ABSTRACT

KENNEDY, RACHEL ERIN. Assessing Numerical Weather Prediction Model Forecast Skill Under Different Weather Conditions Using Surface Observations. (Under the direction of Dr. Sandra Yuter).

Identifying conditions where a weather prediction model has higher and lower forecast skill aids in constraining where refinements in different aspects of model physics will have the highest impact and helps forecast users to account for typical biases. This study analyzes numerical weather prediction model biases under different meteorological conditions, including the amount of observed cloud cover, time of day, region, and season. As well as comparisons of the numerical values of weather variables between observations and forecasts at a given valid time, we also examined errors in the timing of low pressure system passages and precipitation events. The Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) is a regional operational weather forecasting model developed and used by the US Navy. The Global Forecast System (GFS) is the the US National Oceanic and Atmospheric Administration's global operational weather forecast model. We compare COAMPS and GFS model output to surface observations at 333 weather stations and 37 buoys across the North America with an additional 88 stations for GFS only in the Intermountain West. Focus is on the the summer season from May - Sep 2022 and the winter season of Nov 2022 - Feb 2023. The findings of this study will be given to the Naval Research Laboratory (NRL) and used to direct model developers to variables with the largest biases and identify potential causes of these biases.

We use several types of evaluation metrics for forecasts at 48-hour lead times. Matched time comparisons of model output minus observations are a standard metric but apparent biases can result from small timing and location errors. We compare bulk distributions of a given variable over a season to determine if biases detected in matched-time comparisons are similar when the timing criteria is relaxed. We have also developed two metrics to quantify timing errors. We use pressure tendency to determine the timing of low pressure center passages and assess to what degree the model forecasts low pressure center passages too early or too late. In addition, we examine forecast and observed start and stop times as well as durations of precipitation events.

We found that there are regional and time of day differences in several types of model biases. Winter afternoon temperatures (3PM local time) temperatures are often too cold with the exception of locations in the Plains which have smaller errors. At 7AM local time in

winter, there is a south to north gradient in temperature biases with locations to the south tending to be too warm while those north are too cold. Station median temperature biases were usually larger in winter than in summer. In winter, biases were larger in conditions with less cloud coverage than in all cloud coverage conditions. Temperature biases at 7AM in the winter were found to be influenced by the amount of observed cloud cover, changing from a warm bias when all conditions were examined to a cold bias when only conditions with <25% cloud cover were examined. Dewpoint errors in the western US tend to be too dry while those in the eastern US are too moist. Biases for temperature, low pressure passage timing, and wind speed/wind direction all tended to be larger in mountainous terrain as compared to flatter areas.

We defined extreme temperature events as outside of the 10th and 90th percentiles of the 30-year hourly climatology. Typically, the severity of these events was underestimated in the forecasts where observed temperatures were > 90th percentile as too cool and where observed temperatures were < 10th percentile as too warm. Median forecast biases for the subset of extreme events were larger than for all events.

In regions outside of mountainous terrain, most low pressure passages were forecast to occur within +/- 2 hours of the observed low pressure passage time. Both COAMPS and GFS had notable timing errors for precipitation events with start times typically forecast to occur too early and end times too late yielding precipitation event durations that were several hours or more too long.

The model physics are identical and the initializations are very similar between two COAMPS regional model runs that cover California with different grid spacing. We examined temperature errors for COAMPS NEPAC (15.5 km grid) and COAMPS CENCOOS (3.65 km grid). The finer grid CENCOOS model performed worse in temperature forecasts, particularly in the summer at both 7AM and 3PM local time, than the coarser grid NEPAC. Wind speed errors large enough to impact aviation tended to occur more frequently at buoys, along the coast, and near mountainous terrain at both grid resolutions.

Comparison of temperature and dewpoint biases for both models suggests that errors in temperature and dewpoint forecasts only minimally influence each other. Comparison of temperature biases and wind errors shows no real relationship, indicating errors in wind speed/direction are not the largest cause of temperature biases. Stations with larger temperature biases and more frequent wind speed/direction errors tended to be located in complex mountainous terrain.

Assessing Numerical Weather Prediction Model Forecast Skill Under Different Weather Conditions Using Surface Observations

by
Rachel Erin Kennedy

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Marine, Earth, and Atmospheric Science

Raleigh, North Carolina
2023

APPROVED BY:

Dr. Matthew Parker

Dr. Sarah Larson

Dr. Sandra Yuter
Chair of Advisory Committee

ACKNOWLEDGEMENTS

Thank you to my adviser Dr. Sandra Yuter for all of her time and guidance over the last two years. This project would not have been possible without her support. Thank you to my committee members, Dr. Matthew Parker and Dr. Sarah Larson, your feedback and ideas about my thesis were invaluable. Thank you to the Naval Research Lab - Monterey for their help and feedback throughout this project. The research was supported in part by ONR grant N00014-21-1-2116. Thank you to my research group Environmental Analytics but specifically to Dr. Matthew Miller, Laura Tomkins, Kevin Burris, and Jordan Fritz for listening to me and talking through different issues as they came up with me. Special thanks goes to Dr. Miller and Jordan Fritz for allowing me to use their codes and contributing to the analysis of model biases in this study. Lastly I would like to thank my friends and family but special thanks goes out to my mom, my sister, Miranda Silcott, Elizabeth Hartsog-Chakrabarty, Valerie Ryba, and my cats. None of this would have been possible without their support and I would be lost without their guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Numerical weather prediction models are an essential tool for weather forecasters and provide key information that is used create accurate and timely forecasts for the general public and a wide variety of specialized users including agriculture, aviation, sporting events, and military operations. It is well known that model forecast skill for a given lead time varies among models and for different types of weather events and seasons. Both the National Weather Service and commercial weather forecasters make bias adjustments to correct for known forecast deficiencies as part of post-processing model output (Glahn and Lowry 1972). But depending on how these bias adjustments are calculated, they can make individual forecasts better or worse. For example, a bias adjustment that is calculated based on daily averages may make diurnally-varying errors worse as the average bias adjustment may overcompensate for one portion of the diurnal variation. Experienced forecasters are often aware of local biases in the models, even after post-processing, that directly impact their forecast region and adjust their predictions accordingly.

In addition to the recent advent of rapid and sophisticated machine learning, there have been many differing opinions over the years on whether it is more fruitful to refine the underlying model physics (Wong et al. 2020), improve inputs to the model via upgrades to data assimilation (Schultz et al. 2021), or if post-processing bias adjustments and/or

blending outputs of an ensemble of models is the better route to more skillful deterministic forecasts (Vaittinada Ayar et al. 2021). For the purposes of this study, we take the position that improving the physical representation of the atmosphere within a numerical weather prediction model will yield the longest lasting benefits. Modeling centers need information on model strengths and weaknesses and diagnosis of likely error sources to know how best to invest finite resources in improvements in the underlying numerical weather prediction algorithms.

1.1 Background and Motivation

Operational numerical forecast model performance is usually evaluated over seasons (e.g. Colle et al. 1999; Sobash and Kain 2017), annually, over decades (e.g. Sukovich et al. 2014), or for particular events (e.g. Chien et al. 2002) by groups outside of and within the modeling centers responsible for a particular model. Most model evaluation efforts utilize one or more verification scores such as skill score, anomaly correlation coefficient, and root-mean-square error (Casati et al. 2008). Commonly, model output is compared to reanalysis, which is helpful in the sense that there is a value for every grid box but problematic since reanalysis is not truly an independent data set. Reanalysis data uses observations at set locations as a starting point and interpolates between observation points to reproduce values for those locations (Parker 2016). These values are calculated using theory based equations that are used in first-guess forecasts, meaning reanalysis data are partially determined by NWP forecasts and are not based purely on observations (Parker 2016). Typical metrics to assess model performance, such as 500 hPa geopotential heights (e.g. Sun et al. 2023; Ji et al. 2021) are useful to address synoptic scale features but have only an indirect relation to hour by hour surface conditions where people live and work. It is beneficial to evaluate forecasts at hourly time scales since the duration of weather conditions such as timing of precipitation and periods of excessive heat and cold can have large impacts on transportation and outdoor work and leisure activities.

This study evaluates strengths and weaknesses in the US Navy's regional weather forecasting model, the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) (Hodur 1997) relative to surface observations at 333 airports (plus an additional 88 for GFS only) and 37 buoys in North America during summer season of May - September 2022 and winter season of November 2022 - February 2023. The specific months selected for summer versus winter also represent the approximate dry season (summer) and wet season

(winter) for California which is a location of interest for the Naval Research Laboratory based in Monterey, CA (NOAA 2023). We acknowledge that, given the limited number of months used to represent the summer/winter seasons, biases calculated in this study may be impacted by large scale patterns present during the requested time periods. Most of the analysis examines the 48-hour lead time forecasts. A parallel analysis is undertaken for NOAA's Global Forecasting System (GFS) model (AWS 2022) to provide context on how the magnitude and regional distributions of biases vary between COAMPS and GFS. This study builds on work by Patel et al. (2021) on assessing winter season diurnal temperature forecasts from the NOAA's GFS and High Resolution Rapid Refresh (HRRR) models as a function of observed cloud cover.

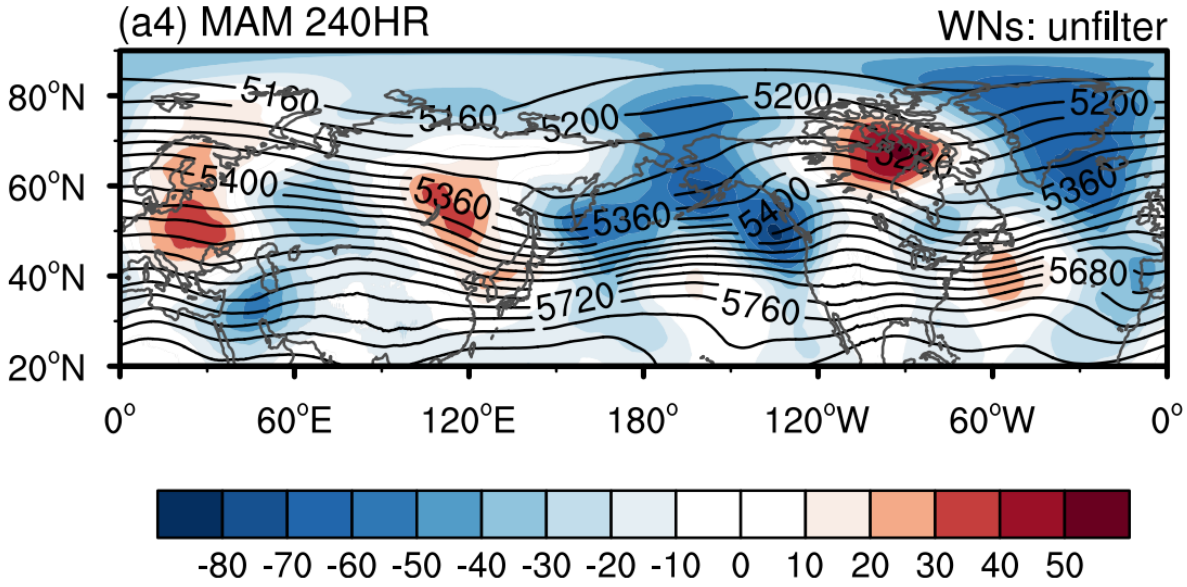

Figure 1.1: Example of model verification for seasonal (March/April/May 2021) mean 500 hPa geopotential height forecasts using the CMA-GFS model forecast field at a leadtime of 240 hours. Contours represent the mean seasonal (MAM) geopotential height forecast values and shading represents the corresponding model forecast geopotential height errors over the same time period. The highest amount of errors are found in between troughs and ridges, with blue shading representing the model forecasted heights too low and red shading representing the model forecasted heights too high. Adapted from Sun et al. (2023), their Figure 2 A4.

