (mb)**	**F06 SLP (mb)**	**SLP Trend (mb/3hr)**	**F00 Error (mb)**	**F06 Error (mb)**
GFS	1001.7	999.5	-1.1	0.7	0.5
NAM	1001.7	999.1	-1.3	0.7	0.1
OBS	**06 UTC SLP (mb)**	**12 UTC SLP (mb)**	**Trend (mb/3hr)**		
	1001	999	-1		

Table 5.4. Same as table 5.3, except for the 12 UTC deterministic GFS and NAM runs.

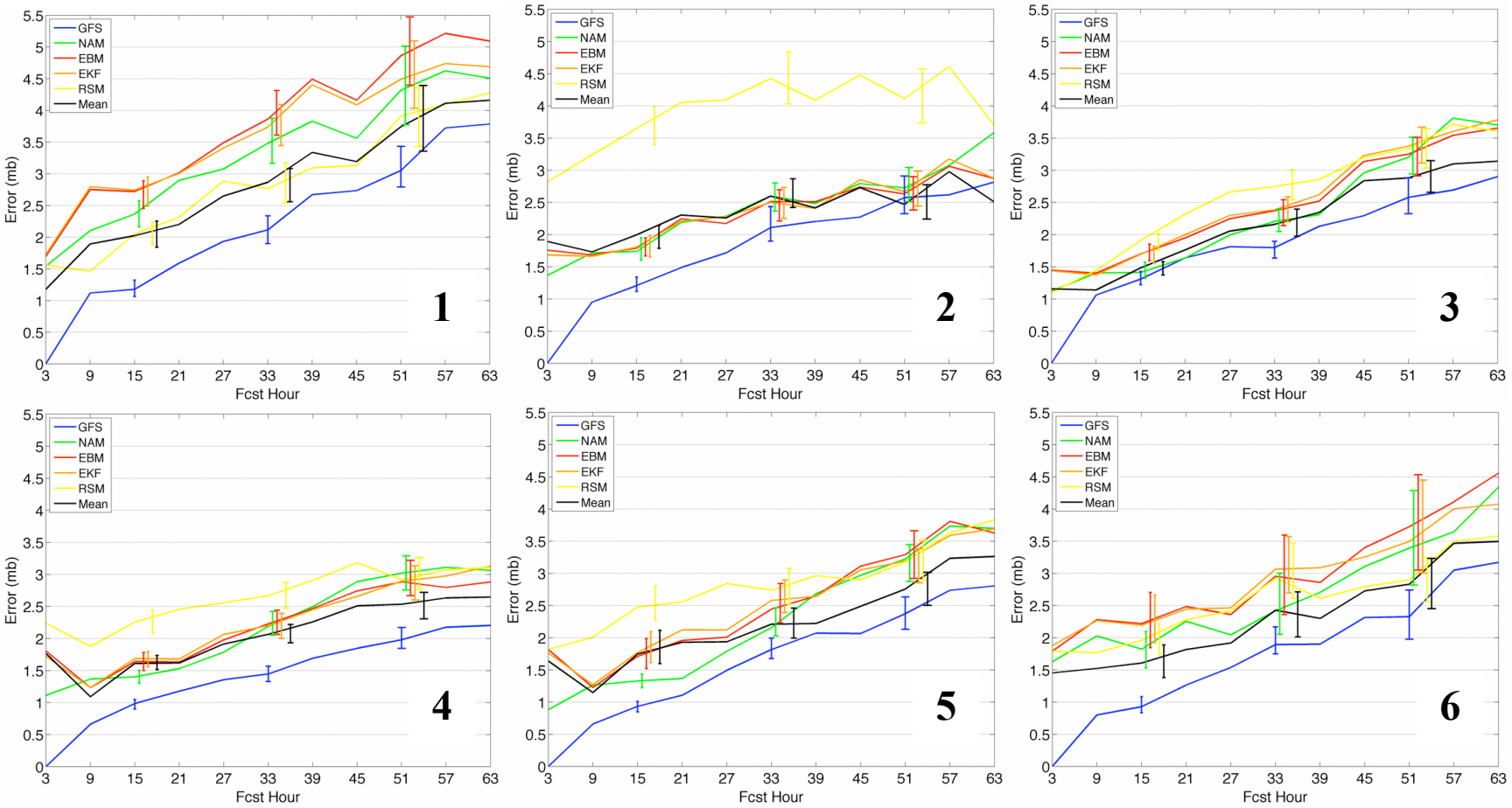

Figure 5.1. Central pressure mean absolute error (in mb) versus forecast hour averaged for the 2004-2007 cool seasons for the GFS, NAM, and SREF (and its subcomponents) given by the colored lines for each region (1-6).

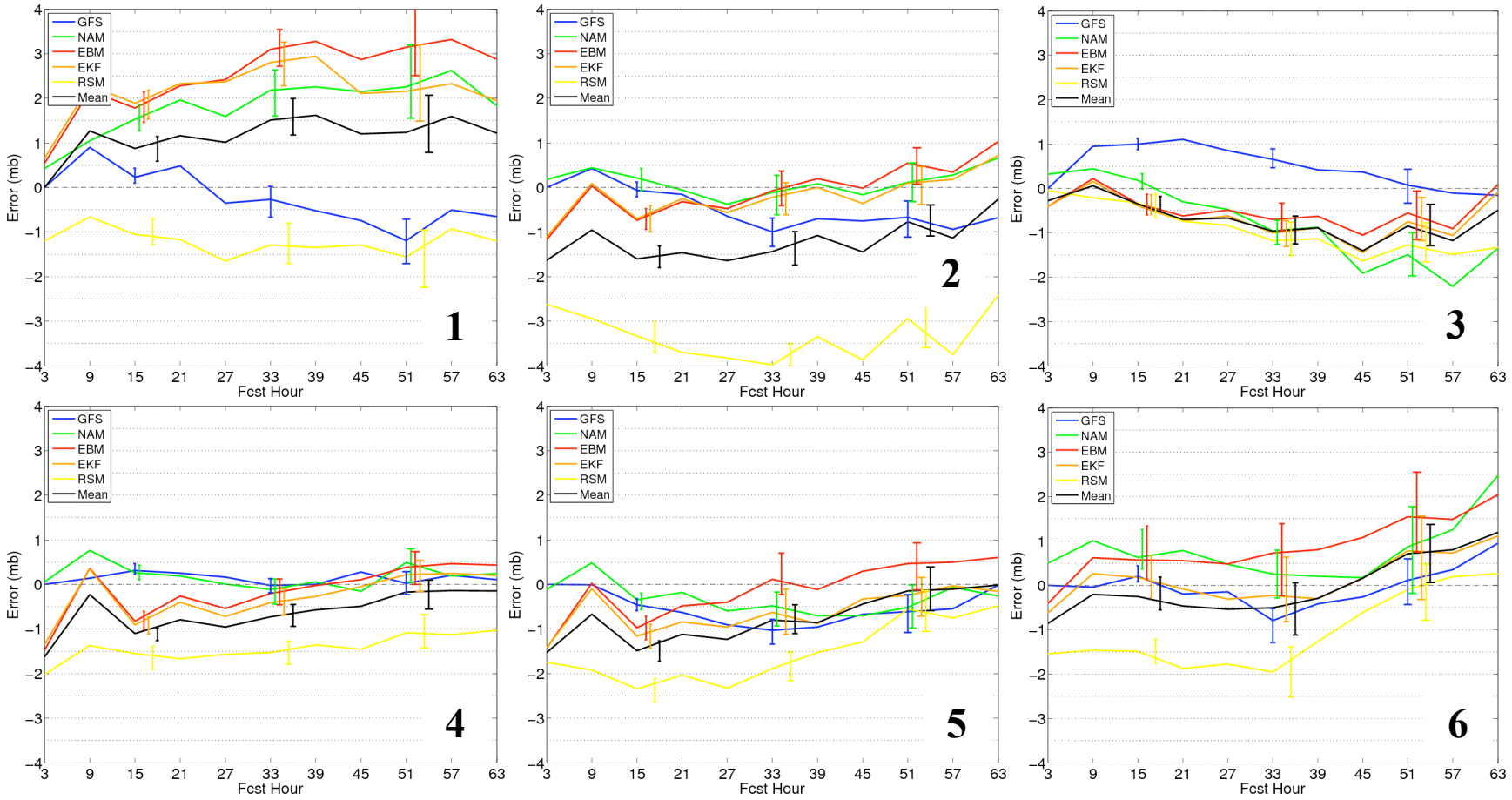

Figure 5.2. Same as Fig. 5.1 except for cyclone central pressure bias (in mb).

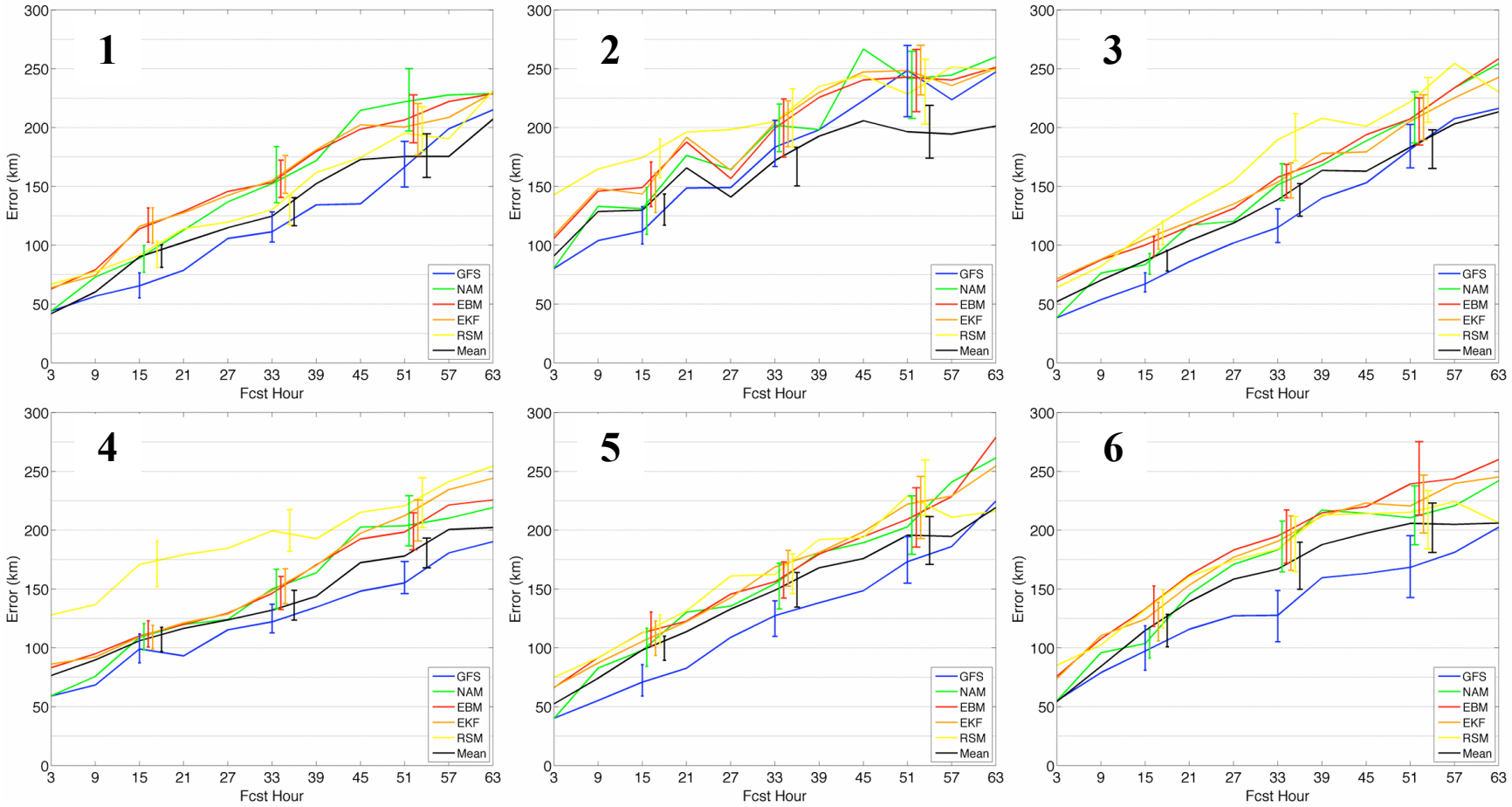

Figure 5.3. Same as Fig. 5.1 except for cyclone central displacement error (in km).

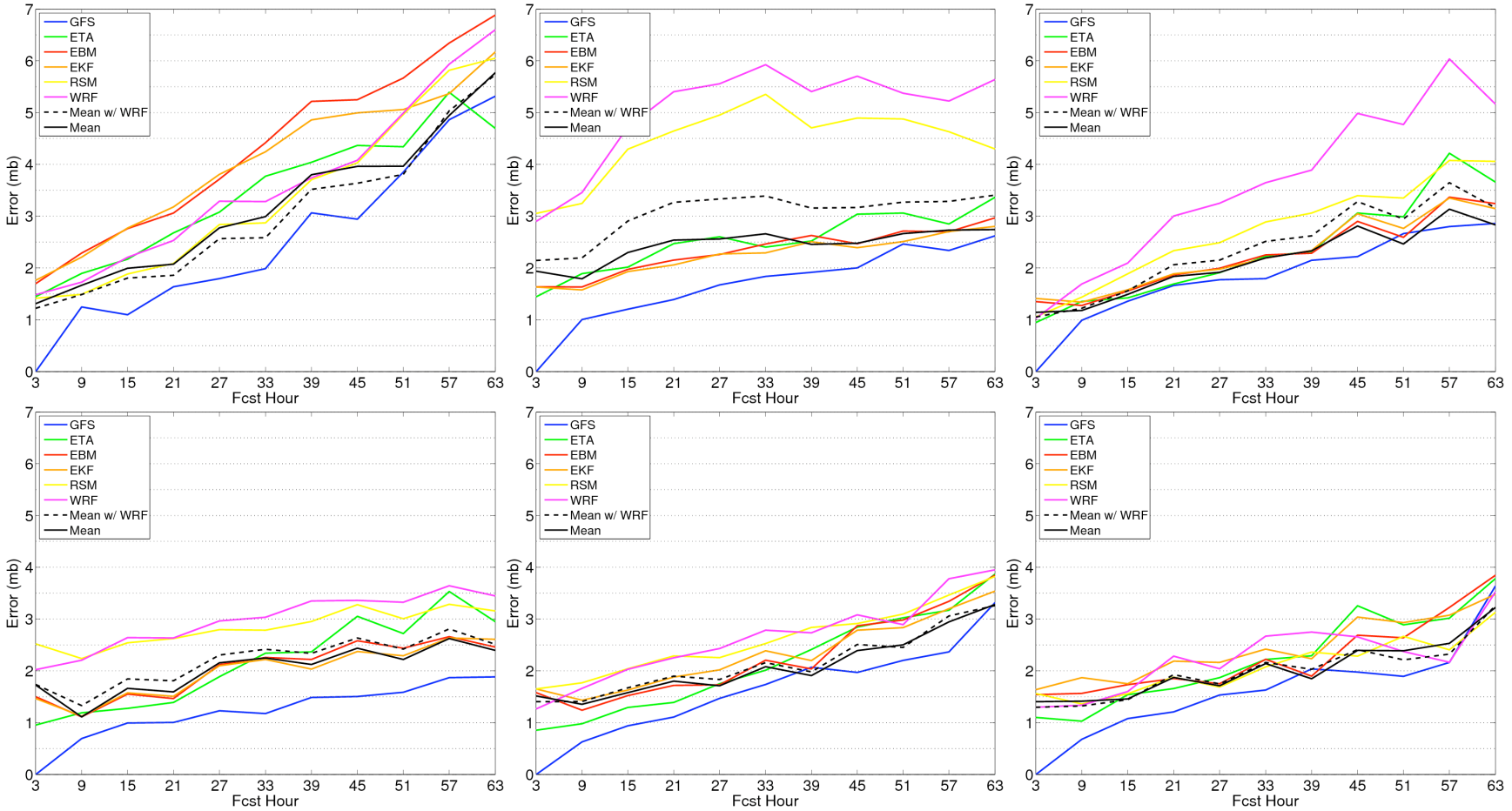

Figure 5.4. Central pressure mean absolute error (in mb) versus forecast hour averaged for the 2006-2007 cool season for the GFS, NAM, and SREF (and its subcomponents, including 6 WRF members) given by the colored lines for each region (1-6).