Standard forecast verification methods are typically defined using statistical methods to assess some form of skill score, mean square error, observation and forecast variance, bias,

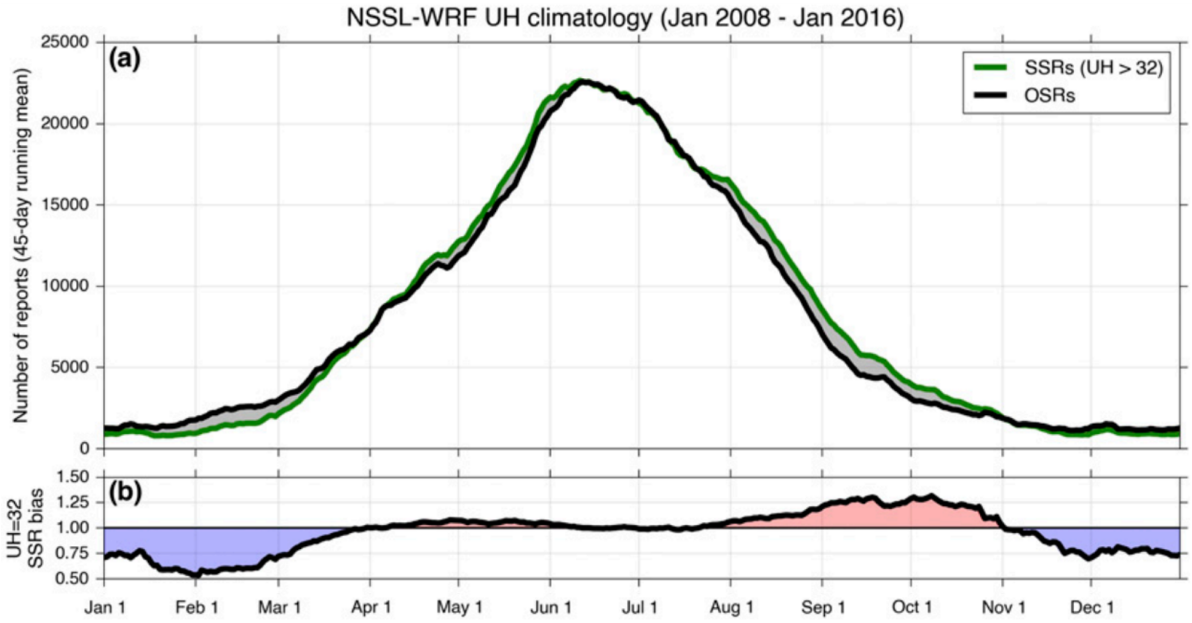

Figure 1.2: Example of model verification using model created surrogate storm reports (SSR) compared to observed storm reports (OSR). A) Number of observed storm reports and surrogate storm reports on a 45 day rolling window using data from January 2008 - January 2016. B) Surrogate storm report biases, blue is an underforecast and red is an overforecast. Adapted from Sobash and Kain (2017), their Figure 3 A/B.

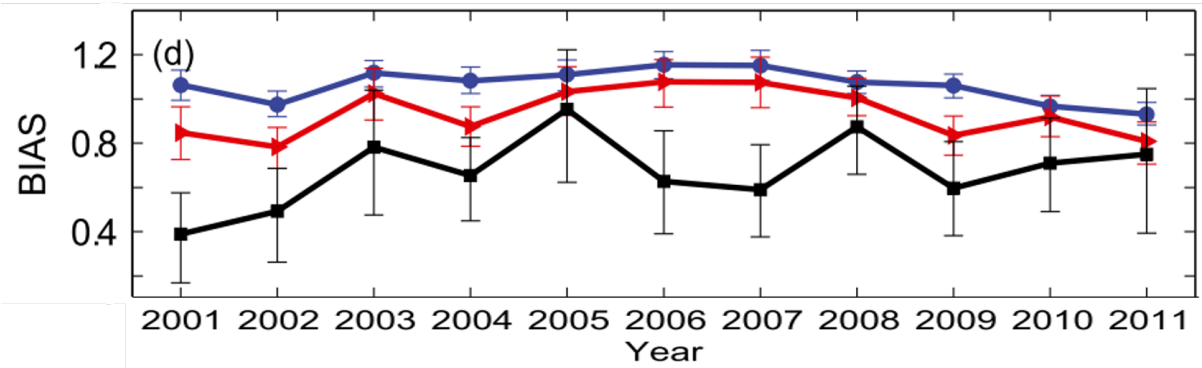

Figure 1.3: Example of model verification of the frequency biases associated with the calculation of the 1 inch precipitation threat score (blue), where a threat score depends on how similar the forecast and observed precipitation events are for any given amount of precipitation, the top 1.0% of extreme precipitation events threat scores (red), and the top 0.1% of extreme precipitation events threat scores (black). The frequency bias is calculated from the number of forecast verses observed events at a given threshold (i.e. 1 inch, top 1.0%, or top 0.1%) with a bias of 1 corresponding to an unbiased forecast, a bias > 1 meaning the model is overforecasting event occurrence, and a bias of < 1 meaning the model is underforecasting event occurrence. Adapted from Sukovich et al. (2014), their Figure 5D.

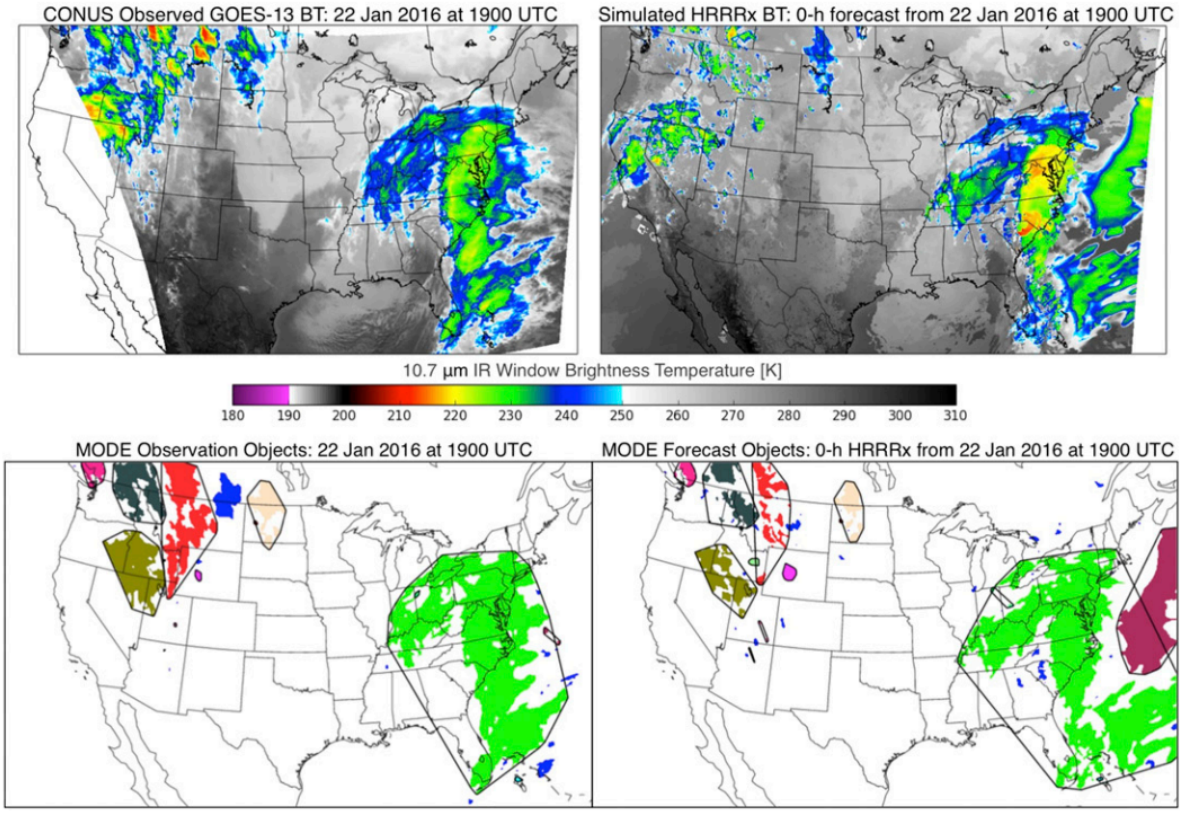

Figure 1.4: Example of model verification for the amount of forecast cloud cover. Top row shows satellite observed and HRRR-model simulated IR brightness temperatures for 22 January 2022 at 1900 UTC. Bottom row shows extracted cloud feature objects and their bounding region perimeters. Adapted from Griffin et al. (2017), their Figure 3.

and correlation (Casati et al. 2008). Newer evaluation metrics are providing verification results that are more tailored to the applications of forecast users (e.g. agriculture, energy sector, emergency management). The World Meteorological Organization in collaboration with the National Meteorological Services developed a set of verification tools to standardize the forecast verification process and create consistency in results (Casati et al. 2008). The development and testing of forecast verification techniques is an ongoing process that has allowed verification metrics to become more refined and to provide better information on model forecast skill.

Previous work on operational regional and global model verification has mostly focused on synoptic scale features in 500 hPa geopotential heights (e.g. Sun et al. 2023) (Fig. 1.1). There has been some work analyzing model biases in specific environments such as Evans et al. (2018) which examines the model forecast skill for thunderstorm-supporting environments in regional models such as RAP (Benjamin et al. 2016) and HRRR (Dowell et al. 2022). Other studies (e.g. Sobash and Kain 2017), have assessed the forecast skill of convection-allowing model (CAM) forecasts by creating surrogate severe probability forecasts and comparing these forecasts to storm reports from the Storm Prediction Center (Fig. 1.2). Other areas of long-term interest for model evaluation are quantitative precipitation forecasts (Sukovich et al. 2014) (Fig. 1.3) and forecasting the amount of observed cloud cover (Griffin et al. 2017) (Fig. 1.4). The Model Evaluation Tools (MET) verification package developed by the Developmental Testbed Center was originally designed to evaluate mesoscale model case studies and employs some object-based feature verification tools (Bullock et al. 2016).

Li et al. (2022) analyzed the wintertime surface air temperature forecast skill of three NWP models: Model for Prediction Across Scales-Atmosphere (MPAS-A), China Meteorological Administration (CMA) model, and the European Centre for Medium-Range Forecasts (ECMWF) model at a two week lead time to assess forecast skill on a subseasonal timescale. They found that skill and bias varied regionally. For wintertime surface temperatures, ECMWF had a dominant cold bias in western North America, CMA had a dominant warm bias in eastern North America, and MPAS had a dominant cold bias across all of North America. Western North America displayed the lowest forecast skill likely due to the complex topography of this region while eastern North America displayed the highest forecast skill (Li et al. 2022).

Dutra et al. (2021) used hindcasts which are made by a recent NWP model version to predict weather in previous years before that version was available. They used 29 years of ECMWF hindcasts to assess systematic temperature biases across CONUS in April-May

and June-July at leadtimes of 1, 2, 3, 4, 5, and 6 weeks. From April to May, ECMWF hindcasts showed predominantly cold biases across the United States but transitioned to predominantly warm biases from June-July. Temperature biases at the time of the daily minimum temperature were found to be too warm, indicating the ECMWF was not forecasting temperatures to drop as low as observations overnight (Dutra et al. 2021). They noted that there was a higher forecast skill in the winter than in the summer, hypothesizing that this is related to more persistent synoptic scale activity in the winter in comparison to the summer (Dutra et al. 2021).

Køltzow et al. (2019) assessed the skill of short-range forecasts using all available hourly leadtimes from the model run for conditions in the Arctic of four NWP models: ECMWF Integrated Forecasting System, Applications of Research to Operations at Mesoscale (AROME)-Arctic, Canadian Arctic Prediction System (CAPS), and AROME-Meteo France. This study was limited by the relatively small number of weather observing stations in the Arctic region and found that forecast skill in the Arctic tended to be lower than it was in the middle and lower latitudes (Køltzow et al. 2019). The study found that Arctic temperature errors tended to increase in magnitude when no cloud cover was observed, indicating a potential problem with the model's parameterization of turbulent mixing and/or in its representation of the stable boundary layer (Køltzow et al. 2019). Temperature forecast skill was also found to be sensitive to associated forecasts of surface wind speed, the amount of snow cover, and sea ice (Køltzow et al. 2019).

Massey et al. (2016) assessed the diurnal trend of temperature biases across the Intermountain West depending on surface soil moisture and the amount of observed cloud cover. This was done by comparing operational 48-hour WRF (4DWX-DPG) forecasts using nested grid spacing domains of 30, 10, 3.3, and 1km to observations taken from September to October from 2011 to 2013. The authors found that the model diurnal temperature range was underpredicted when soil moisture was too moist but diurnal temperature range forecast skill increased when soil moisture was drier (Massey et al. 2016). They additionally found that 4DWX-DPG had an overall warm bias in the early morning/overnight and an overall cold bias in the afternoon (Massey et al. 2016). Temperature biases were found to be larger under clear conditions and smaller under cloudy conditions (Massey et al. 2016). Sensitivity tests noted that temperature biases were reduced when a corrected soil moisture parameterization was applied to the model (Massey et al. 2016).