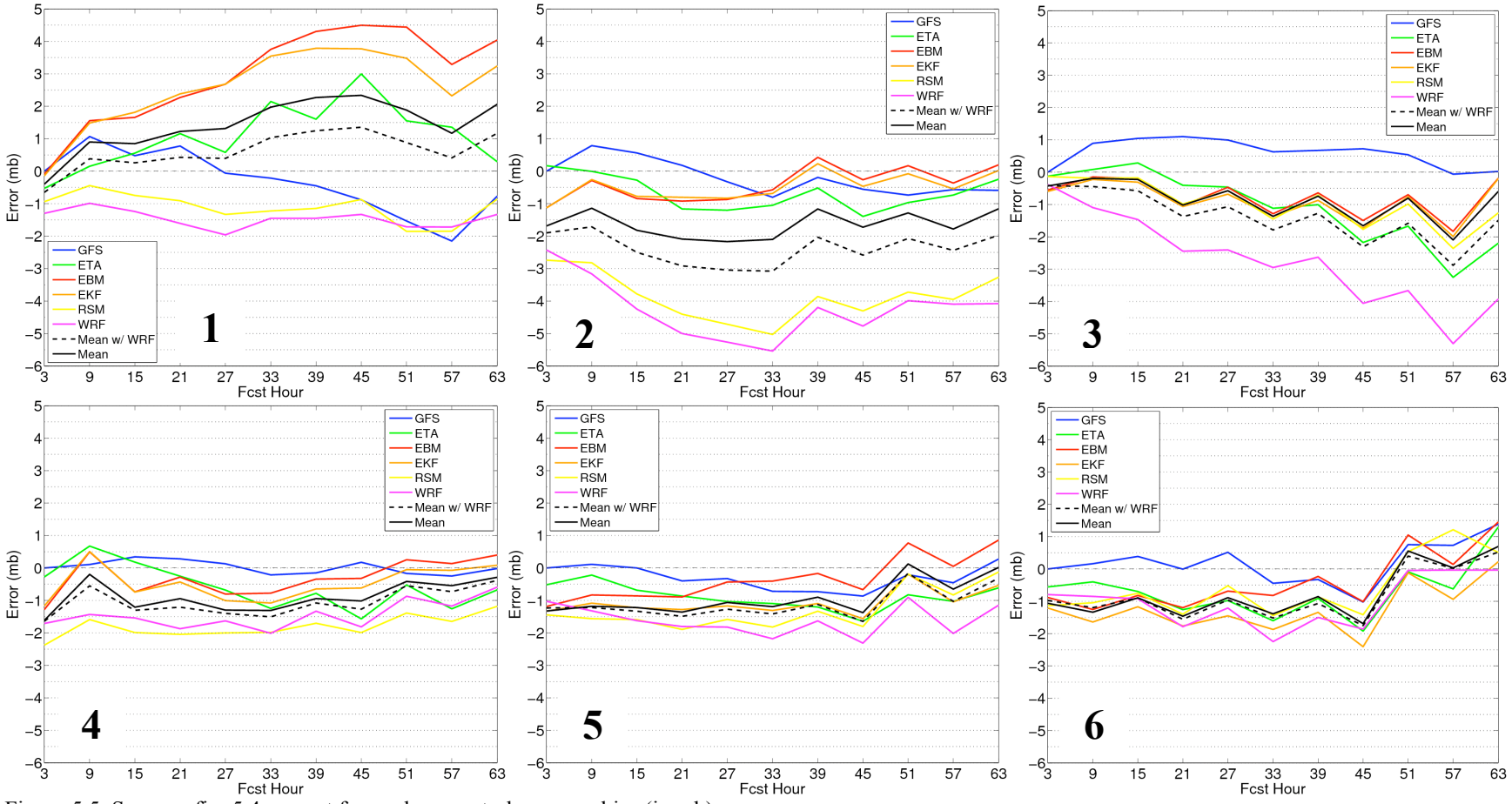

Figure 5.5. Same as fig. 5.4, except for cyclone central pressure bias (in mb).

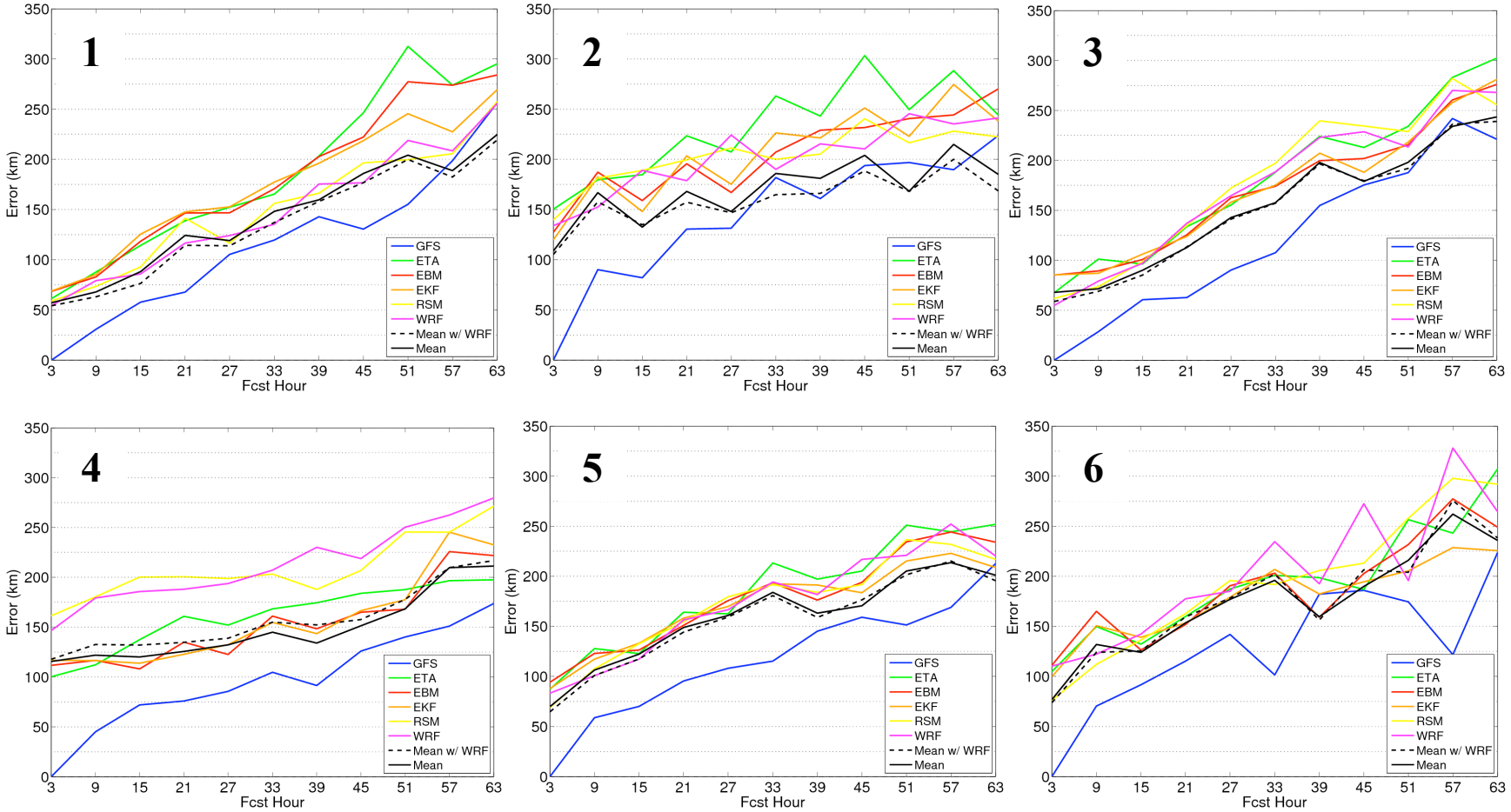

Figure 5.6. Same as fig. 5.4, except for cyclone central displacement error (in km).

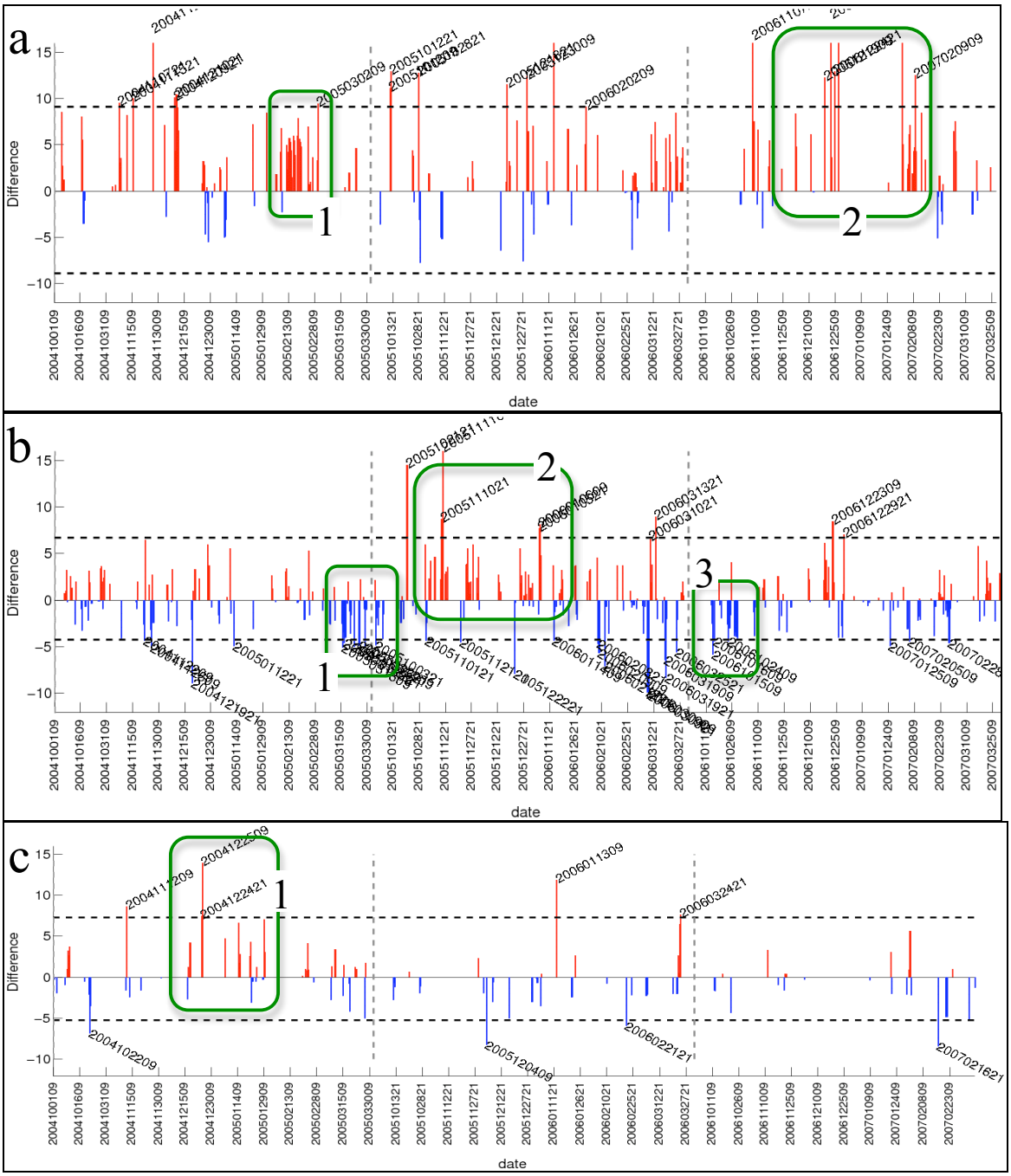

Figure 5.7. Time series of SREF cyclone central pressure error minus GFS cyclone central pressure error, at hour 48 for (a) region 1, (b) region 4, and (c) region 6, during the 2004-2007 cool seasons (separated by gray vertical lines). Horizontal black dashed lines denote 2 standard deviations above and below the mean SREF/GFS error difference for each region, with differences exceeding this threshold highlighted by the model initialization date. The solid black line is the 30-day running mean of cyclone central pressure errors. Interesting periods for text discussion are highlighted with green boxes and numbered.

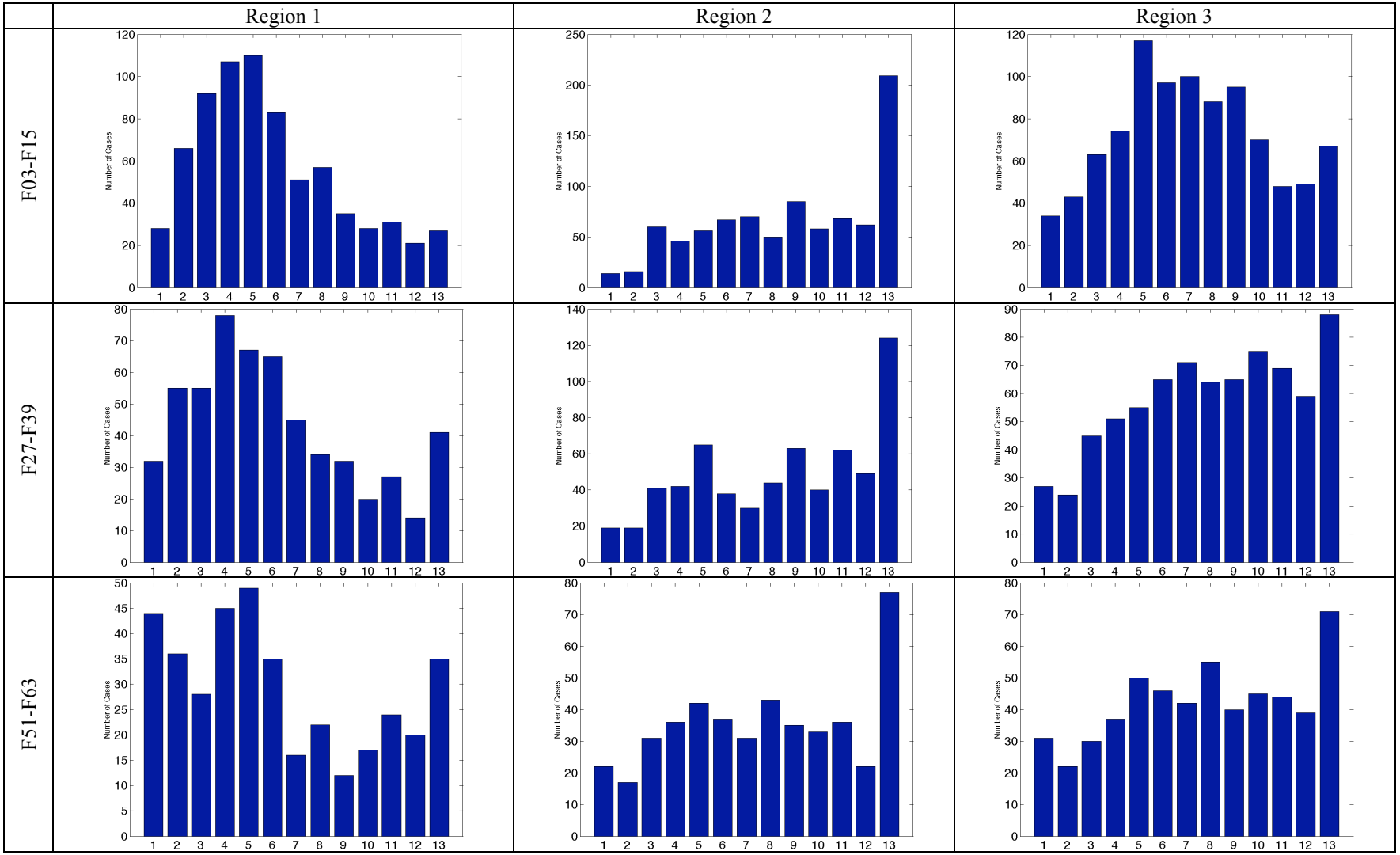

Figure 5.8. SREF rank histograms of cyclone central pressure for all regions for hours 3-15, 27-39, and 51-63.

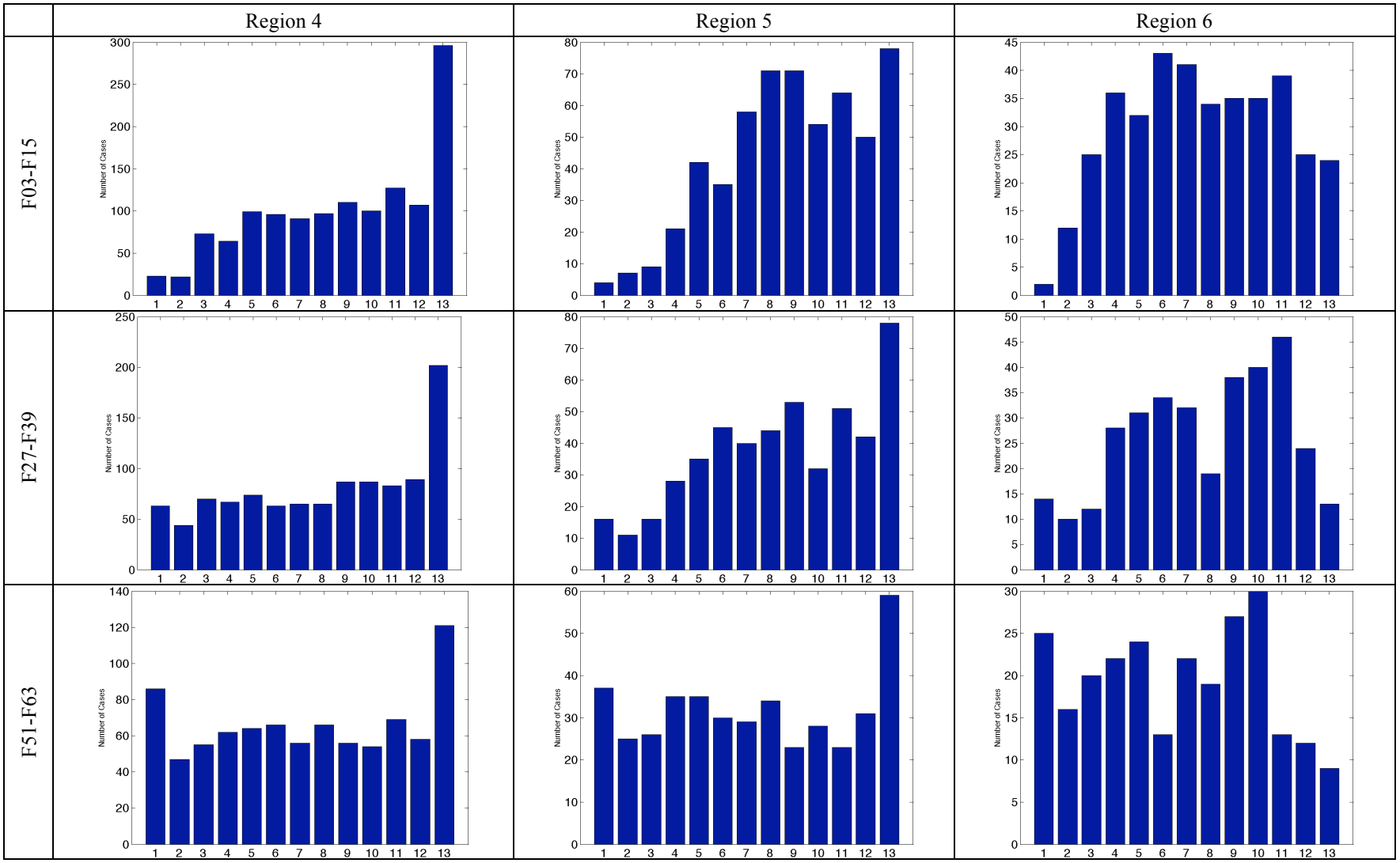

Figure 5.8 continued.

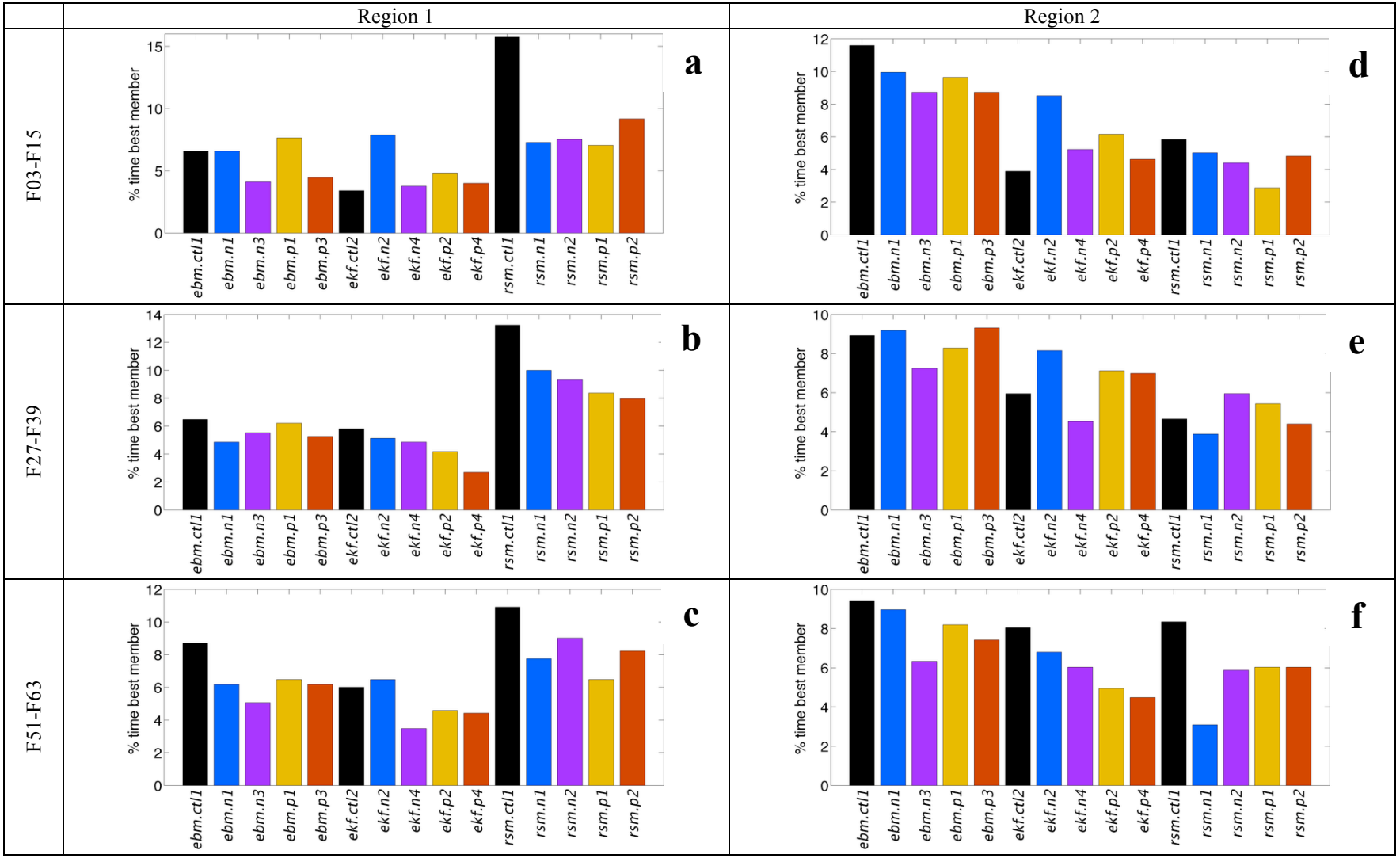

Figure 5.9. Histograms showing the percent time each member in SREF is best, when forecasting cyclone central pressure, for regions 1 (left) and 2 (right).

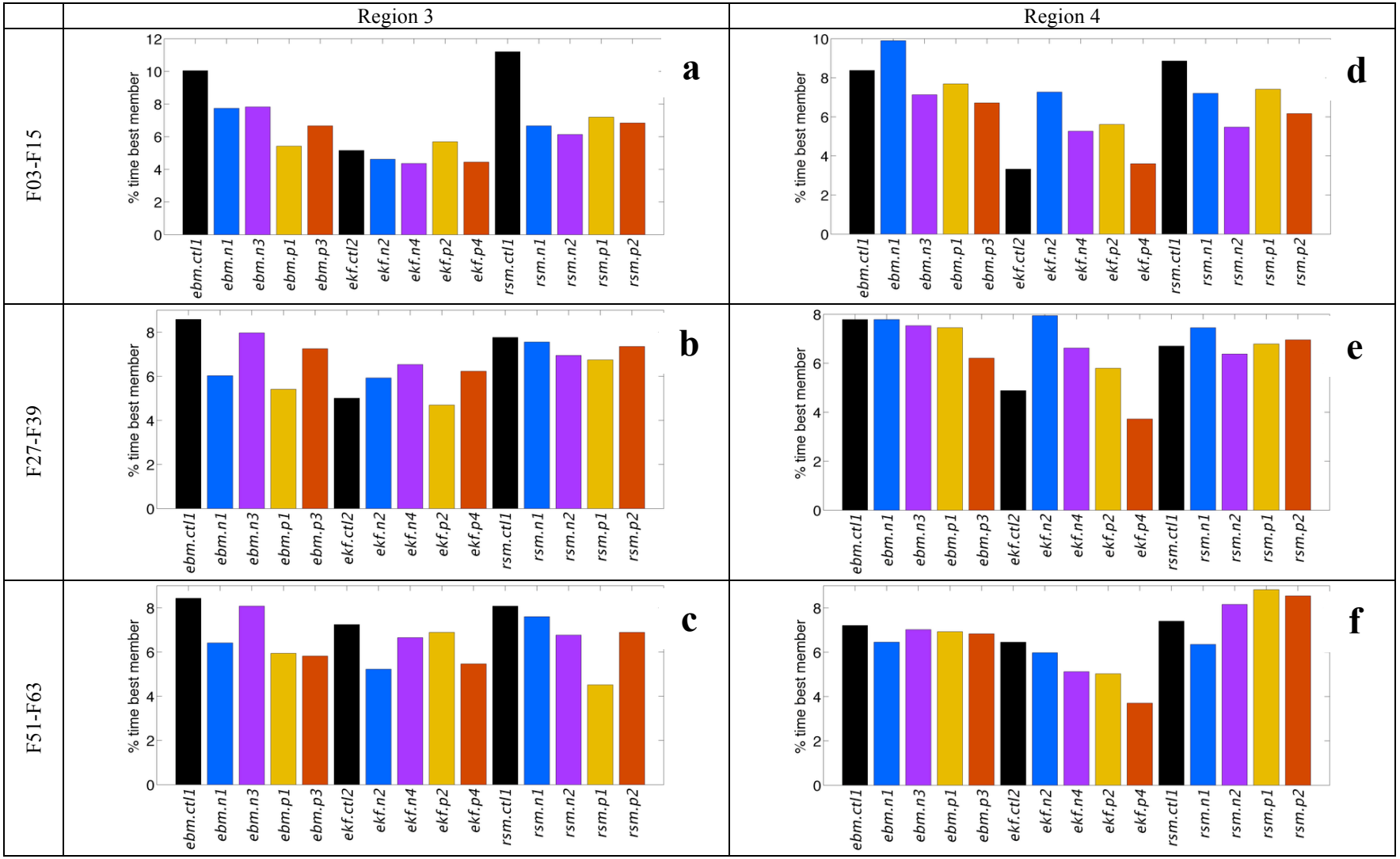

Figure 5.10. Same as Fig. 5.9 except for regions 3 and 4.

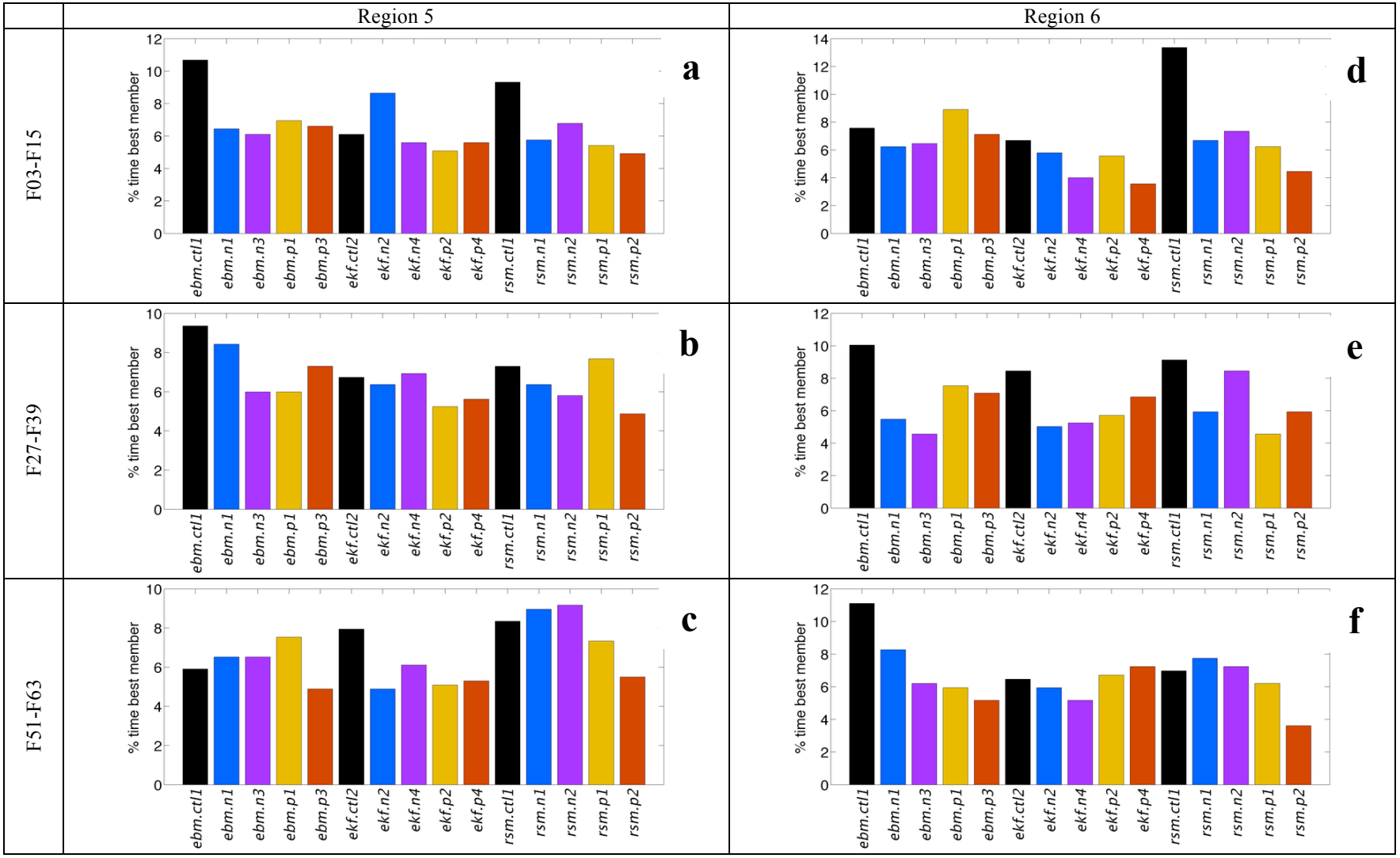

Figure 5.11. Same as Fig. 5.9 except for regions 5 and 6.

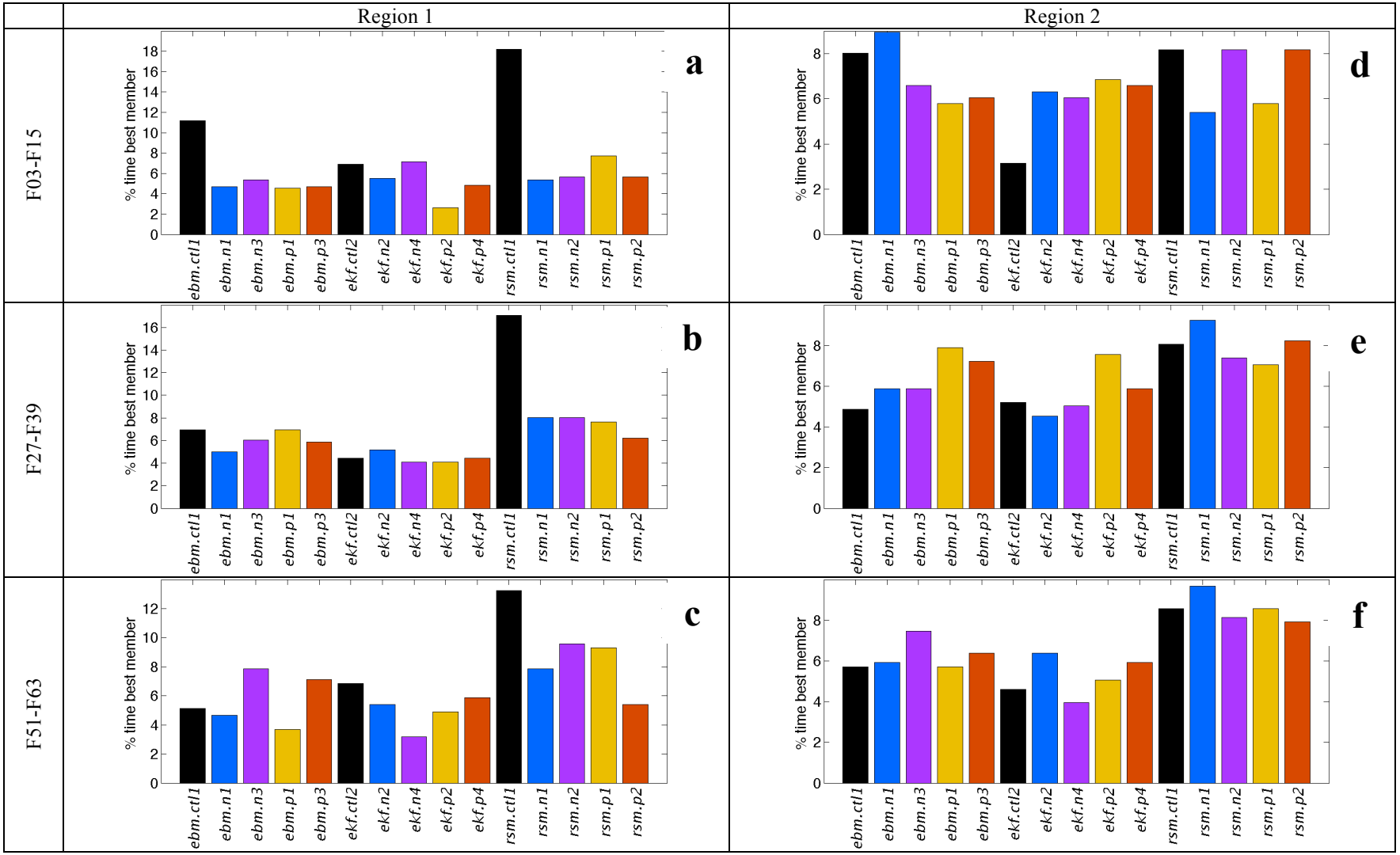

Figure 5.12. Histograms showing the percent time each member in SREF is best, when forecasting cyclone position, for regions 1 (left) and 2 (right).