Lu et al. (2011) assessed the skill of the MM5 (a fifth generation mesoscale model from Pennsylvania State University and NCAR), COAMPS (research mode), and the WRF model over the southeast United States and the northern Gulf of Mexico at leadtimes of 24 and 36

hours during the 2003 warm season and 2004 cool season. Horizontal grid spacing for all three models was 27km. Specifically focusing on COAMPS, the authors found that in the warm season COAMPS displayed a weak warm bias in the Southeast and cool biases located in the Appalachians and in the Plains states (Lu et al. 2011). In the cool season, COAMPS had cool biases extending from the Appalachians and the Mississippi Valley through Texas and warm biases throughout the rest of the Southeast and over the Gulf of Mexico (Lu et al. 2011). In both the warm and cool seasons COAMPS tended to overforecast precipitation with precipitation forecast skill decreasing as the intensity of the precipitation increased (Lu et al. 2011).

Cheng and Steenburgh (2005) assessed the forecast skill of the 48-hour CIRP WRF (12.5km grid spacing) and ETA models (12km grid spacing) in the Western United States from June to August 2003.The authors found that both the CIRP WRF and ETA models displayed positive warm biases with the largest biases occurring in the early morning and in the late afternoon (Cheng and Steenburgh 2005). Biases were found to vary diurnally with the largest warm biases occurring in the early morning for both models (Cheng and Steenburgh 2005). Although both models were found found to be too dry (dewpoint too low compared to observations), dry biases in the ETA model were greater than those in the CIRP WRF (Cheng and Steenburgh 2005). The authors additionally assessed wind speed forecast skill, finding that the CIRP WRF overforecast wind speed intensity while the ETA underforecast wind speed intensity (Cheng and Steenburgh 2005).

The utilization of higher resolution models, defined as a model with ≤10 km grid spacing (Mass et al. 2002), can reduce errors in wind speed, wind direction, and precipitation amounts associated with complex terrain by better resolving sharp gradients in elevation, gaps, and multiple ridges that modify air flows and orographic enhancement (Mass et al. 2002; Casaretto et al. 2022). In coarser resolution models, mountain top elevations are usually underestimated and smoothed out compared to the actual terrain. Grid resolution is also important to the models ability to correctly resolve mesoscale convection. Coarser resolution models parameterize convection (e.g. Kain 2004) while at finer grid spacings ≤4-6 km mesoscale convection can be handled explicitly (Mass et al. 2002). For regional modeling over the Pacific Northwest, Mass et al. (2002) found that forecast skill did improve when between a model with a 36 km grid and the same model with 12 km grid. However, comparisons between forecasts using a 12 km grid versus a 4 km grid did not yield a significant increase in forecast skill. Hence, while increasing model resolution improves the representation of topography and convection it does not guarantee that the associated forecast will be more accurate (Mass et al. 2002; Hoadley et al. 2004).

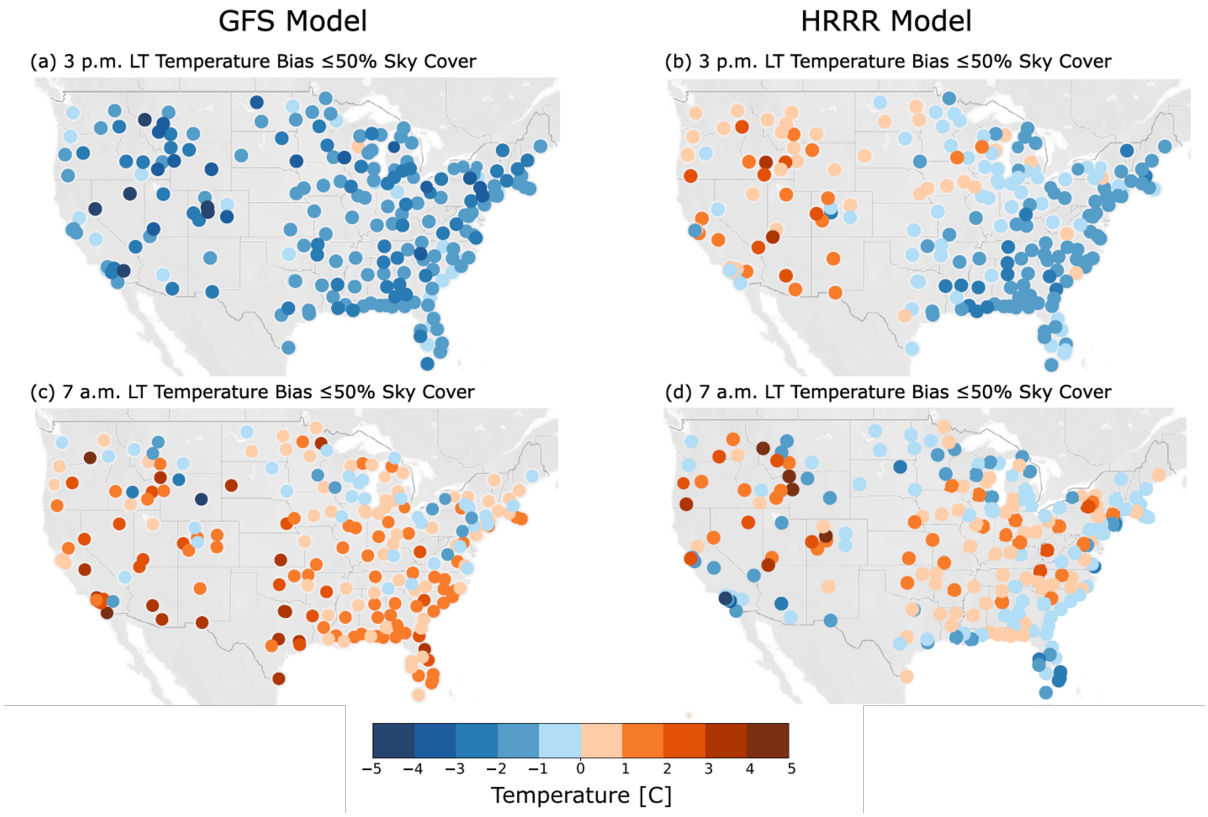

Figure 1.5: Bias of 36-h lead time temperature forecasts from GFS (a, c; left column) and HRRR (b, d; right column) for 210 airports over the period 1 Nov 2019 to 31 March 2020 for observed sky conditions with < 50% cloud cover valid at 3PM (a, b; top row) and 7AM (c, d; bottom row) local time. Blue colors indicate cold bias, red colors indicate warm bias. Adapted from Patel et al. (2021),their Figure 1.

This study follows and builds on the methodology of Patel et al. (2021). Patel et al. (2021) analyzed numerical weather prediction model forecast biases for the GFS and the HRRR during the winter season from 1 November 2019 to 31 March 2020. Their study examined the diurnal trend of temperature biases at 7AM and 3PM local time as functions of the amount of observed cloud cover, observed wind speed, and, when applicable, snow depth. Biases were examined at 7AM and 3PM to represent the diurnal variation of biases where 7AM represents the diurnal low temperature and 3PM represents the diurnal high temperature. They used 210 ASOS stations located at airports across the contiguous United States and examined temperature biases for the 36-hour leadtime. During the winter, the GFS model was on average about 1°C too warm at 7AM and 2°C too cold at 3PM in conditions with <50% cloud cover. At 3PM, sites across the US were consistently too cold in GFS but in the HRRR there was a regional variation with locations east of the Great Plains too cold and western locations too warm (Fig. 1.5). At 7AM, regional temperature biases were more variable than at 3PM. The largest magnitude cold biases in GFS and warm biases in HRRR were at locations in the Intermountain West (Fig. 1.5).

A key innovation in the model evaluation strategy of Patel et al. (2021) was the examination of model errors by similar weather conditions on many days as opposed to over date ranges with varied conditions or for a small number of case studies. Subsetting by weather condition helps to emphasize certain physical processes which aid in narrowing the diagnosis of potential error sources. For example, at 7AM site by site temperature biases increased with decreasing cloud cover and decreasing wind speed in both GFS and HRRR implicating deficiencies in the representation of nocturnal temperature inversions as a possible source of the large biases.

1.2 Research Goals

The Navy needs weather forecasts to plan training, ship routes, and combat operations. The Naval Research Laboratory can use information on model strengths and weaknesses to help determine priorities for revisions in model physics and forecast input fields. Additionally, information on systematic biases in the current version of the COAMPS model can be used by forecasters to adjust their predictions.

This study is an exploratory data analysis rather than hypothesis based. We assess several weather-conditioned biases within COAMPS and the GFS during the summer months (May - September 2022) and the winter months (November 2022 - February 2023) for the

US and Canada. This analysis evaluates forecasts of several key variables of interest to Navy operations, including temperatures, dewpoints, and winds. As part of investigating potential error sources, we also examine low pressure passage and precipitation event timing. Investigation of event timing helps to distinguish if a storm's structure is well represented but in the wrong place at the wrong time or if the storm's structure in the model is substantially different from what was observed.

Questions this study will address include:

- How do the temperature biases at 7AM (∼ daily low temperature) and 3PM (∼ daily high temperature) in winter compare to those in summer?
- Do regions and seasons with larger temperature errors also have larger errors in other variables such as dewpoint and wind speed/direction?
- Are seasonal biases in the matched model and observation temperature errors also found in comparisons of seasonal temperature distributions when the distributions of all forecast and observed temperatures are compared to each other?
- Are forecasts of the timing of low pressure passages and precipitation events within a reasonable margin of error?

1.3 Roadmap

In Chapter 2, we describe the data sets used in this study and the evaluation metrics used and developed for different variables. Analysis results will be discussed in Chapter 3 including biases in temperature subset by time of day, cloud cover, and by extreme events relative to a 30-year climatology. We will also examine biases in dewpoints, wind speed and direction, and the timing of low pressure passages and of precipitation events. By analyzing these variables, we will develop a wide ranging assessment of COAMPS overall forecast skill and assess how it is doing in comparison to the GFS under the same evaluation metrics. How COAMPS is performing in comparison to GFS will help us to better understand if a particular error is common to both models or not. Chapter 4 will address differences between COAMPS at two different grid spacings for Central California. This will help us to determine how COAMPS forecast skill changes depending on grid spacing. Chapter 5 will summarize the results of this study and future work.

CHAPTER

2

DATA AND METHODS

2.1 Data

This study analyzes the forecast skill of numerical weather prediction models. Several key terms used throughout this paper include: *initialization time*, the first valid time in a model run, *lead time*, the time between when a forecast for a specific time is initialized and when it is valid, and *valid time*, the specific time that a forecasted value occurs. For example, if a model run were initialized at 00 UTC on 25 May the first time in the model run would be 00 UTC on 25 May. If we then wanted a lead time of 48 hours, we would need to look 48 hours out from 00 UTC on 25 May to 00 UTC on 27 May.

For matched observation and model forecasts (Section 2.3), we assess the 48 hour lead time and for the bulk analysis (Section 2.4) we examine both 48 and 72 hr leadtimes. Model forecast skill typically increases at shorter leadtimes.

2.1.1 COAMPS model

The Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) was developed by the Naval Research Laboratory (NRL). Refinements are incorporated into a new

operational version of the model about once a year. COAMPS was first used in 1996 to analyze and forecast ocean sea-surface temperatures (Chen et al. 2003). It uses nonhydrostatic, compressible equations and a scheme C grid to represent atmospheric components using either real data (in operational and case study modes) or idealized data (in research mode) to model both large scale and mesoscale atmospheric phenomena on both short and long term time frames. Operational COAMPS runs are initialized every 12 hours (0000, 1200 UTC) and forecasts can extend outwards for up to 4 days or as short as 2 days. Parameterization schemes used for COAMPS are listed in Table 2.1. For this work, we used the operational COAMPS version 5.6 (S. Chen, personal communication). COAMPS output is not run through any type of model output statistics (MOS) bias correction before the data is released from the NRL (Chen et al. 2003). Soil moisture is initialized using NASA's Land Information System (LIS) and land cover uses the community NOAH Land Surface Model (LSM).

In this study we examine the forecast biases of several COAMPS model products, temperature, dewpoint, and wind speed/direction, and the timing biases of precipitation and low-pressure passage. Temperature and dewpoint values are forecast in the model at a height of 2m above ground level and wind speed/direction values at a height of 10m above ground level (Chen et al. 2003). Land ASOS stations take temperature/dewpoint data (2m above ground level) and wind speed/direction (10m above ground level). The heights of National Buoy buoy sensors vary but are usually between 3 and 5 m above the ocean surface (https://www.ndbc.noaa.gov/faq/bmanht.shtml).

COAMPS operational regional domains used in this study include the Northwest Atlantic (Eastern North America - NWATL) (Fig. 2.1), Northeast Pacific (Western North America - NEPAC) (Fig. 2.2), and Central California (CENCOOS) (Fig. 2.3). The COAMPS domains in North America do not include the Intermountain West region. COAMPS NWATL, NEPAC, and NWPAC all have a grid spacing of ∼15.5 km. COAMPS CENCOOS has a smaller grid spacing of ∼3.7 km. Model values at airports and buoys are determined using spatial linear interpolation for the ∼15.5 km grids and by nearest grid box for the ∼3.7 km grid.

To assess to what degree COAMPS forecast skill is impacted by changes in model grid spacing, in Chapter 4 we focus on the California region to compare forecasts from COAMPS NEPAC (lower resolution, larger grid spacing) to COAMPS CENCOOS (higher resolution, smaller grid spacing). NEPAC and CENCOOS have nearly identical model physics. The primary difference between NEPAC and CENCOOS model physics is that NEPAC uses a modified Kain-Frisch parameterization scheme to parameterize convection while CENCOOS does not (convection is explicit). The Navy's Fleet Numerical Meteorology and

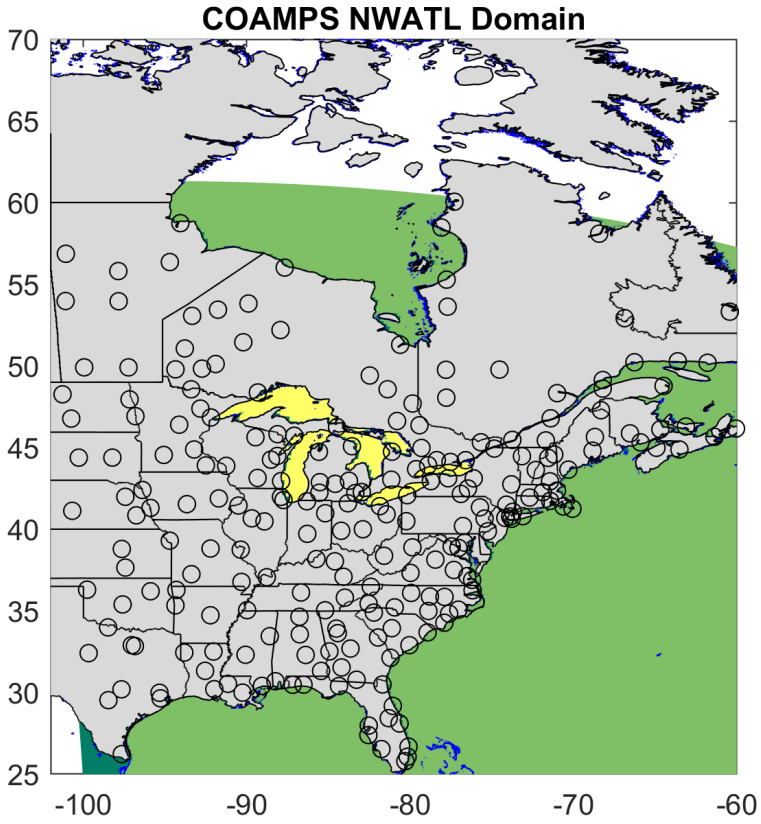

Figure 2.1: COAMPS North West Atlantic (NWATL) model domain. Black circles represent stations within the NWATL domain used in the bias statistical analysis. Gray shading represents the land surface type in the model, yellow shading represents the inland water surface type in the model, and green shading represents the ocean surface type in the model.

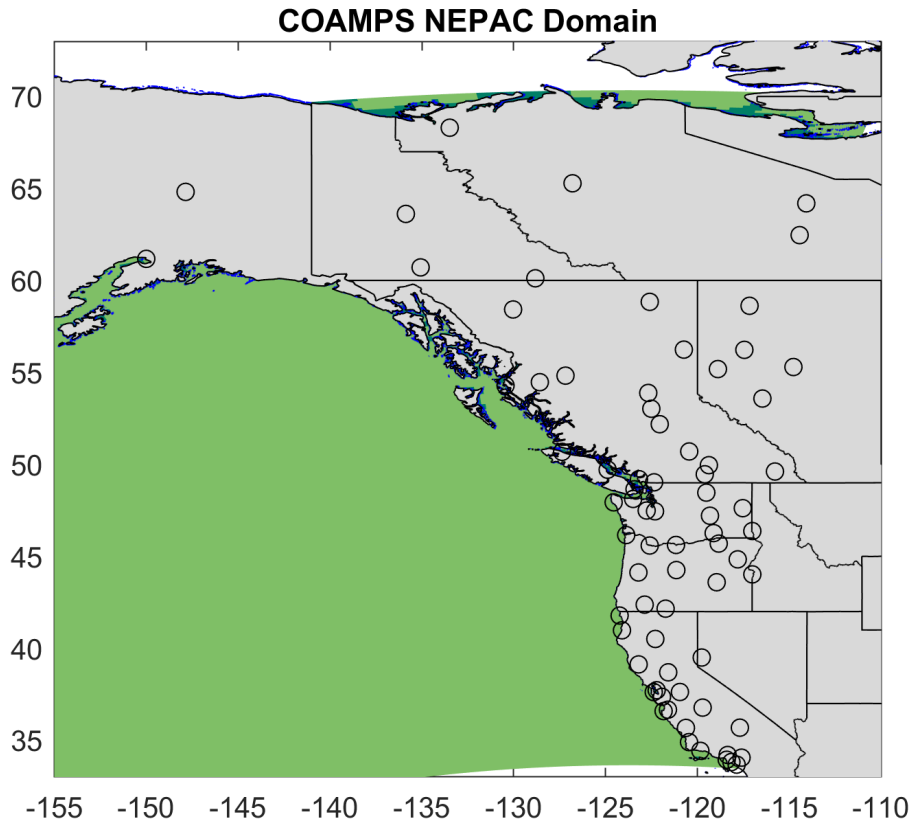

Figure 2.2: Same as 2.1 except for COAMPS Northeast Pacific (NEPAC) model domain.

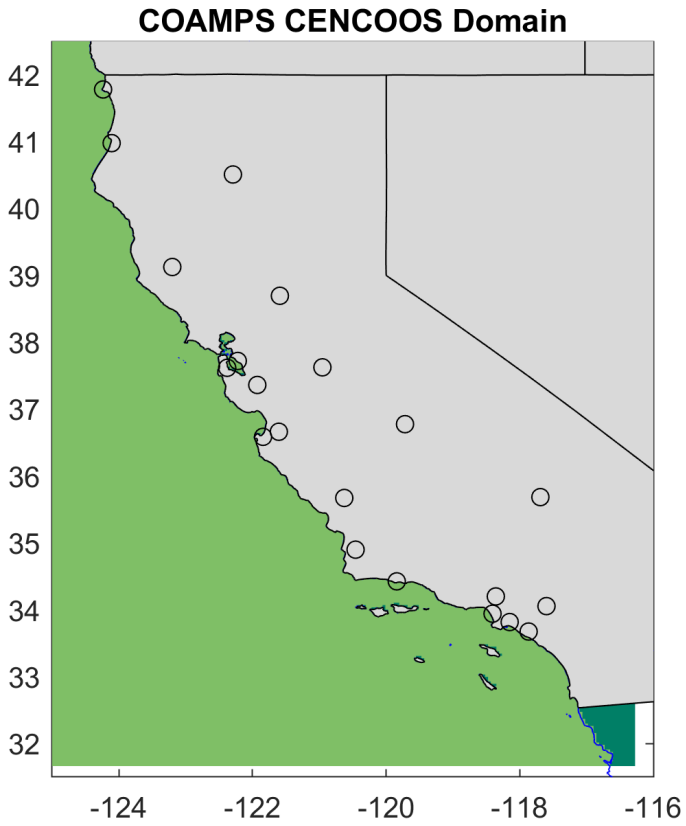

Figure 2.3: Same as 2.1 except for COAMPS Central California (CENCOOS) model domain.

Oceanography Center, which has responsibility for the COAMPS operational model, starts the CENCOOS runs shortly after NEPAC runs in their processing queue (J. Doyle, personal communication).

2.1.2 GFS Model

We used NOAA's Global Forecasting System (GFS) (AWS 2022) as a comparison metric to assess COAMPS forecast accuracy. The GFS model internally uses a 13 km grid and are output for distribution on a 0.25 degree (27-28 km) grid. GFS runs are initialized every six hours at 0000, 0600, 1200, and 1800 UTC. The model output used in this study from May 2022 to February 2023 utilizes GFSv16 which was instituted in March 2021 (AWS 2022). The 0.25 degree GFS data are obtained from the NOAA AWS cloud. GFS model values at airport and buoy locations are determined using spatial linear interpolation as in Patel et al. (2021). Parameterization schemes used for COAMPS and GFS are listed in Table 2.1.

GFS forecasts are run through the National Center for Environmental Prediction (NCEP) post processor, the Unified Post Processor (UPP), before being released to the public. UPP is used for all of NCEPs operational models and converts input model data to standard pressure level/heights and standard output grids (Center 2020). For both land and ocean stations, forecast temperature and dewpoint values are interpolated from the model bottom level to 2m above ground level and wind speed/wind direction values are interpolated to 10m above ground level.

2.1.3 Surface Weather Observations

Numerical weather model forecast output are compared to hourly observation surface data from Meteorological Terminal Air Reports (METAR) Automated Surface Observing Systems (ASOS) at airports, and National Buoy Data Center buoys (Table 2.2). Comparisons to model output in the COAMPS domains use 333 airports and 37 buoys across North America (Fig. 2.4). For GFS, an additional 88 stations across central North America are used for the region not covered by regional COAMPS domains. In accordance with Federal Aviation Administration guidelines (Durham 2020), stations in the Western United States and along the Appalachians are considered to be part of mountainous terrain (Fig. A.1). METAR data are filtered to have only one observation per hour for each variable. Hours with missing temperature data are filled in using the nearest value between the previous non-missing time and the nearest non-missing time. Large gaps (>6 hours) in the data

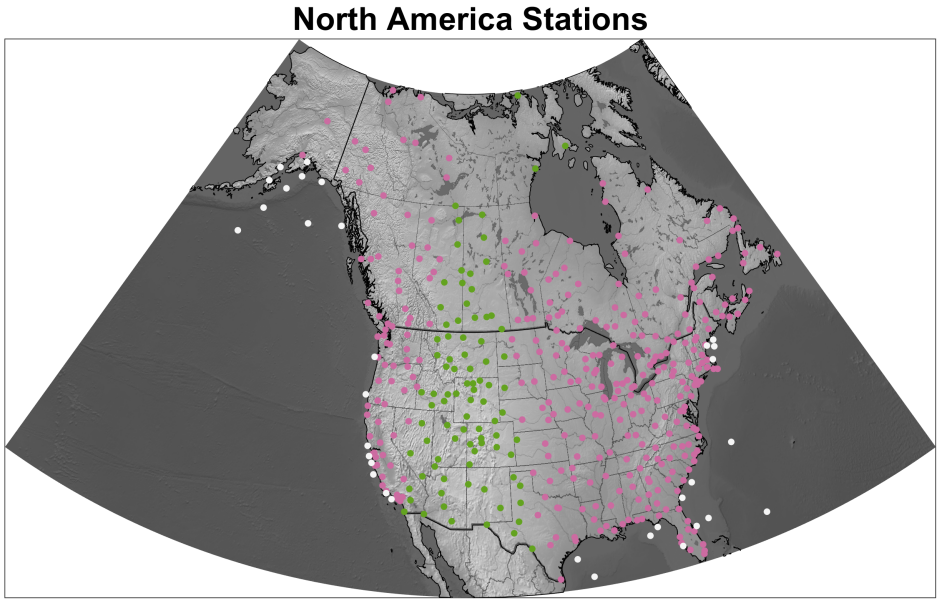

Figure 2.4: Map of ASOS land stations in North America that are compared to both COAMPS and GFS (pink) and GFS only (green). Buoys (white) are compared to both COAMPS and GFS.