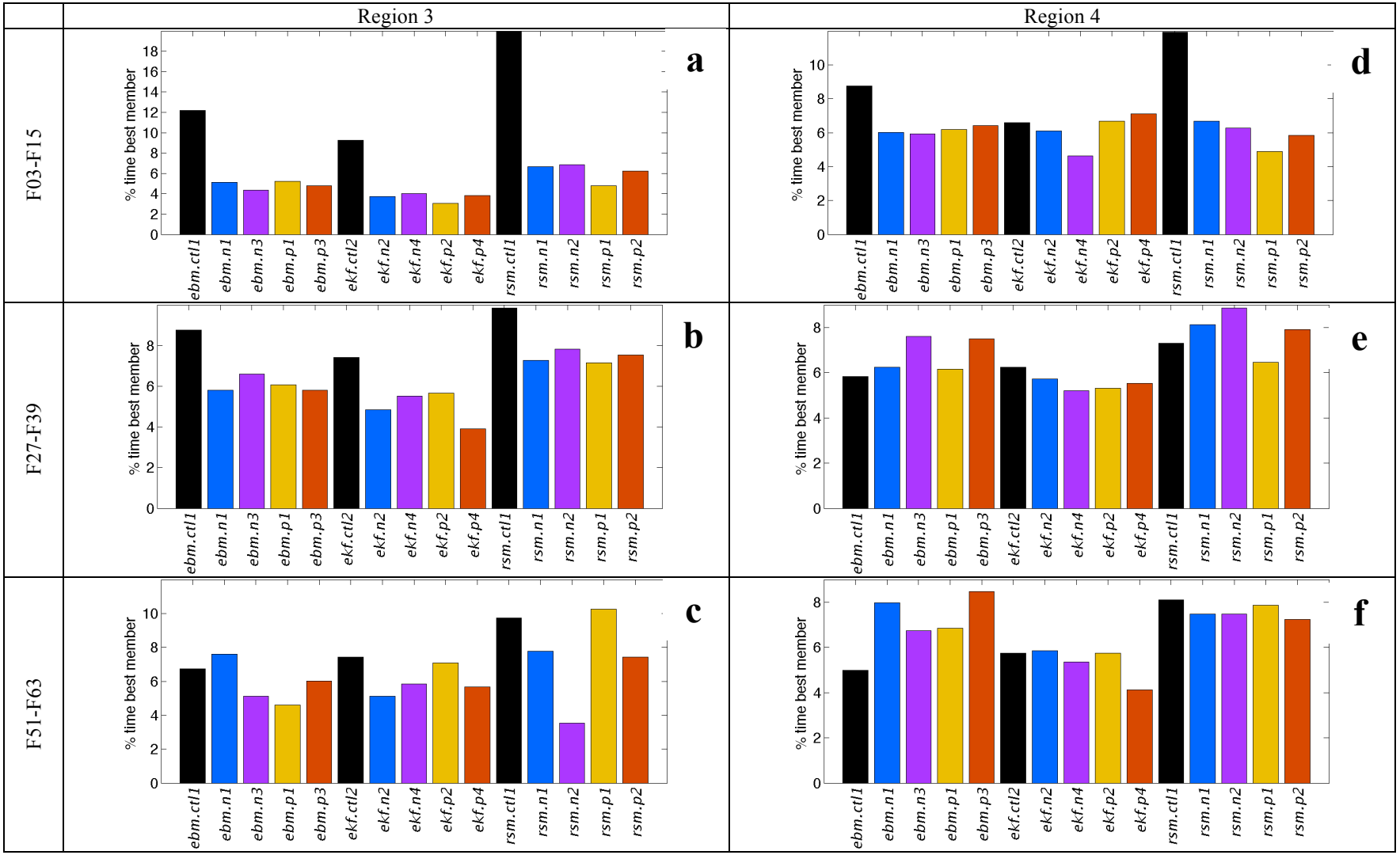

Figure 5.13. Same as Fig. 5.12 except for regions 3 and 4.

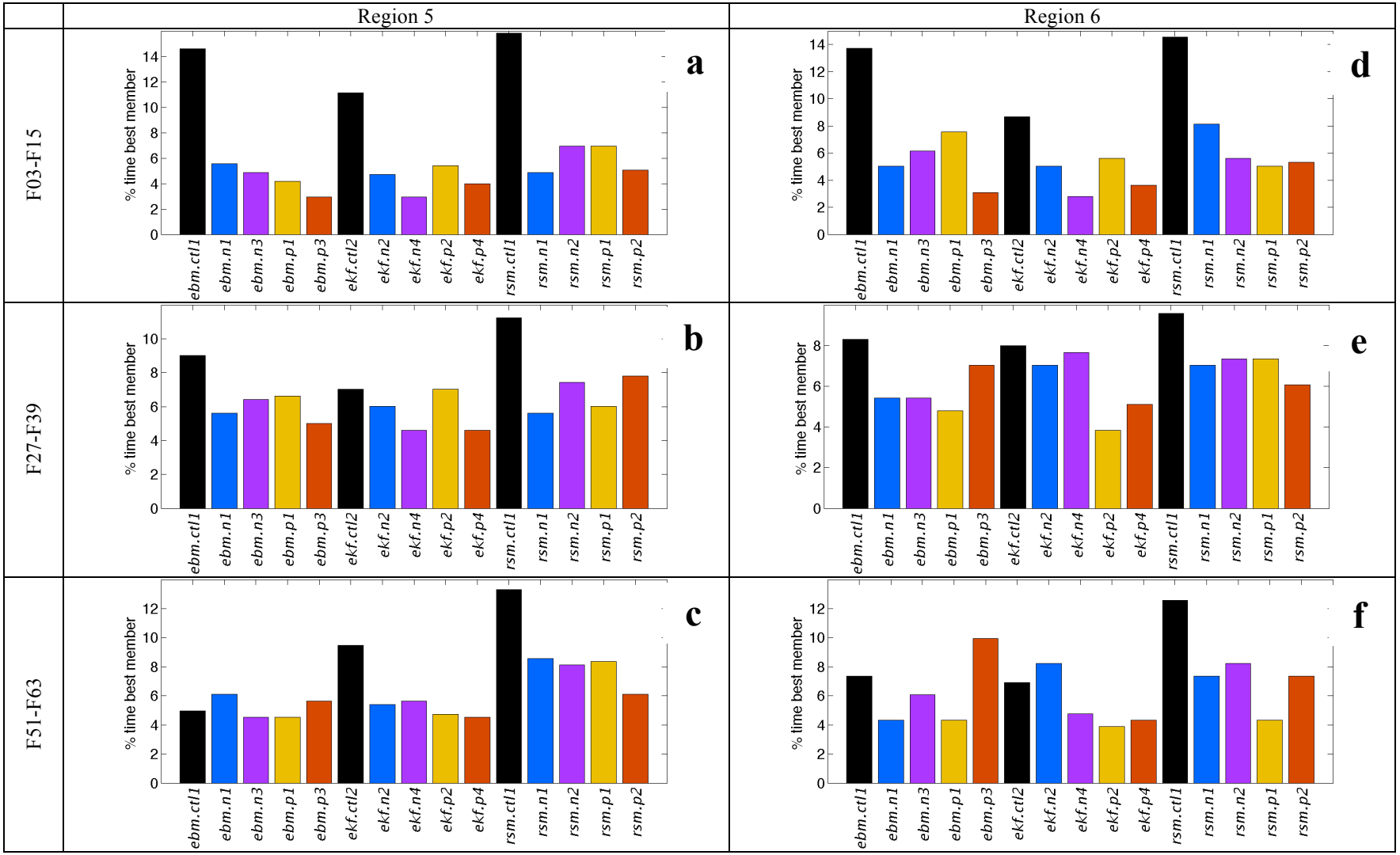

Figure 5.14. Same as Fig. 5.12 except for regions 5 and 6.

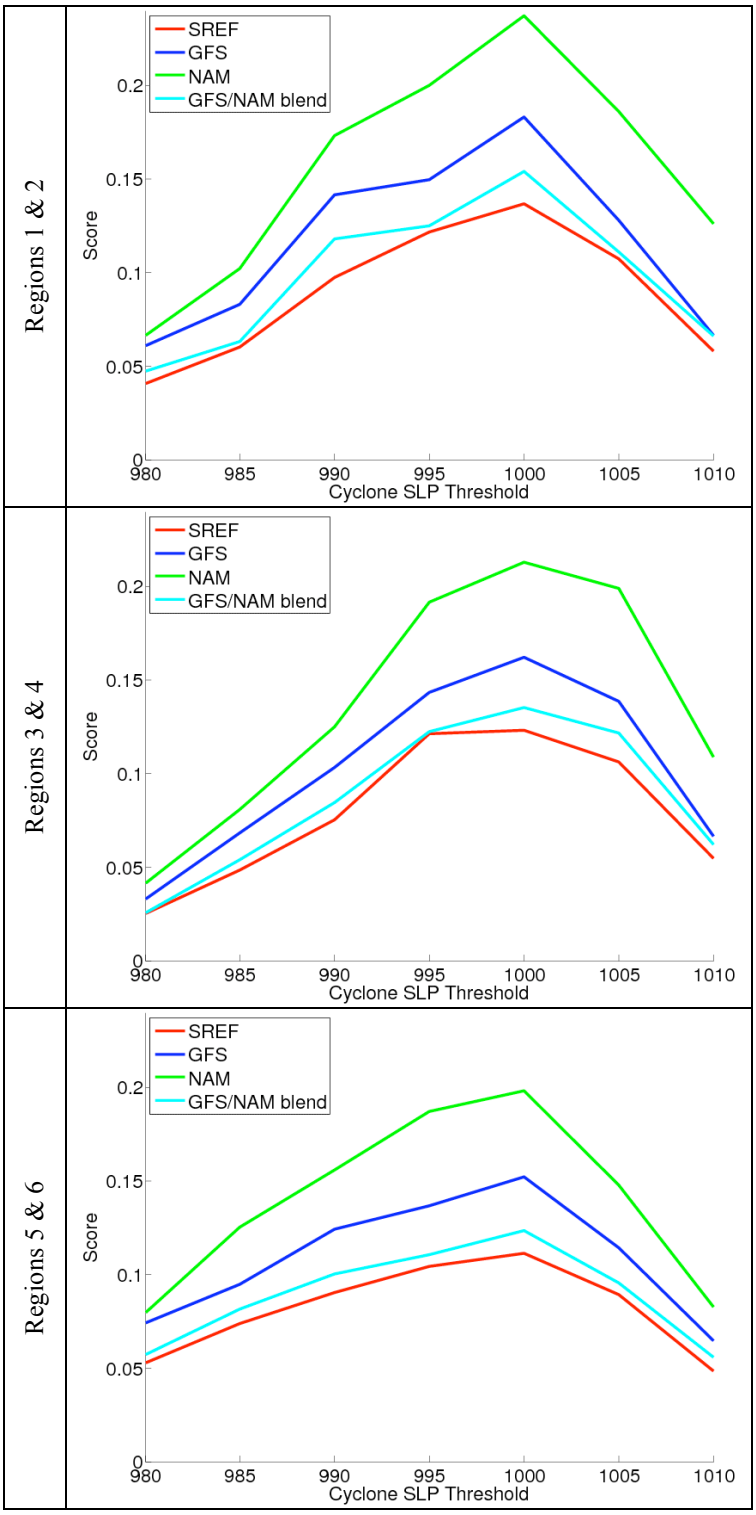

Figure 5.15. Brier score calculated over a range of thresholds (see Eq. 1) for the SREF, GFS, NAM, and GFS/NAM blend for F27-F63.

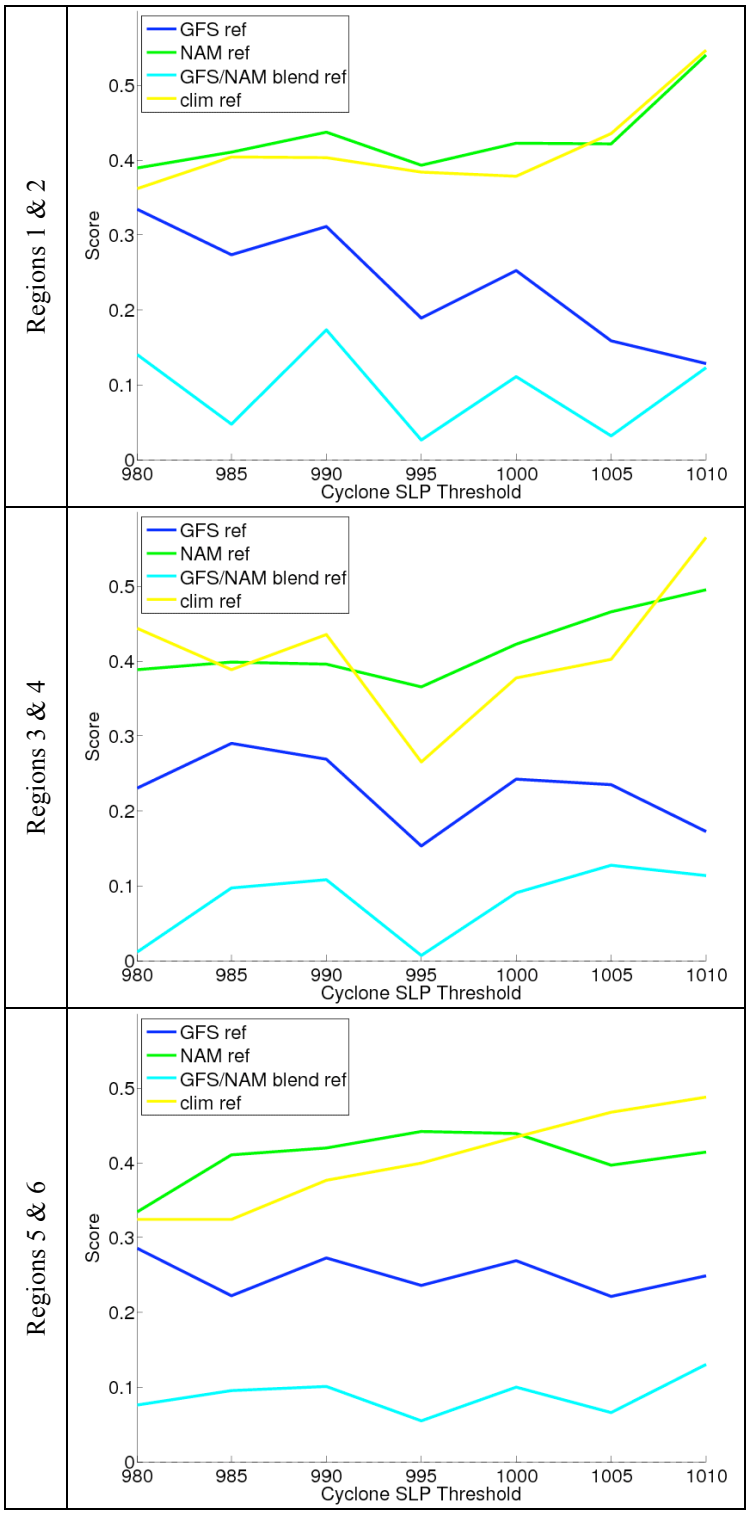

Figure 5.16. Brier skill score (Eq. 6), calculated over a range of thresholds for the SREF, GFS, NAM, and GFS/NAM blend for F27-F63.

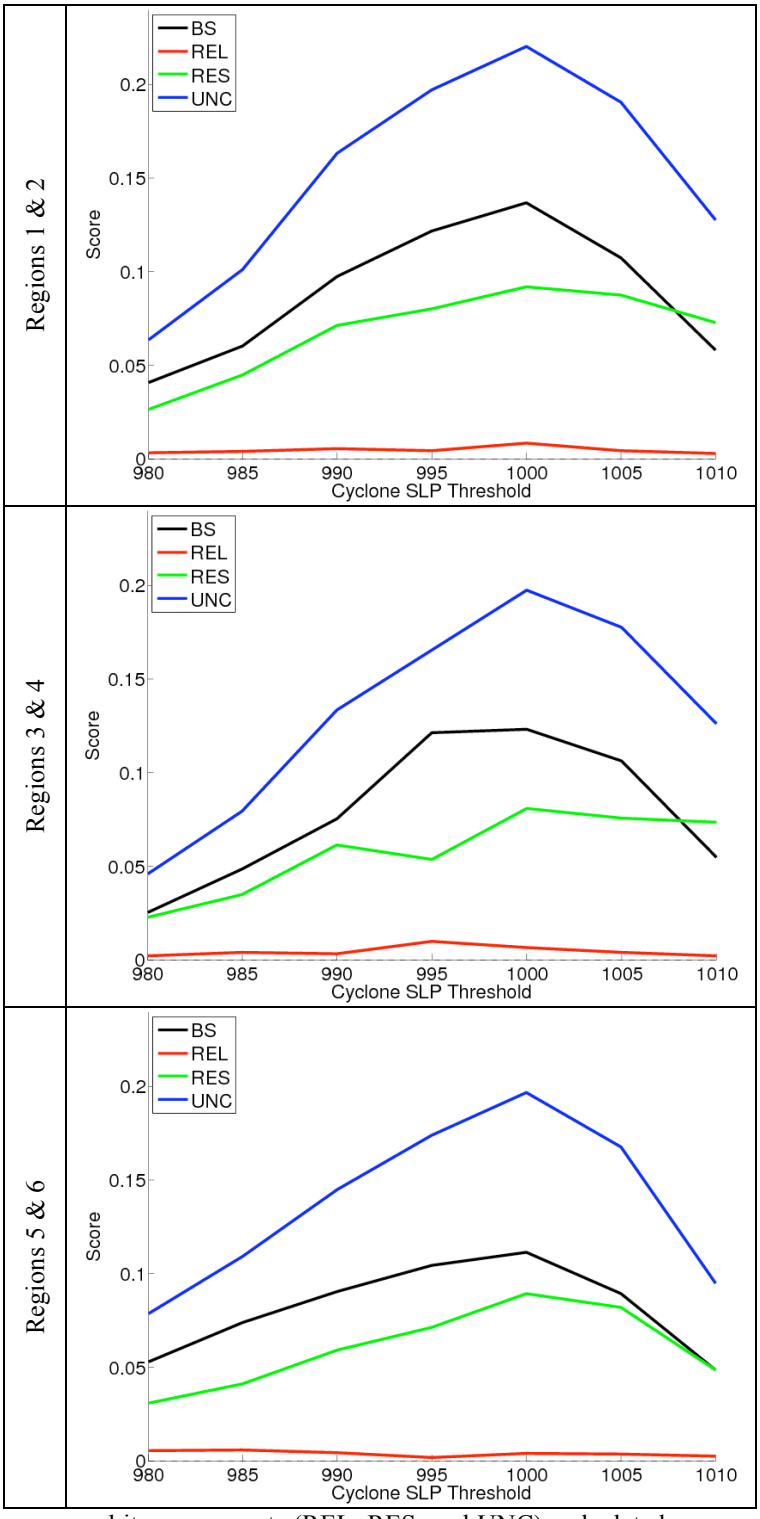

Figure 5.17. Brier score, and its components (REL, RES, and UNC), calculated over a range of thresholds (see Eq. 2-5) for the SREF, GFS, NAM, and GFS/NAM blend for F27-F63.

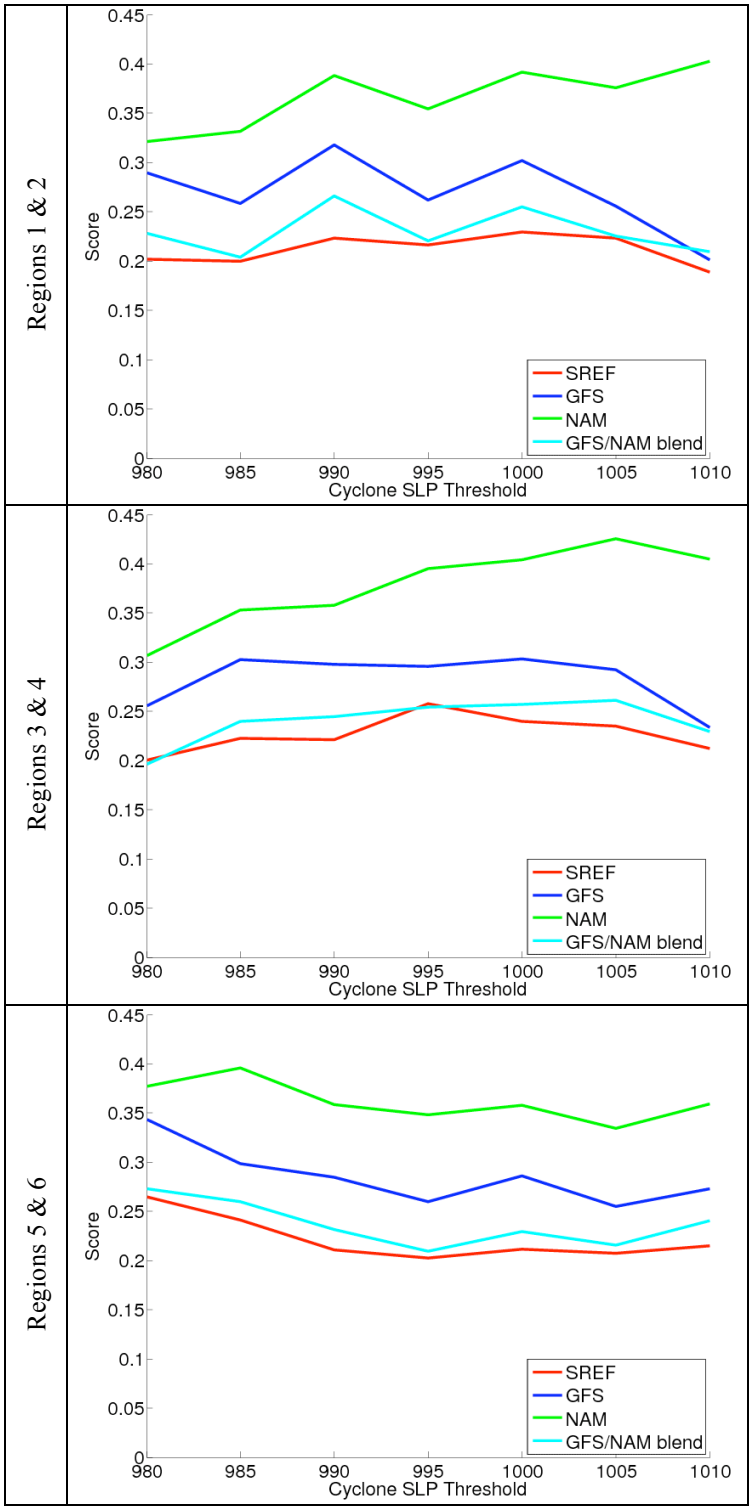

Figure 5.18. Same as Fig. 5.15 except the calculation of Brier score only includes events in which the forecast probability and/or the observed probability of the event occurring was greater than 0.

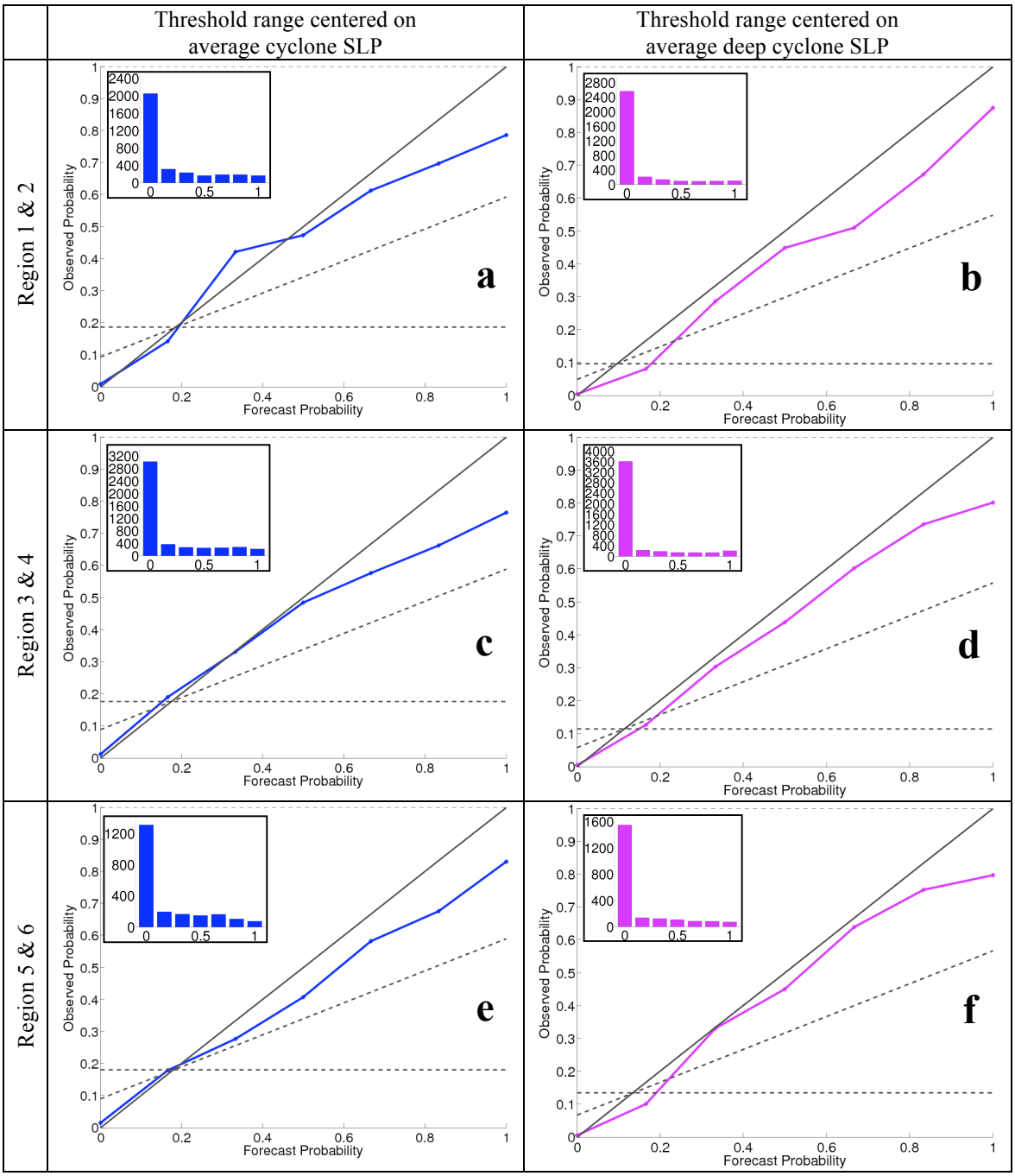

Figure 5.19. Diagrams of reliability of the SREF 15-member ensemble with respect to the average cyclone central pressure in each region (left) and 1.5 standard deviations below the average cyclone central pressure in each region (right), for regions 1-3, for F27-F63. Reliability is represented by the solid curve with markers plotted every 20%. A perfect ensemble forecast is shown by the 1:1 solid line. The tilted dashed line indicates an ensemble with no skill, and the horizontal dashed line indicates the climatology (no resolution).

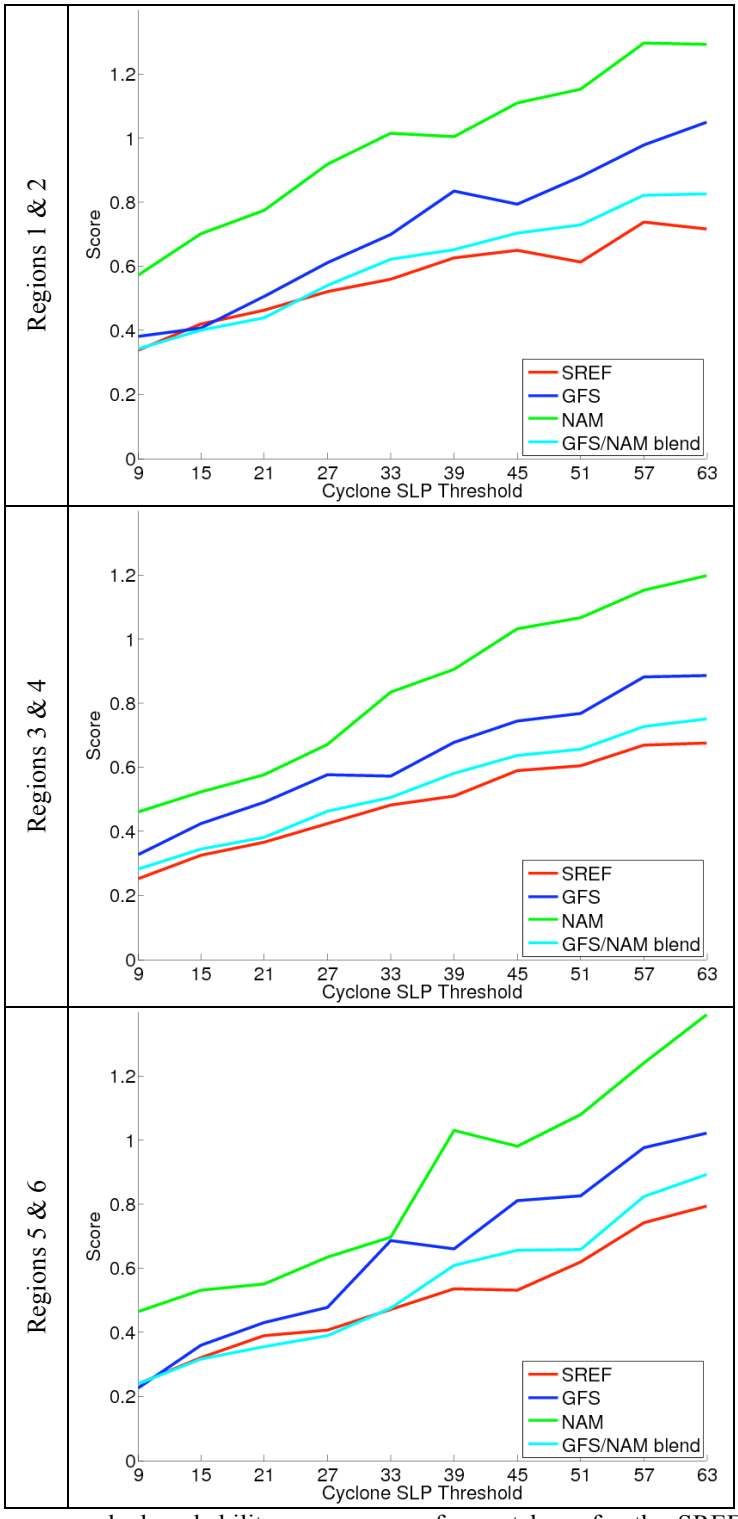

Figure 5.20. Continuous ranked probability score versus forecast hour for the SREF, GFS, NAM, and GFS/NAM blend.

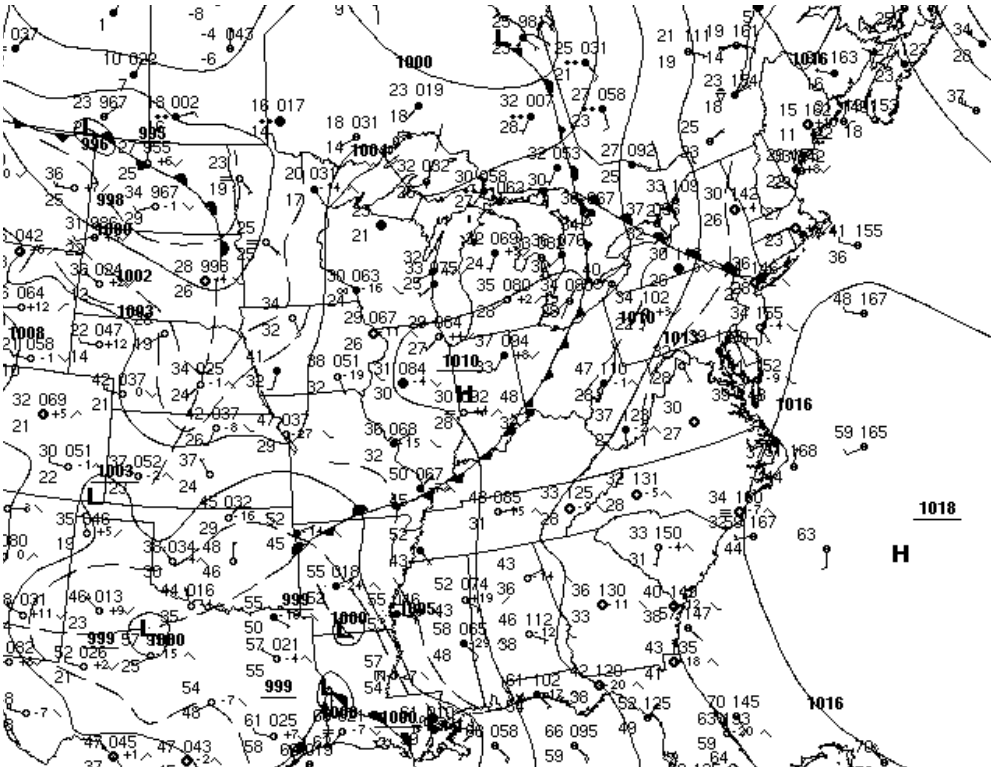

Figure 5.21. Surface analysis and station models (full barb = 10 kts) from the Hydrological Prediction Center (HPC) valid at 0900 UTC 02 February 2006.