	COAMPS	GFS
Grid Spacing	~15.5 km	27.5 km (0.25 deg)
Cumulus Parameterization	Kain-Frisch	SAS-based mass flux
Shallow Convection	Tiedtke-like scheme	SAS-based mass flux (shallow convection)
PBL/Turbulent Mixing	Modified 1.5 order Mellor-Yamada Scheme	Hybrid EDMF PBL and Free Atmospheric Turbulence
Microphysics	Rutledge and Hobbs (1984) Lin et al. (1983)	GFDL (sa)
Surface Layer	Louis et al. (1982) Modified by COARE algortithm	GFS PBL
Radiation (short/long wave)	Fu-Liou Radiation Scheme	RRTMG
Land Surface	NOAH LSM	NOAH LSM
Soil Moisture	NASA LIS	NOAH LSM

Table 2.1: COAMPS and GFS model characteristics and parameterizations. Based on AWS (2022) and J. Doyle personal communication

record are filled in with NaN (Not a Number) instead of using the nearest value. Large gaps in the data are not common in our analysis since we filter to remove stations with large quantities of missing data (discussed below). Missing values for gaps of <5 hours are filled in using the nearest non-missing value as done in Patel et al. (2021). Only five stations in the summer and three stations in the winter have gaps larger than six hours. METAR Aviation Selected Special Weather Reports (SPECI) are excluded. Data from the NOAA National Buoy Center buoys (NCEI 1971) located along the East and West Coasts of North America were used to provide insight into forecast skill over the ocean.

In order for a station to be included in the analysis, the station must have valid observations for at minimum 70% of the total number of hourly observations over the 5 month summer (153 days) or 4-month winter (120 days) period. This threshold criteria for representativeness impacts buoys more frequently than land stations as buoys are located in harsher conditions and instrument malfunctions occur more frequently.

Since model grid spacing is often much coarser than fine scale details in actual coastlines, coastal stations may be located in the model's ocean surface type instead of the land surface type. This misclassification of surface type prevents an accurate representation of the diurnal cycle at these stations as the model is forecasting for an ocean environment but

Available Data for Land and Ocean Stations		
Data Variable	ASOS - Land Station	Buoy - Ocean Station
Temperature	Yes	Yes
Dewpoint	Yes	Yes*
Cloud Cover	Yes	No
Pressure Trace	Yes	Yes
Wind Speed	Yes	Yes
Wind Direction	Yes	Yes

Table 2.2: Available variables for ASOS land stations and buoy data. Asterisk indicates dewpoint data is available for some but not all buoys.

Land Stations Located in the Ocean Surface Type		
NWATL	NEPAC	CENCOOS
KSRQ	KUIL	KCEC
KPHF	KAST	KOAK
KJFK	KCEC	
KACK	KACV	
	KOAK	
	KMRY	
	KSBA	
	KSNA	

Table 2.3: Airport land stations located in the ocean surface type in each COAMPS domain.

observations are taken on land. The number of land stations located in the ocean surface type varies with the grid specification (Table 2.3). For the state of California, NEPAC has six coastal stations located in the ocean surface type whereas CENCOOS has two coastal stations located in the ocean surface type. These misclassified stations are excluded from most of the analysis that follows. When these misclassified stations are included they are clearly labeled.

2.1.4 Historical Observations

To assess forecasts for stronger warm and cold events, the Integrated Surface Database (ISD) Lite (National Centers for Environmental Information 2021) is used to create thirty-year hourly air temperature climatologies for each station. This database is made up of global hourly and synoptic observations going back to the 1960's and 1970's. Similar to

NOAA's climate normal's (Arguez et al. 2012), our station by station climatologies use thirty-one years worth of data from 1991 to 2021.

Hourly temperature climatologies are calculated using the 31 observed hourly ISD Lite temperatures for a specified calendar day and time for each airport in our data set for North America. For example, 00 UTC on Jan 1 1991, 00 UTC on Jan 1 1992,..., to 00 UTC on Jan 1 2021. Figure 2.5 shows the hourly temperature climatology for Raleigh-Durham airport (KRDU). There are 8760 hours in a year, excluding leap-days. These "8760" hourly climatologies are often used as a first guess weather forecast particularly for long-lead time forecasts in applications that are sensitive to the diurnal cycle such as energy demand forecasting. Numerical forecast models are intended to add value compared to a forecast based on climatology by forecasting for events outside of the typical climatological range. Hence, we isolate periods when events outside of the 90th to 10th percentile climatological range occur as a separate weather condition to assess model skill.

2.2 Subsetting by Weather Conditions

We define several types weather conditions to subset data for this analysis: cloud cover amounts (Section 2.2.1), and temperature events outside the 90th to 10th percentile climatological range (Section 2.2.2). Filtering by weather conditions helps narrow potential "diagnosis" of bias sources.

2.2.1 Amount of Observed Cloud Cover

Following Patel et al. (2021), forecast data and observations are filtered into five different categories by the amount of observed cloud cover present at the same valid time. The cloud cover categories used in this study are: all recorded cloud cover observations, no cloud cover reported (CLR), less than 25% cloud cover (CLR, FEW), less than 50% cloud cover (CLR, FEW, SCT), and greater than 50% cloud cover (BKN, OVC)(Table 2.4). Subsetting by cloud cover category can help rule in and rule out possible error sources. For example, if the 7AM temperature biases are larger for periods with less cloud cover than for overcast conditions it suggests that the model's handling of radiation temperature inversions may be a factor in these errors (Patel et al. 2021).

Cloud Cover Categories				
All Conditions	CLR Conditions	< 25% Cloud Cover	< 50% Cloud Cover	> 50% Cloud Cover
All observations	CLR only 0 Oktas	CLR, FEW 0-2 Oktas	CLR, FEW, SCT 0-4 Oktas	BKN, OVC 5-8 Oktas

Table 2.4: Cloud cover categories used filtering by amount of observed cloud cover.

2.2.2 Temperature Periods Outside of 90th and 10th Percentiles

We use station by station hourly 31-year climatologies (Section 2.1.4) to determine the specific subset of hours when observed or forecast temperature values are outside of the observed 90th and 10th percentiles. An example showing how hours outside of the 90th and 10th percentiles are defined relative to the 31-year climatology from Norfolk, Virginia is presented in Fig. 2.6.

We examine four different kinds of events: when the observed temperature is above the observed 90th climatology percentile, when the observed temperature is below the observed 10th percentile, when the forecast temperature is above the observed 90th climatology percentile, and when the forecast temperature is below the observed 10th percentile.

In both the wintertime period (11/22 - 2/23) and the summertime period (5/22 - 9/22), there were more observed hours with temperatures greater than the 90th climatology percentile than observed hours with temperatures less than the 10th climatology percentile. The geographic distribution of the number of hours where observed temperatures were greater than the 90th climatology percentile is shown in Fig. 2.7 and the geographic distribution of the number of hours where observed temperatures were below the 10th climatology percentile is shown in Fig. 2.8. Both matched model - obs at the same valid time (Section 2.3) and bulk analysis statistics (by examining the overall distribution of temperatures) (Section 2.4) are calculated for events outside of the 90th and 10th percentiles.

2.3 Matched Model and Observations at Same Valid Time

2.3.1 Leadtime-ish Method

Following Patel et al. (2021), we examine the diurnal variation of meteorological biases by examining model errors at the time of the approximate local daily climatological high

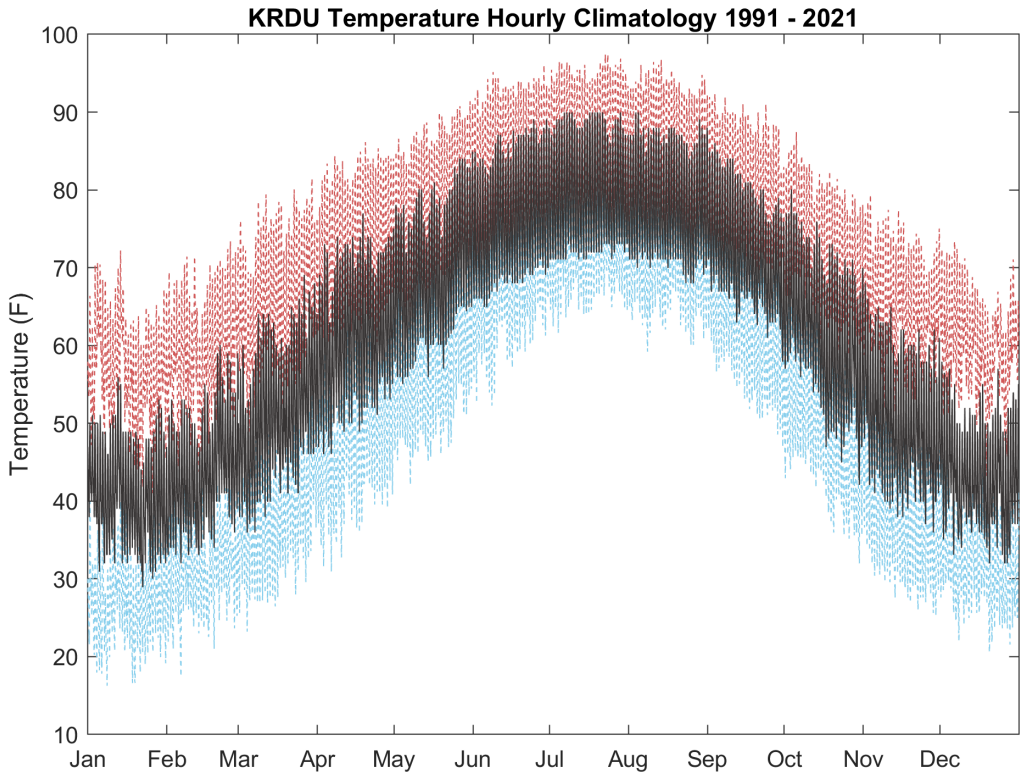

Figure 2.5: Hourly climatology for KRDU (Raleigh, NC) from 1991 to 2020. Solid black line represents the median hourly temperature over this period. Dashed red line represents the 90th percentile for temperatures over this period. Dashed blue line represents the 10th percentile for temperatures over this period

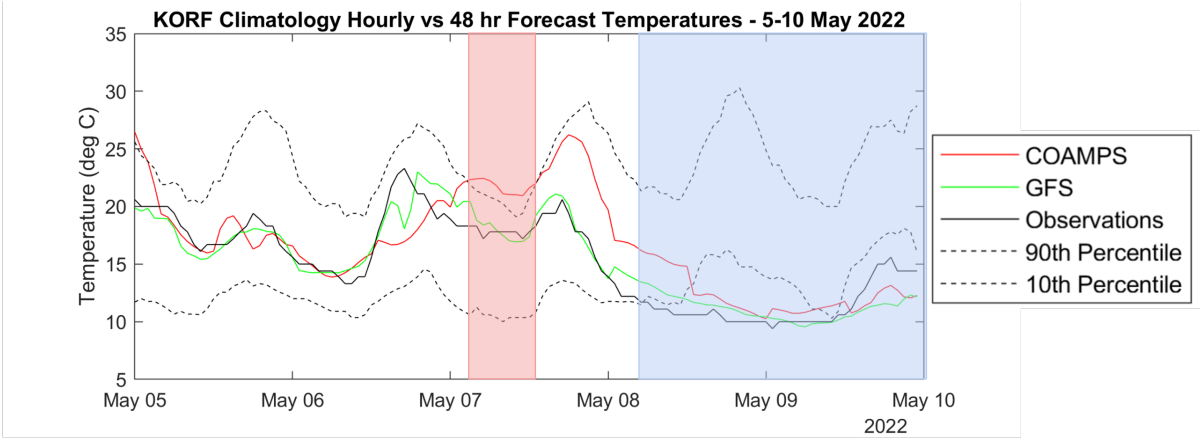

Figure 2.6: Time series of forecast temperatures and observed temperatures from 5-10 May 2022. GFS forecast is marked in green, COAMPS forecast is marked in red, observations are marked in solid black, and the black dashed lines represent the 90th (upper line) and 10th (lower line) percentiles. Times when the COAMPS forecast temperature is greater than the climatological 90th percentile are shaded in red and times when the observations and/or forecasts are below the climatological 10th percentile are shaded in blue.