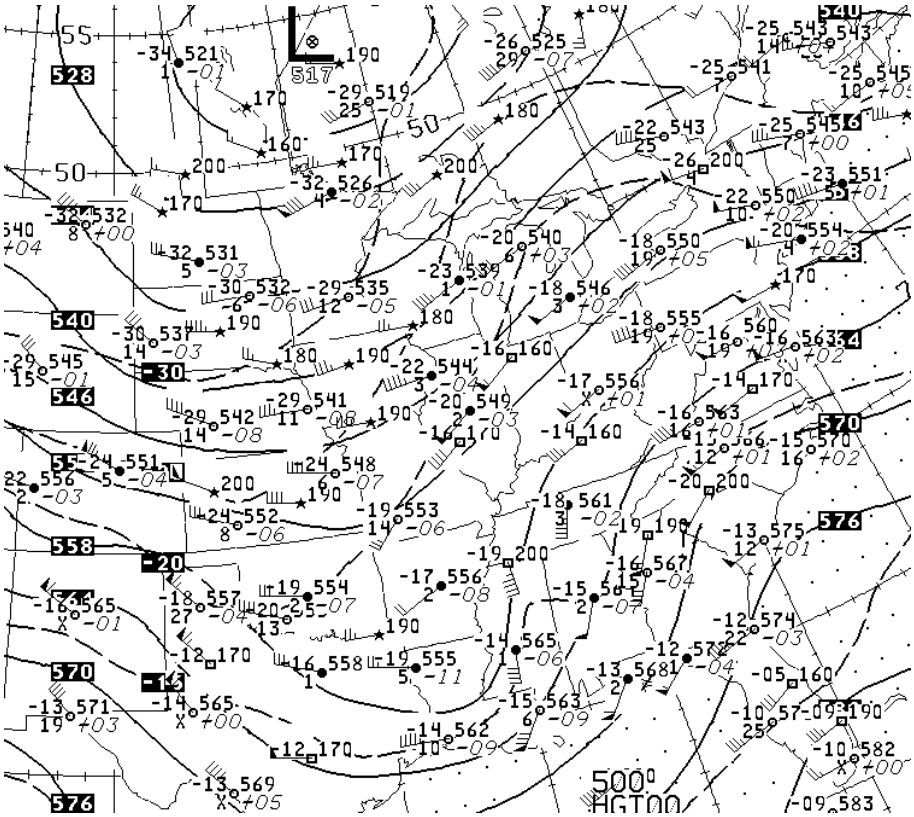

Figure 5.22. 500 mb height analysis (every 60 m) from HPC valid at 1200 UTC 02 February 2006.

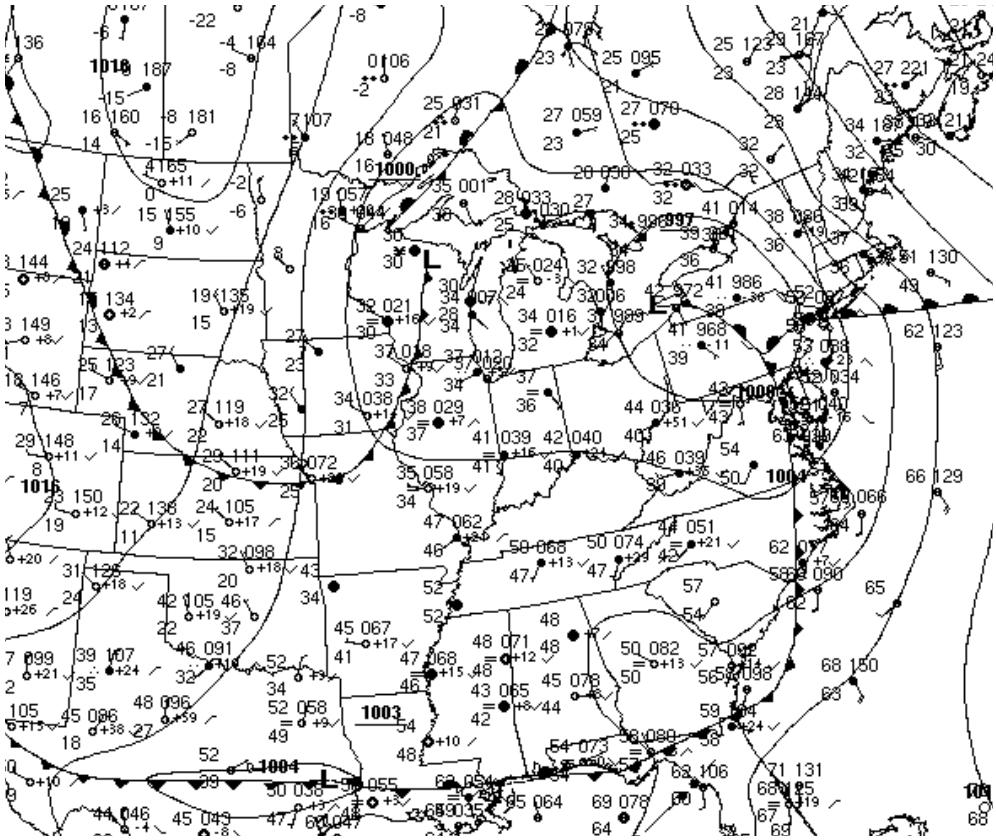

Figure 5.23. Same as Fig. 5.21, except at 12 UTC 03 February 2006.

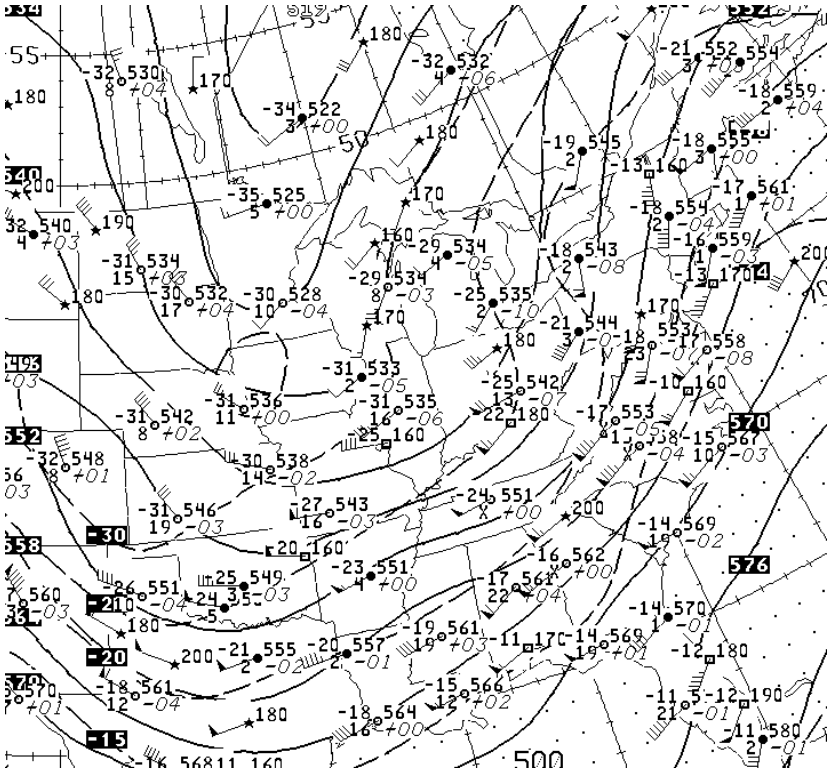

Figure 5.24. Same as Fig. 5.22, except at 12 UTC 03 February 2006.

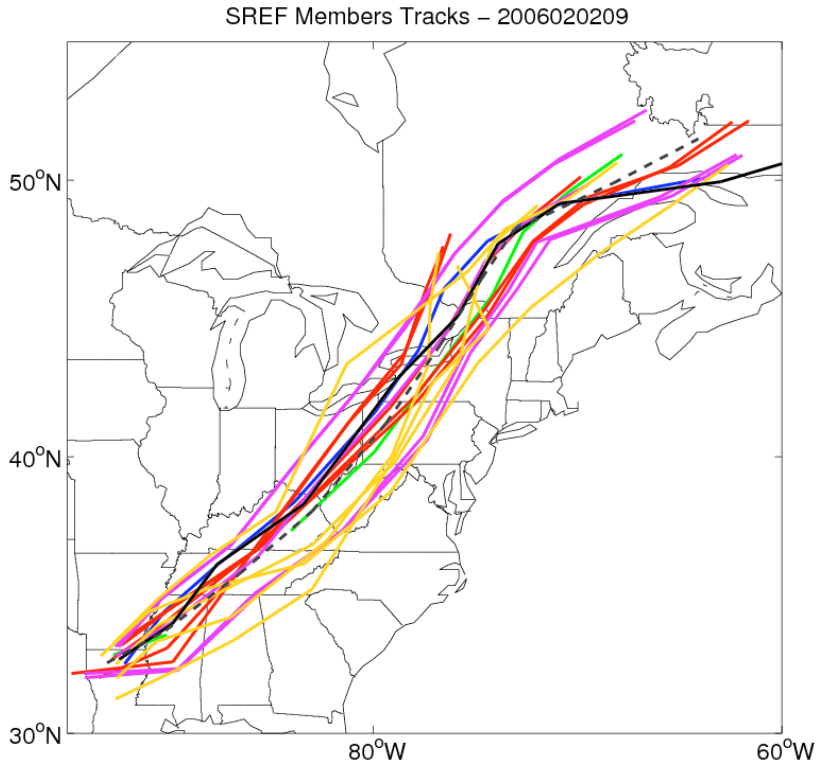

Figure 5.25. Cyclone tracks for the operational model runs initialized at 0900 UTC 02 Feb. 2006, with the SREF EBM members (red), EKF members (purple), and RSM members (yellow), GFS (blue), NAM (green), and the observed track (black) from F03 to F51.

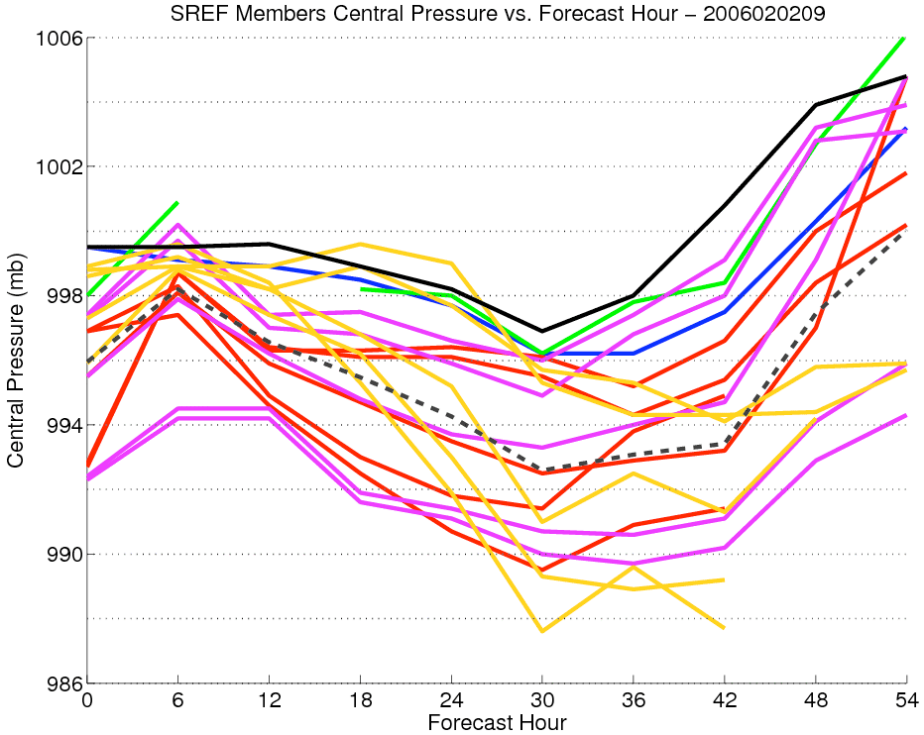

Figure 5.26. Same as Fig. 5.25 except showing forecast central pressure versus forecast hour for the SREF components, GFS, and NAM.

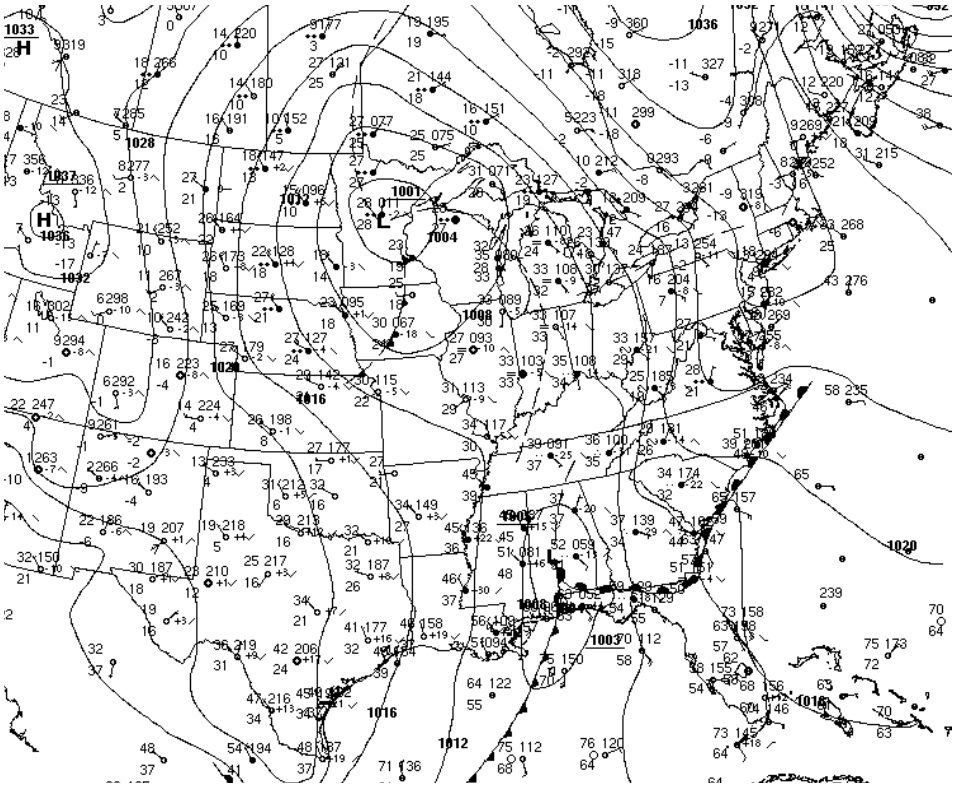

Figure 5.27. HPC surface analysis valid at 1200 UTC 15 December 2005.

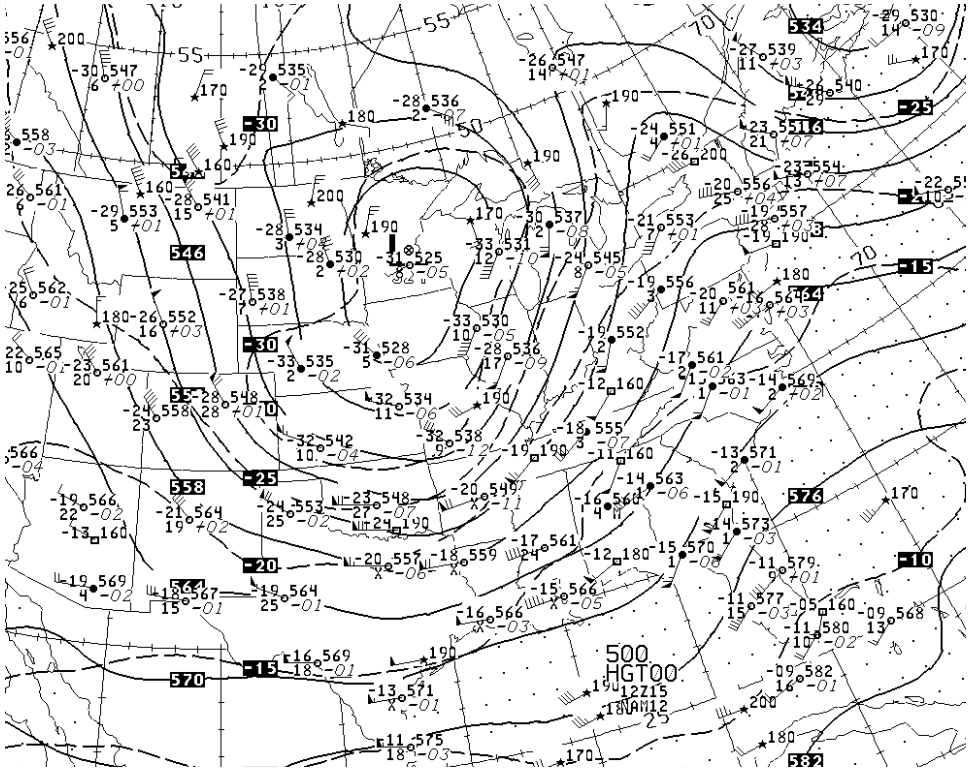

Figure 5.28. 500-mb analysis from HPC valid at 1200 UTC 15 December 2005.

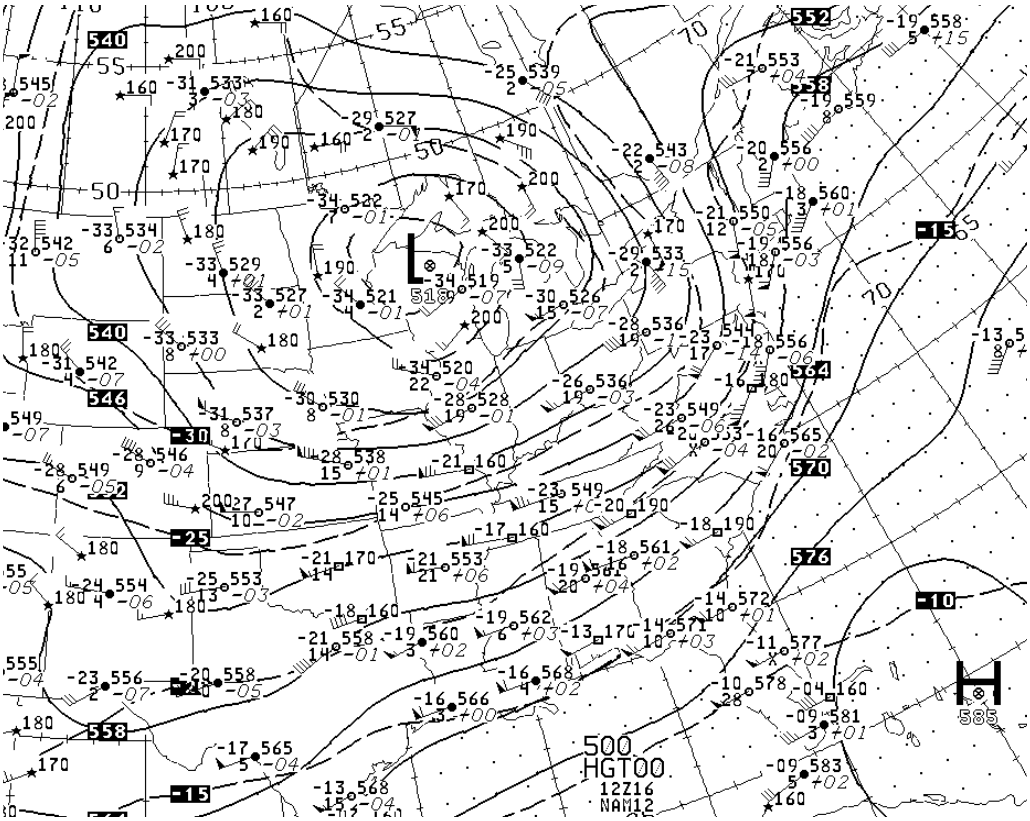

Figure 5.29. HPC surface analysis valid at 1200 UTC 16 December 2005.

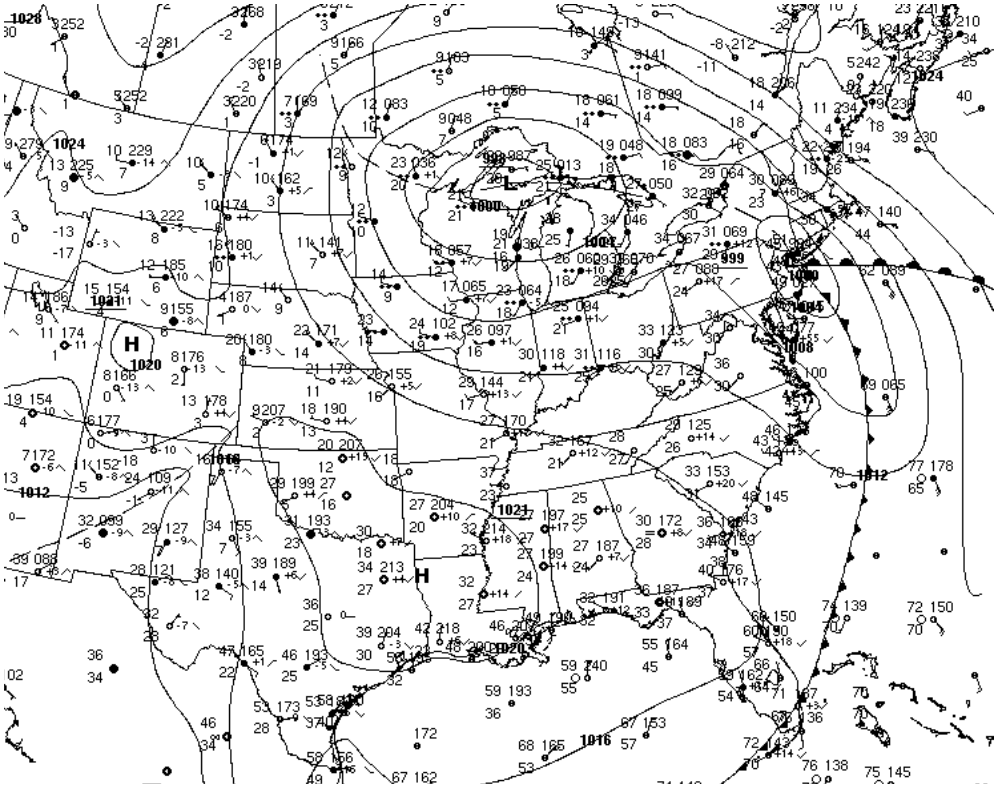
Figure 5.30. HPC 500 mb height analysis valid at 1200 UTC 16 December 2005.

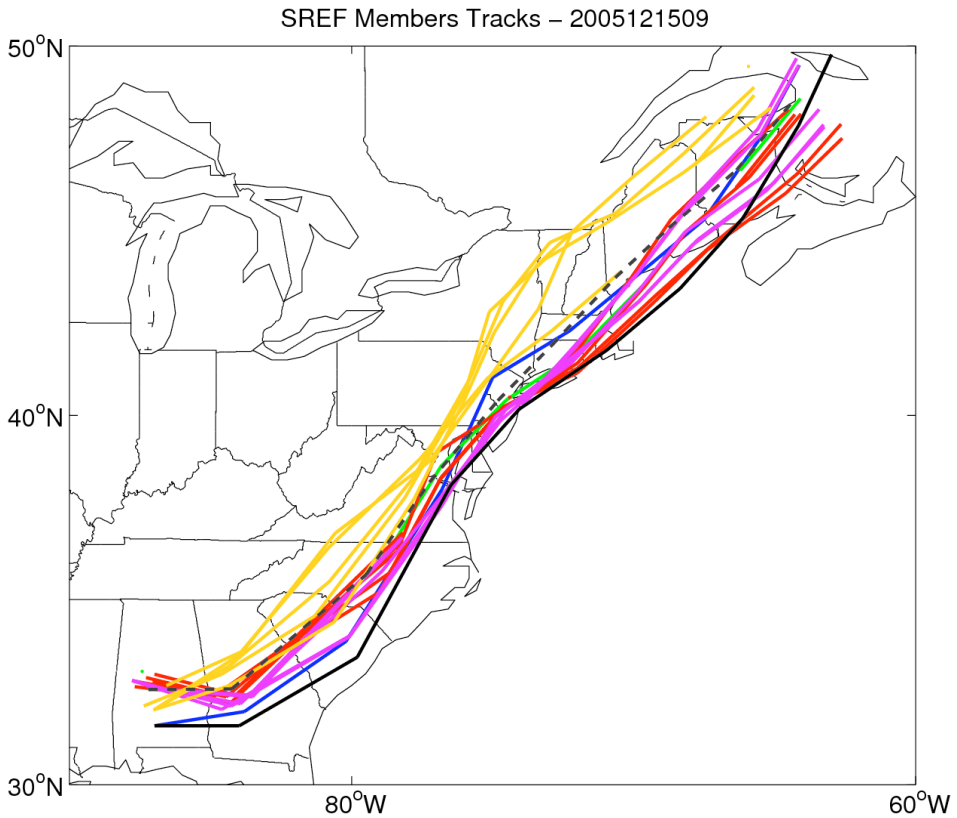

Figure 5.31. Same as Fig. 25, except for the event initialized 0900 UTC 15 December 2005.

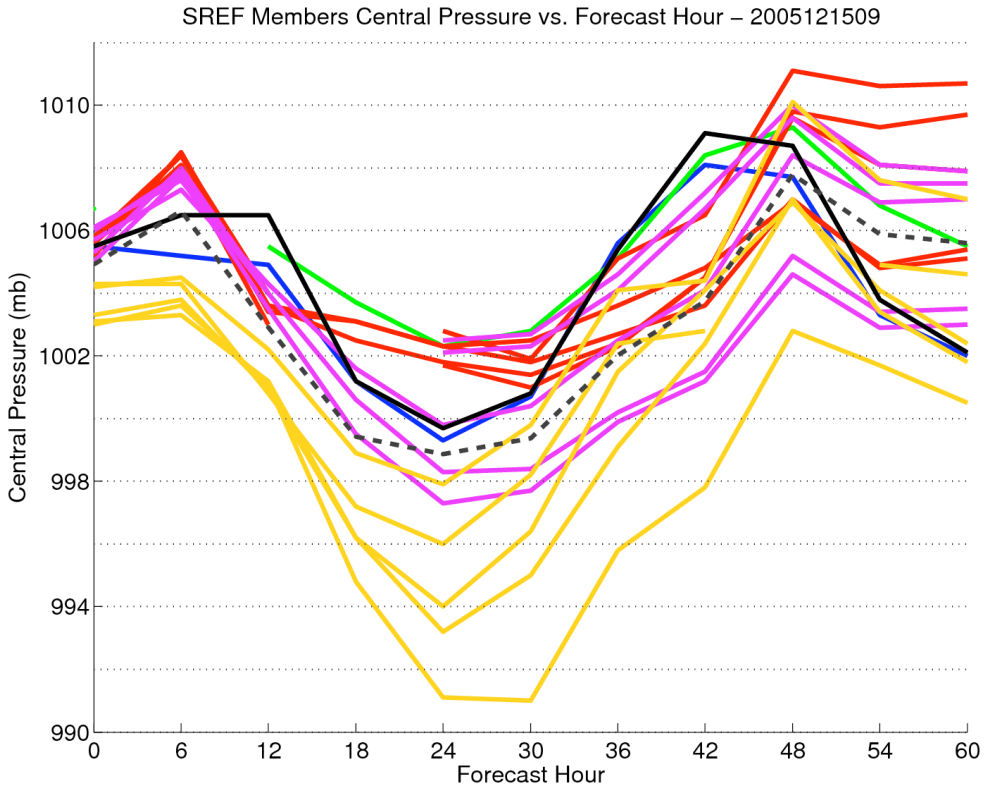

Figure 5.32. Same as Fig. 26, except for the event initialized 0900 UTC 15 December 2005.

VI. SUMMARY/CONCLUSION

The goal of this thesis was to establish a climatology of extratropical cyclone forecast errors for the cool seasons between 2002 and 2007 in NCEP operational models, including the GFS, NAM, and SREF. Another objective was to diagnose the large-scale evolution of these errors in an attempt to begin to understand the origin of these errors and how they evolve.

Forecast errors are a result of errors in the initial conditions as well as errors in the model. To evaluate the accuracy of the initial conditions of several models, the GFS analysis, NAM analysis, and NARR were verified against surface station observations when a cyclone was within 500 km (Fig. 3.4). The GFS analysis has the lowest absolute error in SLP, while the NARR has the greatest absolute error in SLP. All three analyses have a positive bias in SLP, though the GFS bias is much smaller than the NAM and the NARR. Based on these results, the GFS was used as the analysis for the cyclone central pressure. Since there was no way to determine the most accurate model analysis for the cyclone position, an average of the NAM and GFS cyclone positions was used for this study.

For all regions the cyclone central pressure absolute errors are ~0.5-1.5 mb higher in the NAM than the GFS (Fig. 3.9). Except for region 2, all regions have a more positive bias in the NAM than in the GFS. For example, in the Pacific (region 1) the positive bias peaks at ~3.5 mb at F54 in the NAM, while the Pacific bias is not significantly different than zero after F18 in the GFS. Over the 60 hour forecast, cyclone displacement in the NAM is also larger in every region than the GFS.

Both the magnitude and direction of cyclone displacement was examined. It was found that forecast cyclones in the Pacific tend to be displaced to the east and south of the observed cyclones (Fig. 3.14). In eastern Canada (region 3), cyclones are usually forecast to the west and south of the observed cyclones. In the central U.S. (region 4) cyclones were mostly displaced to the north in both models. In the Atlantic (region 6), a large portion of cyclones are displaced to the west of the observed cyclone, particularly in the NAM.

Cyclone errors in the 00/12 UTC runs were also compared to those in the proceeding 06/18 UTC runs (Fig. 3.20). Both the NAM and GFS show the largest improvement of the 06/18 UTC cyclone SLP forecasts, over the preceeding 00/12 UTC forecasts, for the Pacific (region 1). This improvement is slightly higher in the 12 UTC versus 18 UTC runs. There is also an improvement in the cyclone position forecast from 00/12 UTC to 06/18 UTC, for all regions, in both the GFS and NAM.

Over the 5 cool seasons studied (2002-2007), the NAM has a significant decrease in eastern Pacific cyclone central pressure absolute errors of more than 0.5 mb per year, though there is little improvement in the other regions (Fig. 3.12). The NAM has a significant reduction in the positive central pressure bias over the Pacific of more than 1 mb per year, as the central pressure bias decreased from ~4.5 mb in the 2002-2003 season to nearly zero by the 2006-2007 season. Meanwhile, the GFS over the same period has little improvement in any region, except a slight decrease in absolute errors in eastern Canada. The other regions all experience the same trend in the negative direction for the NAM, so that by the 2006-2007 cool season regions 2-6 have an increasing negative bias. The GFS has a slight trend of increasing negative bias in the Pacific. Although the central

pressures have been getting better in the NAM, the NAM cyclone displacement errors have been increasing over the 5-year period except region 2, while the GFS errors have a slight decreasing trend.

It is also important to understand how the strength of a cyclone is related to the cyclone errors. It was found that deep cyclones in all regions have larger absolute errors of central pressure than all cyclone events (Fig. 3.21). Unlike for all cyclones, deep cyclones in eastern Canada (region 3) are forecast with less error by the NAM than by the GFS, while the rest of the regions have larger absolute error in the NAM. Also, region 3 has a bias near zero in the NAM for deep cyclones, but a 1-2 mb positive bias in the GFS, while other regions (except for region 2) have a larger positive bias in the NAM than in the GFS. For all regions, deep cyclones are generally forecast with less displacement error than for all cyclones in both the NAM and GFS.

After F60, the GFS absolute central pressure errors over the western Atlantic (region 6) increase rapidly, and eventually become comparable to the eastern Pacific (region 1) errors (Fig. 3.24). Between F102 and F120, the central Pacific and central Atlantic errors are larger than any other region. Between F60 and F120, the Pacific and the Rockies (region 2) generally have a negative bias, with a positive bias in the rest of the regions. For deep cyclones, absolute central pressure errors over the western Atlantic grow rapidly after F72, exceeding that of other regions by F84 (Fig. 3.25). These errors correspond to a large positive pressure bias. In fact, all regions (except for region 2) experience an increasing positive bias throughout the 120 hour forecast, particularly after F48. Cyclone displacements are slightly less for deep cyclones in the GFS extended forecast than for all cyclones.

In order to advance today's operational models, the source of such cyclone forecast errors needs to be understood. Composites of the largest cyclone error events at F24 in the Pacific, and at F96 in the eastern U.S., show the evolution of SLP and 500 mb height error anomalies across the U.S. (Fig. 4.1). It was found that events with large errors over the Pacific at F24 result in errors propagating across southern Canada, but not across the U.S. By tracing back the events that result in large F96 errors in the eastern U.S., it was found that such events do not appear to originate solely from large errors over the Pacific, but from errors associated with a cyclone that moves southeastward out of south-central Canada and across the Great Lakes (Fig. 4.3). The errors associated with this composite cyclone grow very rapidly between F72 and F96, resulting in large errors over the Northeast U.S. by F96. These results suggest that events with large eastern U.S. errors can originate from rapidly growing errors within the eastern third of the U.S.