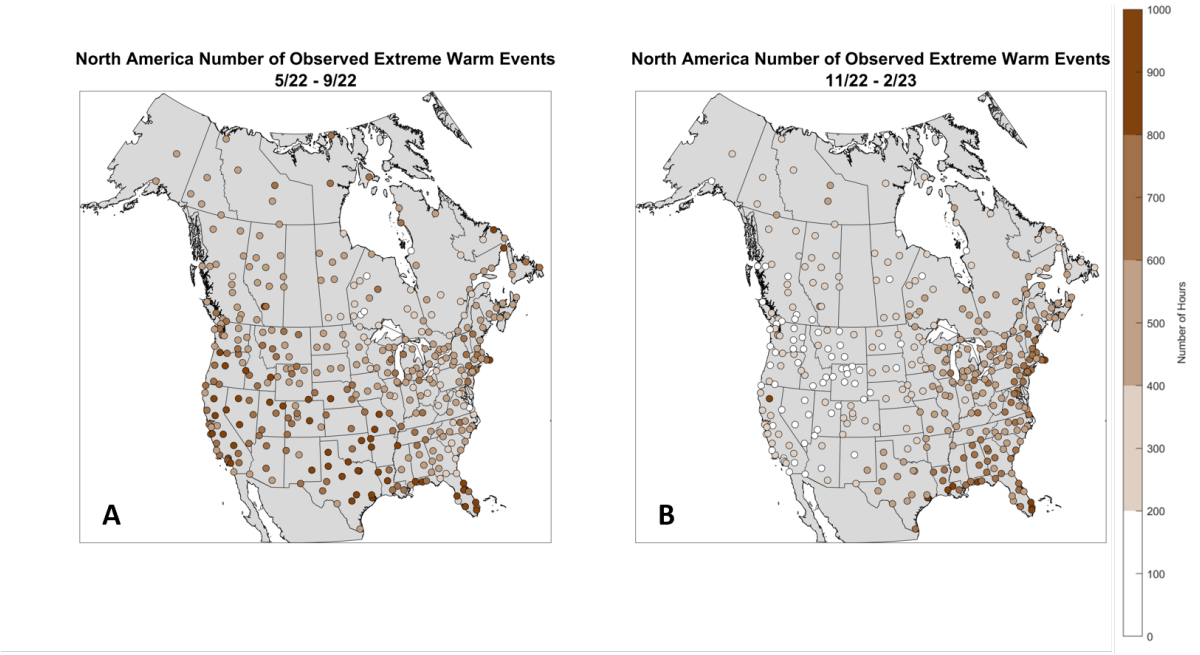

Figure 2.7: Number of observed hours with temperatures >90th climatological percentile in North America for A) 5/22 - 9/22 and B) 11/22 - 2/23.

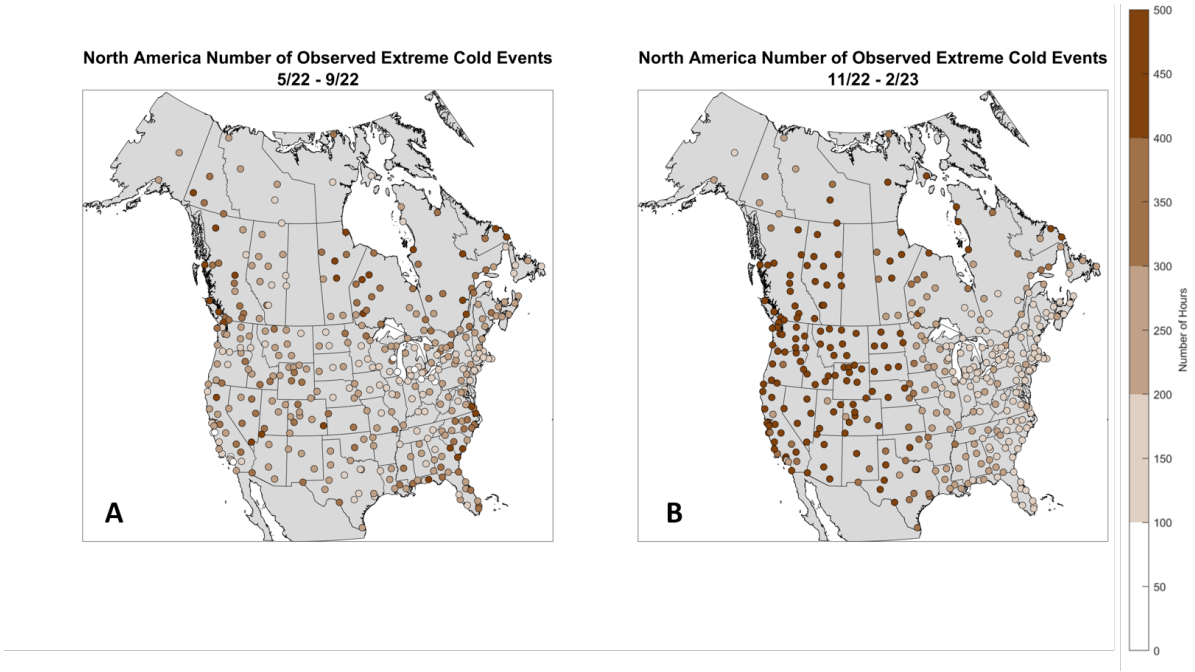

Figure 2.8: Number of observed hours with temperatures <10th climatological percentile in North America A) 5/22 - 9/22 B) 11/22 - 2/23. Note that the range of the color scale is half of what it is in Fig. 2.7.

temperature (3PM LT) and low temperature (7AM LT) temperatures in North America in winter. The equivalent local clock times shift to 4 PM LT and 8 AM LT in the summer with Daylight Savings Time. The longitudinal boundaries and UTC times used for different time zones are shown in Table 2.5.

Related to the initialization times, the diurnal low and high temperatures do not always coincide with a 24/48/72 hour leadtime. For example, if the models are initialized at 00 and 12 UTC, then hours other than 00 and 12 UTC do not have an exact corresponding 24, 48, or 72 hour forecast. This becomes an issue when examining model error statistics at 3PM and 7AM local time as the equivalent UTC times do not occur at 00 and 12 UTC consistently across all regions in our domain. To account for this issue, we use the "leadtime-ish" method developed by Patel et al. (2021). In order to provide one forecast value for each valid time within the requested time frame, the leadtime-ish method first locates the 48-hour forecast within one model initialization run. It then finds the eleven nearest hours directly ahead of the 48-hour forecast and considers those eleven valid forecasts to be part of the overall 48-hour leadtime forecast. For example, the following would all be considered part of the 48-hour leadtime-ish forecast; a leadtime of 37 hours at 13 UTC, 38 hours at 14 UTC, 39

hours at 15 UTC, 40 hours at 16 UTC, 41 hours at 17 UTC, 42 hours at 18 UTC, 43 hours at 19 UTC, 44 hours at 20 UTC, 45 hours at 21 UTC, 46 hours at 22 UTC, 47 hours at 23 UTC, and 48 hours at 00 UTC. For a station at longitude of -95°, the UTC time corresponding to the daily minimum temperature is 13 UTC (Table 2.5). In this case, the 37-hour lead time forecast is used to represent the 48-hour lead time in the model error statistics. The leadtime-ish method then repeats every 12 hours, moving on to the next initialization run when the requested leadtime is reached, until the end of the time period. This method allows for comparison of model values to the observed values at the approximate climatological low and high temperatures over a 24-hour period.

Corresponding to time zones, there is a 3 hour difference in the forecast leadtime associated with specific local times in the Eastern time zone versus the Pacific time zone. For example, in the winter season, using a model initialized at 1200 UTC, 7AM local time corresponds to 1200 UTC (a 48 hour lead time) in the Eastern time zone and 1500 UTC (a 51 hour lead time) in the Pacific time zone. Geographic dot plots of biases (see Section 3.1.2) show larger biases in the western United States than the eastern US. The primary cause of these differences is more likely to be elevated, complex terrain in the west and not a few additional hours of leadtime. Among factors contributing to forecast biases, the machine learning method used by Bouallègue et al. (2023) found that station elevation was key contributor (second in importance in 10-m wind forecasts and 6th in importance in 2-m wind speed forecasts) out of 43 factors considered. To untangle the roles of terrain versus few hour lead time differences, regions outside the US where mountains are to the east and flatter areas are to the west would need to be examined which is beyond the scope of this thesis.

2.3.2 Calculation of Matched Time Model Error and Model Bias

Matched time model-observation errors are calculated by comparing the forecast model value to the observed value at the same valid time (Equation 2.1) A positive model error indicates that the model forecast a higher value than was observed. A negative model error indicates that the model forecast has a lower value than was observed. We use distributions of model errors to evaluate the model performance at specific locations and within geographic regions.

$$\text{Model Error} = \text{Forecast Value} - \text{Observed Value} \tag{2.1}$$

UTC Times Used Based on Longitude		
Longitude Range	Time of Estimated Daily Maximum Temperature	Time of Estimated Daily Minimum Temperature
> -82.5°	20 UTC	12 UTC
-82.5° < lon ≥ −97.5°	21 UTC	13 UTC
-97.5° < lon ≥ −112.5°	22 UTC	14 UTC
-112.5° < lon ≥ −127.5°	23 UTC	15 UTC
-127.5° < lon ≥ −175°	0 UTC	16 UTC

Table 2.5: UTC times used to represent daily high and low temperature by longitude for North America. Additionally, regions in Northern Canada from longitude > -141.5°W but < -101°W and at a latitude > 60°N uses a max temperature time of 21 UTC and a minimum temperature time of 13 UTC.

The overall station median bias is calculated by finding the median model error over the specified time period. The median bias is often a better metric than mean bias as the median is not as impacted by distribution outliers and better represents skewed distributions. For normal distributions median bias equals mean bias. Station by station analysis of the median bias is useful to detect geographic regional bias patterns in model output.

2.4 Bulk Distribution Analysis

An issue with error analysis based on matched forecast value and observed value at the same valid time is that it can misconstrue magnitude errors with timing errors. Under the matched valid time analysis, if a storm was forecast to arrive 4 hours too early but otherwise accurately forecast temperatures associated with the storm, the biases would be a result of the timing offset even though the magnitudes were accurately forecast. Bulk distribution analysis helps to separate out timing errors from magnitude errors. When persistent model biases are present, they will be present in the bulk comparisons between the model forecast distribution and the observed distribution.

The bulk distribution analysis method looks at the statistical distributions of key variables (i.e. temperature, dewpoint, wind speed and direction) at different leadtimes. For example, at a given station using the 48 hour leadtime-ish method, the temperature distributions of all initialized runs within a time period of one month are examined in comparison to the observed temperature distribution to identify how similar or dissimilar the distributions are. This analysis is done over all forecast hours found using the leadtime-ish

method and does not focus only on distributions at 7AM or 3PM local time. Bulk analysis can be done over varying time periods from one month to several months and for single stations at a time or for regions (e.g. Fig. 2.9). Bulk analysis comparisons can highlight if the model is better or worse at forecasting the overall distribution of a variable in a specified month or season. COAMPS forecast bulk analysis distributions are compared to the bulk analysis distribution of observations and to the forecast value distributions from NOAA's GFS, HRRR, and NAM. As part of digging into potential error sources, bulk distribution analyses were done for selected land-based stations and for buoys located along the coast of North America (Section 3.1.2).

2.5 Low Pressure Event Timing

From a user perspective, a basic metric of model performance is how well storm timing is forecast. We developed a new method to determine storm timing using pressure tendency (change in pressure per hour) to identify the time of the observed and the forecast local minimum pressure at each station. The local minimum in a surface pressure time series associated with low pressure events is usually sharper and more clearly associated with a specific hour than a local maxima for high pressure events.

Analysis of pressure tendency is ill-suited to the leadtime-ish method (Section 2.3.1) to match observation and model output for a specified lead time. Two consecutive model runs initialized at different times (e.g. 00 UTC and 12 UTC) can yield different forecast values for the same valid time due to changes in the initial conditions input into the model at the time of model initialization By piecing together model runs from different initialization times, the leadtime-ish method could display a "model initialization shock", a sharp jump, either upwards or downwards, in the forecast values which would not be representative of the variable's tendency for that station. Model initialization shock is more relevant when there are large differences in first-guess pressure values input into the model, such as when a low-pressure system passes through the region or frontal passage occurs. We found that model initialization shocks yielded spurious pressure changes that were difficult to filter out. We examine model pressure tendencies that are calculated over each of the 8 initialization runs that are relevant for the valid time separately rather than for specific leadtimes. For example, for a valid time of 4 Feb 2023 at 12 UTC, the following initialization times would be used 00 UTC 1 Feb, 12 UTC 1 Feb, 00 UTC 2 Feb, 12 UTC 2 Feb, 00 UTC 3 Feb, 12 UTC 3 Feb, 00 UTC 4 Feb, 12 UTC 4 Feb.