In order to determine what large-scale flows lead to cyclone overdeepening and underdeepening, tracks of large positive and negative eastern U.S. cyclones were compared (Fig. 4.5). Cyclones with positive errors appear to originate near the Great Lakes or south-central Canada, dropping southeast across the northeast U.S. (Fig. 4.5a). There also appears to be a secondary coastal development. Cyclones with negative errors appear to mainly track southwest to northeast through the eastern third of the U.S. (Fig. 4.5b). These differences show that a deterministic model like the GFS can have predictability problems with different large scale flow configurations.

The purpose of an ensemble is to vary initial conditions and/or physics in order to capture uncertainty in the atmosphere, estimating predictability. Also, unlike a deterministic forecast, an ensemble captures multiple possible outcomes, and is therefore

expected to outperform a deterministic model over a period of time. It was found that the SREF 15-member mean, as well as the 3 groups of members (EBM, EKF, RSM), do not have more skill than the GFS in any regions (Figs. 5.1-5.3). The SREF mean provides the best forecast out of all member groups and the NAM, but the GFS has lower absolute errors for much of the time in all regions (Fig. 5.1). A possible reason for this is the fact that the RSM member group continually has a large negative bias in every region at all forecast hours (Fig. 5.2). This results in the SREF 15-member mean having a negative bias in most regions. The GFS also has lower cyclone displacement than the SREF mean for all regions except for region 2 (Fig. 5.3).

Several statistical tools were used to evaluate the ensemble in a probabilistic way, such as the rank histogram, best member diagrams, and CRPS. In regions 1, 3, and 6, the SREF is overdispersed (Fig. 5.8). In regions 4 and 5, SREF forecasts are underdispersed at F51-63. There is a significant positive (negative) bias in region 1 (2-5). The best-member diagrams reveal that the SREF members are not equally skillful, and the ensemble may not accurately portray the true uncertainty in the forecast (Figs. 5.9-5.14). Decomposition of the Brier score revealed that the SREF overforecasts the probability of deep cyclone events in all regions, and has poor resolution for average cyclone events, which increases the Brier score (less skill) (Fig. 5.17). Finally, the Brier skill score reveals that a blend of the GFS and NAM has nearly the same amount of predictive skill as the SREF (Fig. 5.16).

It is hoped that this thesis can improve our understanding of extratropical cyclone errors, and therefore motivate others to view the importance of extratropical cyclone verification. For example, F72 position errors for western Atlantic (region 6) tropical cyclones today are much lower than (about half of) position errors for western Atlantic extratropical cyclones in the GFS (Fig. 6.1). Over the past 5 years, the tropical cyclone displacement errors have decreased by ~40%, but there has been essentially no trend in extratropical cyclone displacement errors over the past 4 cool seasons. This suggests that while work has been done to improve tropical cyclone forecasts, more work needs to be done to understand extratropical cyclone errors and improve forecasts. While tropical cyclones have a great impact on society and the economy, extratropical cyclones are a concern year-round, and are impacting the U.S. nearly every day of the year.

Future work for this project includes tracking warm season cyclones, cyclone relative composites of cyclone events, more case studies, and tracking cyclones within the NCEP global ensemble forecast system (GEFS). To learn why errors grow during a cyclone event, some model simulations can be done. For example, for a large error event on the East Coast, a model can be run with added observations over the Pacific to see if it improves the forecast of the event at long lead times (similar to Zhang et al. 2002). Also, to understand why the SREF appears to overpredict cyclone deepening rates in some storms, experiments can be done with modified initial perturbations. This will reveal if initial perturbations are too large. This future work will likely improve our understanding of extratropical cyclones, and our ability to accurately forecast them.

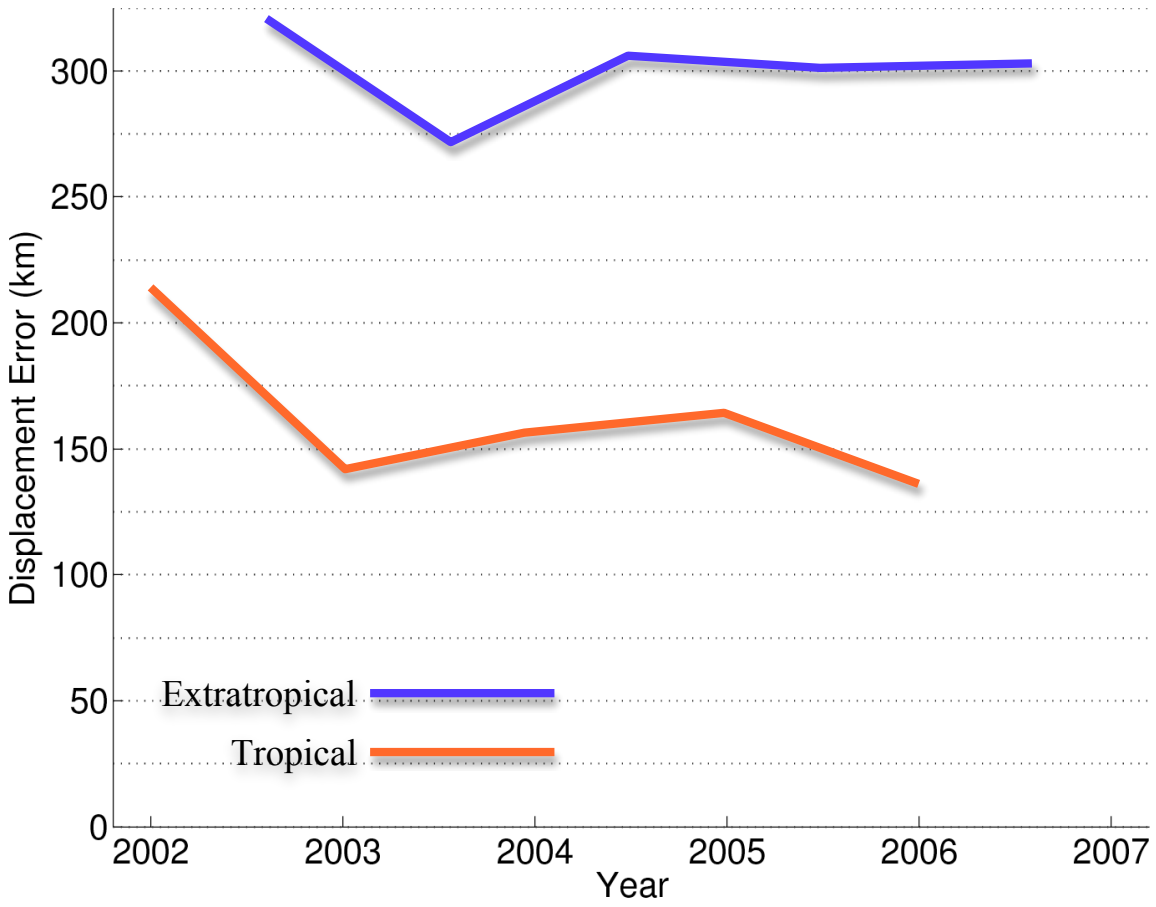

Figure 6.1. Western Atlantic extratropical (blue) and tropical (orange) cyclone displacement errors in km at hour 72 for the GFS, for each year from 2002 to 2006. Extratropical cyclones are plotted in between years since each cool season spans from Oct. of each year to Mar. of the following year. Tropical cyclone errors courtesy of Timothy Marchok.

REFERENCES

Brennan, M.J., and G.M. Lackmann, 2005: The Influence of Incipient Latent Heat Release on the Precipitation Distribution of the 24–25 January 2000 U.S. East Coast Cyclone. *Mon. Wea. Rev.*, **133**, 1913–1937.

Buizza, R., and P. Chessa, 2002: Prediction of the U.S. Storm of 24–26 January 2000 with the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **130**, 1531–1551.

Chang, E.K.M., and I. Orlanski, 1993: On the Dynamics of a Storm Track. *J. Atmos. Sci.*, **50**, 999–1015.

Colle, B.A., B.F. Smull, and M.J. Yang, 2002: Numerical Simulations of a Landfalling Cold Front Observed during COAST: Rapid Evolution and Responsible Mechanisms. *Mon. Wea. Rev.*, **130**, 1945–1966.

Colucci, S.J., 1976: Winter Cyclone Frequencies over the Eastern United States and Adjacent Western Atlantic, –. *Bull. Amer. Meteor. Soc.*, **57**, 548–553.

Davis, C.A., and K.A. Emanuel, 1991: Potential Vorticity Diagnostics of Cyclogenesis. *Mon. Wea. Rev.*, **119**, 1929–1953.

Du, J., G. DiMego, M. S. Tracton, and B. Zhou 2003: NCEP short-range ensemble forecasting (SREF) system: multi-IC, multi-model and multi-physics approach. Research Activities in Atmospheric and Oceanic Modelling (edited by J. Cote), Report 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 1161, 5.09-5.10

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New Dimension of NCEP Short-Range Ensemble Forecasting (SREF) System: Inclusion of WRF Members, Preprint,WMO Expert Team Meeting on Ensemble Prediction System, Exeter, UK, Feb. 6-10, 2006, 5 pages

Du, J., S.L. Mullen, and F. Sanders, 1997: Short-Range Ensemble Forecasting of Quantitative Precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Eckel, F.A., and C.F. Mass, 2005: Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting. *Wea. Forecasting*, **20**, 328–350.

Eichler, T., and W. Higgins, 2006: Climatology and ENSO-Related Variability of North American Extratropical Cyclone Activity. *J. Climate*, **19**, 2076–2093.

Elmore, K.L., D.M. Schultz, and M.E. Baldwin, 2006: The Behavior of Synoptic-Scale Errors in the Eta Model. *Mon. Wea. Rev.*, **134**, 3355–3366.

Froude, L.S.R., L. Bengtsson, and K.I. Hodges, 2007: The Predictability of Extratropical Storm Tracks and the Sensitivity of Their Prediction to the Observing System. *Mon. Wea. Rev.*, **135**, 315–333.

Froude, L.S.R., L. Bengtsson, and K.I. Hodges, 2007: The Prediction of Extratropical Storm Tracks by the ECMWF and NCEP Ensemble Prediction Systems. *Mon. Wea. Rev.*, **135**, 2545–2567.

Gyakum, J.R., 1983: On the Evolution of the *QE II* Storm. I: Synoptic Aspects. *Mon. Wea. Rev.*, **111**, 1137–1155.

Hacker, J.P., E.S. Krayenhoff, and R.B. Stull, 2003: Ensemble Experiments on Numerical Weather Prediction Error and Uncertainty for a North Pacific Forecast Failure. *Wea. Forecasting*, **18**, 12–31.

Hamill, T.M., and S.J. Colucci, 1997: Verification of Eta–RSM Short-Range Ensemble Forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Hirsch, M.E., A.T. DeGaetano, and S.J. Colucci, 2001: An East Coast Winter Storm Climatology. *J. Climate*, **14**, 882–899.

Hoskins, and A. J. Simmons, 1975: A multi-layer spectral model and the semi-implicit. *Quart. J. Roy. Meteor. Soc.*, **101,** 637-655.

Hoskins, B.J., and K.I. Hodges, 2002: New Perspectives on the Northern Hemisphere Winter Storm Tracks. *J. Atmos. Sci.*, **59**, 1041–1061.

Hoskins, B.J., and P.J. Valdes, 1990: On the Existence of Storm-Tracks. *J. Atmos. Sci.*, **47**, 1854–1864.

Kocin, P.J., P.N. Schumacher, R.F. Morales, and L.W. Uccellini, 1995: Overview of the 12–14 March 1993 Superstorm. *Bull. Amer. Meteor. Soc.*, **76**, 165–182.

Langland, R.H., M.A. Shapiro, and R. Gelaro, 2002: Initial Condition Sensitivity and Error Growth in Forecasts of the 25 January 2000 East Coast Snowstorm. *Mon. Wea. Rev.*, **130**, 957–974.

Marchok, T.P., 2002: How the NCEP tropical cyclone tracker works.Preprints, 25th Conference on Hurricanes and Tropical Meteorology, Amer. Meteor. Soc., 21-22.

McMurdie, L., and C. Mass, 2004: Major Numerical Forecast Failures over the Northeast Pacific. *Wea. Forecasting*, **19**, 338–356.

McQueen, J., J. Du, B. Zhou, G. Manikin, B. Ferrier, H-Y. Chuang, G. DiMego, and Z. Toth, 2005: Recent Upgrades to the NCEP Short-Range Ensemble Forecasting System (SREF) and Future Plans. Preprints, 17th Conference on Numerical Weather Prediction/21st Conference on Weather Analysis and Forecasting, Washington DC., Aug. 1-5, 2005, Amer. Meteor. Soc. (paper 11.2).

Mesinger, F., and R.E. Treadon, 1995: "Horizontal" Reduction of Pressure to Sea Level: Comparison against the NMC's Shuell Method. *Mon. Wea. Rev.*, **123**, 59–68.

Miller, J.E., 1946: CYCLOGENESIS IN THE ATLANTIC COASTAL REGION OF THE UNITED STATES. *J. Atmos. Sci.*, **3**, 31–44.

Mullen, S.L., and B.B. Smith, 1990: An Analysis of Sea-Level Cyclone Errors in NMC's Nested Grid Model (NGM) During the 1987-88 Winter Season. *Wea. Forecasting*, **5**, 433–447.

Murphy, A.H., 1973: Hedging and Skill Scores for Probability Forecasts. *J. Appl. Meteor.*, **12**, 215–223.

Pauley, P.M., 1998: An Example of Uncertainty in Sea Level Pressure Reduction. *Wea. Forecasting*, **13**, 833–850.

Reitan, C.H., 1974: Frequencies of Cyclones and Cyclogenesis for North America, 1951–1970. *Mon. Wea. Rev.*, **102**, 861–868.

Roebber, P.J., 1984: Statistical Analysis and Updated Climatology of Explosive Cyclones. *Mon. Wea. Rev.*, **112**, 1577–1589.

Sanders, F., and J.R. Gyakum, 1980: Synoptic-Dynamic Climatology of the "Bomb". *Mon. Wea. Rev.*, **108**, 1589–1606.

Silberberg, S.R., and L.F. Bosart, 1982: An Analysis of Systematic Cyclone Errors in the NMC LFM-II Model During the 1978–79 Cool Season. *Mon. Wea. Rev.*, **110**, 254–271.

Smith, B.B., and S.L. Mullen, 1993: An Evaluation of Sea Level Cyclone Forecasts Produced by NMC's Nested-Grid Model and Global Spectral Model. *Wea. Forecasting*, **8**, 37–56.

Steenburgh, W.J., and C.F. Mass, 1996: Interaction of an Intense Extratropical Cyclone with Coastal Orography. *Mon. Wea. Rev.*, **124**, 1329–1352.

Stensrud, D.J., H.E. Brooks, J. Du, M.S. Tracton, and E. Rogers, 1999: Using Ensembles for Short-Range Forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Toth, Z., and E. Kalnay, 1997: Ensemble Forecasting at NCEP and the Breeding Method. *Mon. Wea. Rev.*, **125**, 3297–3319.

Uccellini, L.W., P.J. Kocin, J.M. Sienkiewicz, R. Kistler, M. Baker: Fred Sanders' Roles in the Transformation of Synoptic Meteorology, the Study of Rapid Cyclogenesis, the Prediction of Marine Cyclones, and the Forecast of New Yokr City's 'Big Snow' of December 1947. In Press. Sanders Monograph, AMS.

Uccellini, L.W., P.J. Kocin, R.S. Schneider, P.M. Stokols, and R.A. Dorr, 1995: Forecasting the 12–14 March 1993 Superstorm. *Bull. Amer. Meteor. Soc.*, **76**, 183–199.

Weigel, A.P., M.A. Liniger, and C. Appenzeller, 2007: Generalization of the Discrete Brier and Ranked Probability Skill Scores for Weighted Multimodel Ensemble Forecasts. *Mon. Wea. Rev.*, **135**, 2778–278

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Zapotocny, T.H., W.P. Menzel, J.A. Jung, and J.P. Nelson, 2005: A Four-Season Impact Study of Rawinsonde, GOES, and POES Data in the Eta Data Assimilation System. Part I: The Total Contribution. *Wea. Forecasting*, **20**, 161–177.

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale Predictability of the "Surprise" Snowstorm of 24–25 January 2000. *Mon. Wea. Rev.*, **130**, 1617–1632.

Zishka, K., and P. Smith, 1980: The Climatology of Cyclones and Anticyclones over North America and Surrounding Ocean Environs for January and July, 1950–77. *Mon. Wea. Rev.*, **108**, 387–401.

Zwiers, F.W., 1990: The Effect of Serial Correlation on Statistical Inferences Made with Resampling Procedures. *J. Climate*, **3**, 1452–1461.