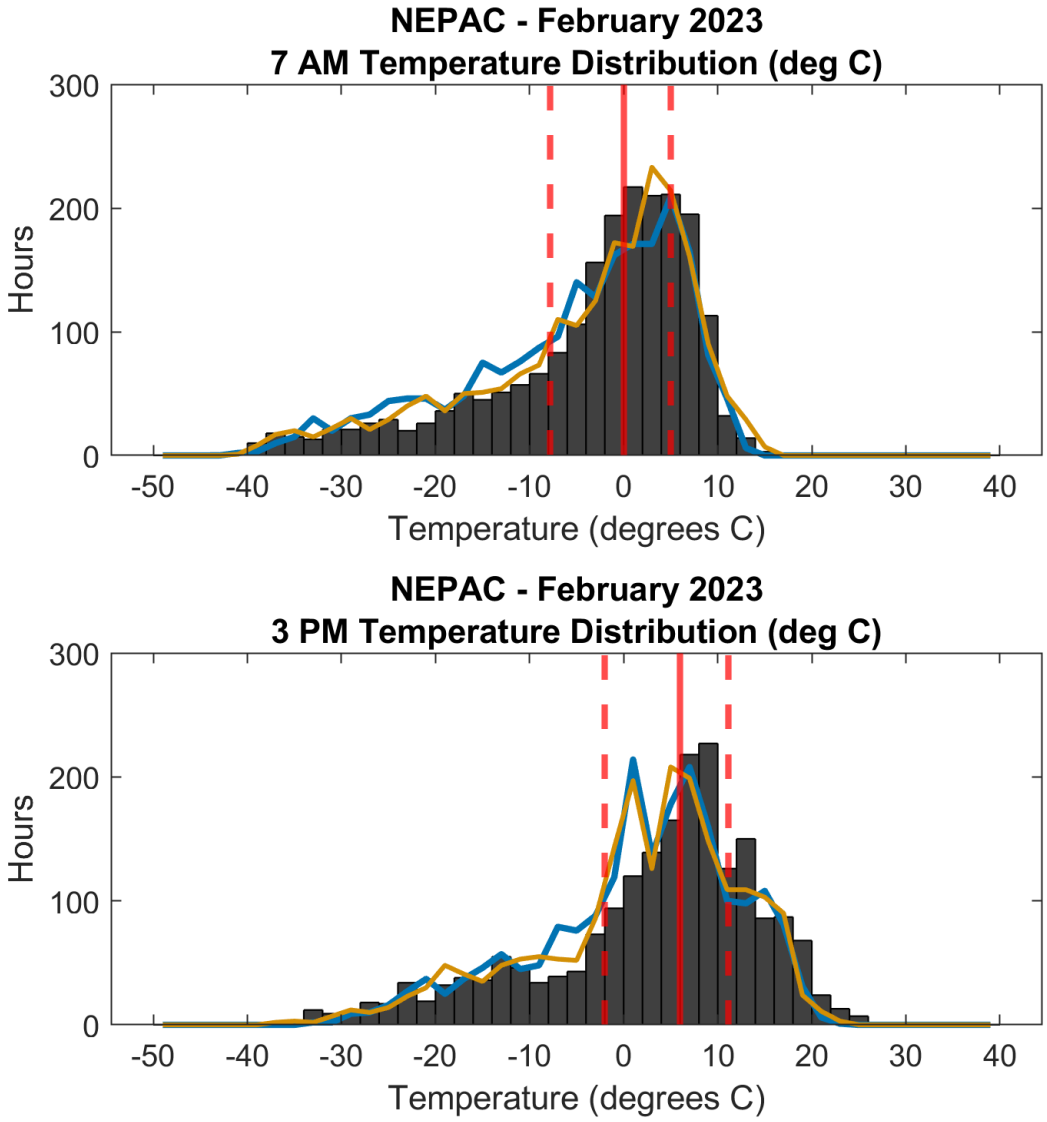

Figure 2.9: 48-hour lead time bulk analysis distributions for the North East Pacific COAMPS domain (Western North America) land stations in February 2023 at (A) 7AM LT and (B) 3PM LT. Solid black boxes represent the distribution of observed temperature, blue line indicates distribution of forecast COAMPS NEPAC temperatures, and orange line indicates distribution of forecast GFS temperatures (orange). Dashed red lines represent the 25th and 75th percentiles of observed temperatures and solid red line represents the median observed temperature value. The y-axis represents the number of occurrences, or number of hours, that a certain value was observed for.

2.5.1 Pressure Time Series Filtering and Pressure Tendency Calculation

In the initial steps in the pressure time series processing, demonstrated in Fig. 2.10 KORF (Norfolk, VA) for 21-24 January 2022, data is converted to a perturbation pressure (Fig. 2.10 A-B) by using a variation of the MATLAB detrend function to remove the mean value over the entire run from each individual pressure trace. Model output and observation surface pressure time series at each location for a given initialization time are processed for the entire run (a complete model run for COAMPS is 97 hours). GFS model runs extend out to 120 hours but to keep the length of time examined between COAMPS and GFS consistent we limited the length of complete GFS model runs to 97 hours to match COAMPS.

Application of frequency filtering on a time series requires further data conditioning. The perturbation pressure data from times 1 to n, where n is the length of the model run, are replicated such that the first third of the new time series is a copy of time 1 to n, and the last third is also a copy from 1 to n. (Fig. 2.10 C). Replicating the data in this manner is a common practice in frequency filtering.

A frequency bandstop filter, from the MATLAB signal processing toolbox, is then applied to remove the atmospheric tide signal (small diurnal up and down variations in the pressure data that are not reflective of the overall pressure pattern) from the data. The bandstop filter is designed to remove oscillations that occur between periods of 11 and 13 hours which is equivalent to 1.84 and 2.18 cycles per day. This serves to remove pressure oscillations that occur approximately every twelve hours which represents the largest magnitude frequency component of the approximate semi-diurnal atmospheric tide signal. It is necessary to mitigate the impact of the atmospheric tide signal to better observe the lower frequency pressure fluctuations associated with synoptic scale low pressure center passages. These steps are shown in Fig. 2.11.

The 1.84 and 2.18 cycles per day bandstop filter is not meant to remove hydrostatic atmospheric pressure changes associated with the diurnal temperature cycle. The effects of the diurnal temperature cycle were found to impact stations especially in the southwest US during the summer months. Future work will assess the utility of adding a secondary filter to reduce diurnal temperature influences on calculated pressure tendency values.

Pressure tendency is calculated using the filtered perturbation pressure data as the change in pressure over a change in time. For example, a five hour pressure tendency would take the pressure value at 04 UTC and subtract the pressure value at 00 UTC from it before dividing the total change in hours between the time times (Equation 2.2)). Pressure tendency values are centered so that the calculated pressure tendency is plotted halfway

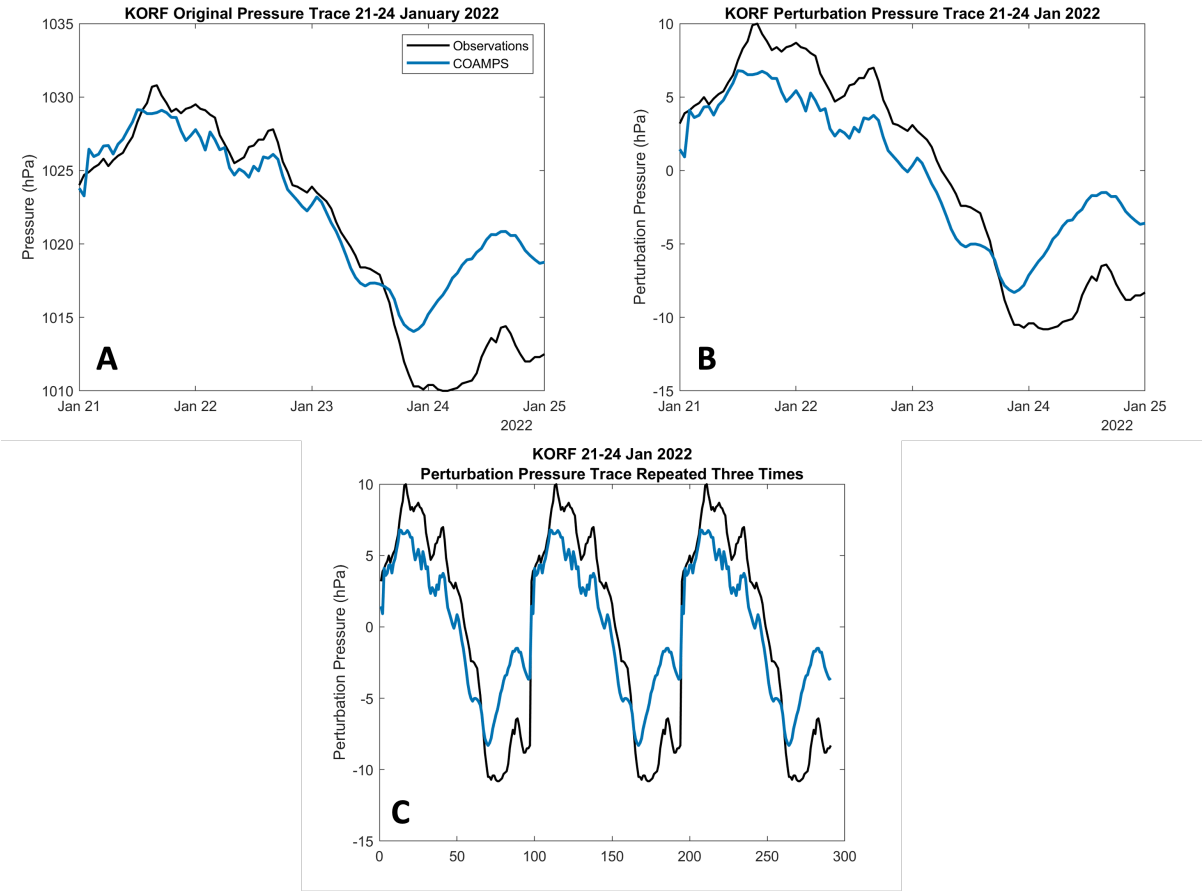

Figure 2.10: An example showing the steps to process pressure time series for COAMPS forecast pressure data (blue lines) and observed pressure data (black lines) for 21-24 January 2022 at station KORF. Panel A) shows the original pressure time series, panel B) shows the perturbation pressure time series, and panel C) shows the perturbation pressure time series repeated 3 times which is used as input to the bandstop filter step (see text for details).

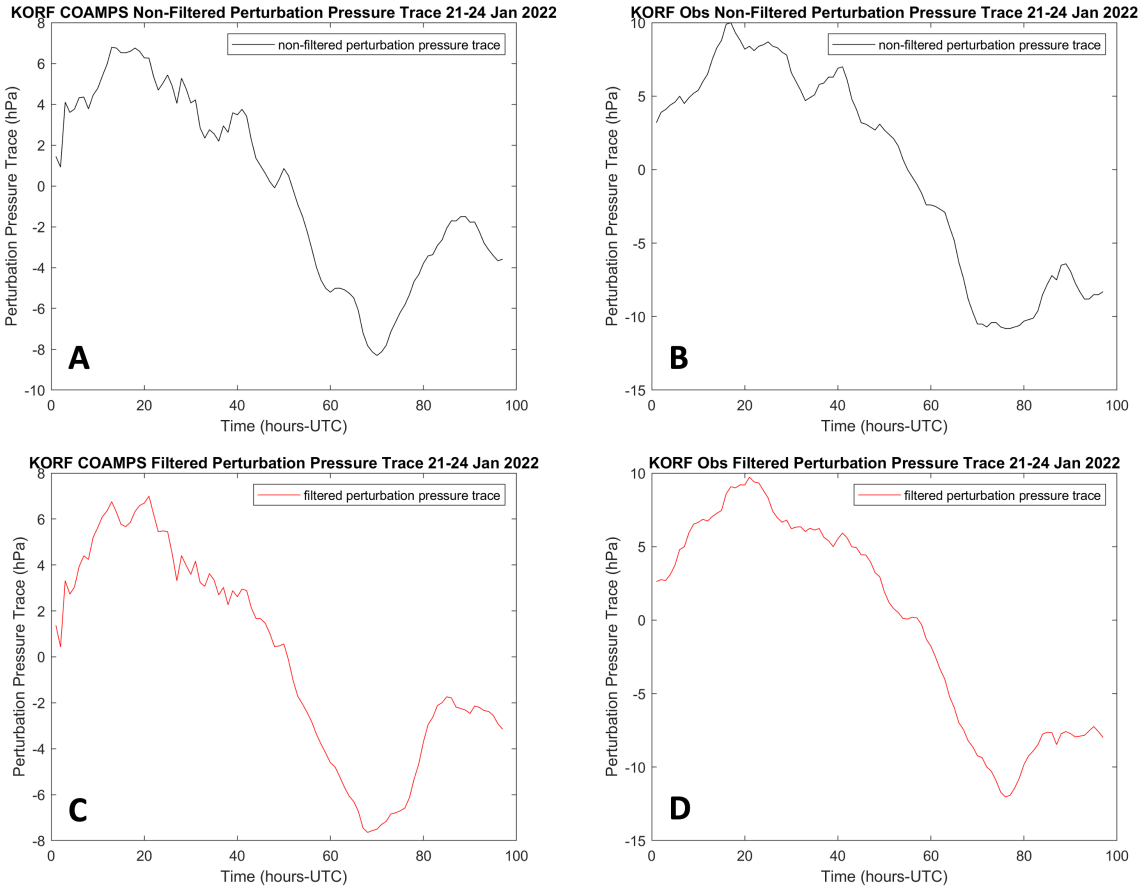

Figure 2.11: An example showing before (A, B; top row) and after (C, D; bottom row) bandstop filtering of COAMPS forecast perturbation pressure trace (A, C; left panel) and observed (B, D; right panel) pressure data from 21-24 January 2022 at KORF (Norfolk, VA).

between the start and end times of the requested period.

$$\text{Pressure Tendency} = \text{Change in Pressure} \,/\, \text{Change in Hours} \tag{2.2}$$

A negative pressure tendency indicates that pressure is decreasing and a positive pressure tendency indicates that pressure is increasing. Pressure tendencies calculated over 1, 3, or 5 hours were noisy and did not reflect the overall pressure pattern whereas the 11-hour pressure tendency was found to smooth the pressure tendency values too much and did not accurately reflect the overall pressure pattern (Fig. 2.12). We used the 9-hour pressure tendency as it provided the smoothest overall change in pressure patterns without oversimplifying the data. This acts as a low pass filter with a primary function of smoothing the data and eliminating the influence of noise and smaller scale weather features on the pressure tendency trend. This helps to remove false low pressure passages (discussed below in subsection 2.5.2) and enables easier identification of true low pressure passages.

2.5.2 Low Pressure Passage and Offset Calculations

A low pressure passage is identified as when the 9 hour pressure tendency switches from being continuously negative to continuously positive, indicating that pressure is no longer decreasing and has started increasing again.

Once a low pressure passage is identified in either the model or the observations it must meet two additional criteria before being considered further. The first criteria is that the observed pressure value (original) at the time of low pressure passage must be less than 1012 hPa. The 1012 hPa threshold is considered representative of the typical sea level pressure. This filter prevents low-pressure passage from being reported when there is not actually a low-pressure system present. The second criteria that must be met is that for the three hours prior to a low-pressure passage, pressure must continuously be decreasing (negative pressure tendency) and for three hours after low-pressure passage, pressure must be continuously increasing (positive pressure tendency). This filter serves to eliminate intermittent noise in the pressure tendency data where there are brief crossovers from negative to positive tendency before pressure continues decreasing again. Brief crossover events from negative to positive pressure tendency would not be considered a low-pressure system passage. A example illustrates the differences between a low pressure passage and several non-low pressure passages based on the 9-hour pressure tendency for KCHO (Charlottesville, Virginia) from 1-4 February 2022 (Fig. 2.13).

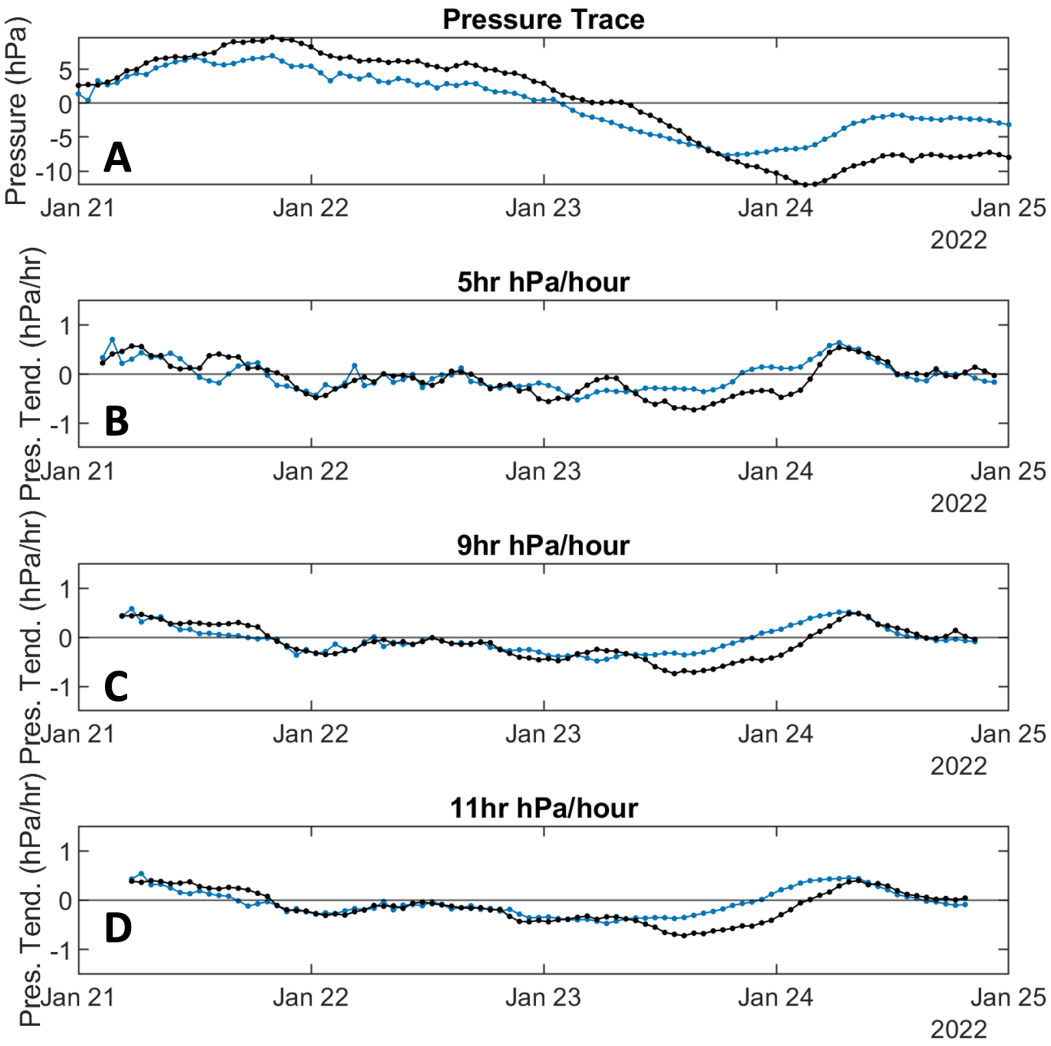

Figure 2.12: Time series of filtered perturbation pressure trace (A), 5-hour (B), 9-hour (C), and 11-hour (D) pressure tendency for 21-24 January 2022 at KORF. Blue line represents COAMPS data and black line represents observed data.

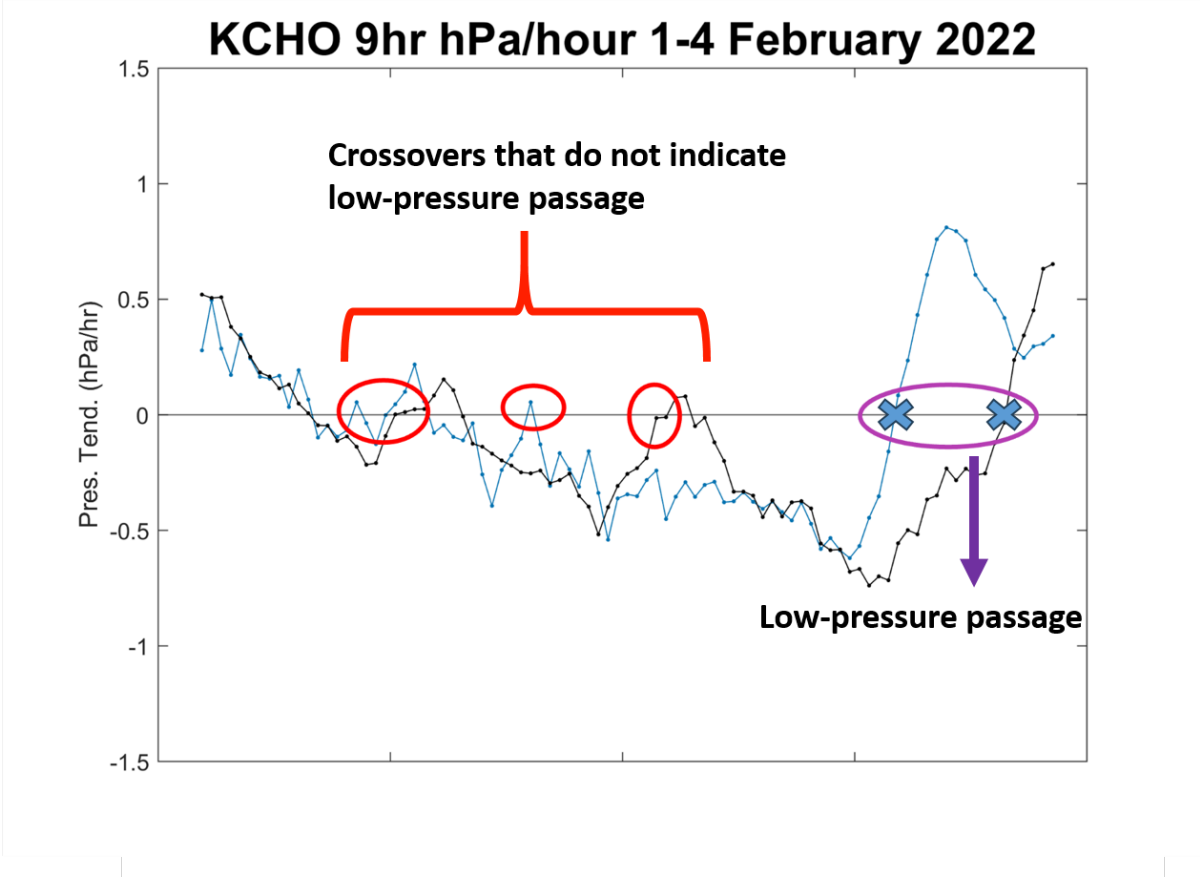

Figure 2.13: Time series of 9-hour observed (black) and COAMPS forecast pressure tendency (blue) for KCHO (Charlottesville, Virginia) for 1-4 February 2022 (COAMPS initialization time is 1 Feb 2022 at 00 UTC). Pressure tendency crossovers from negative to positive that do not continue to positively increase for at least three hours after switching sign (circled in red) do not represent low pressure passages. Pressure tendency crossover where pressure switches from being continuously negative to continuously positive for more than 3 hours before/after pressure tendency changes sign indicates a forecast/observed low pressure passage (circled in purple.

We developed a pairing code to match a forecast low pressure passage hour with an observed low pressure passage hour to see if they represent the same event. If a forecast low pressure passage occurs within +/- 24 hours of an observed low pressure passage they are considered to represent the same event. If a forecast low pressure system passage does not have an associated observed passage, it is considered a "forecast passage - no event" and if an observed low pressure system passage does not have an associated forecast passage it is considered to be an "observed passage - no forecast." The time of the observed low pressure passage is subtracted from the time of the forecast low pressure passage with the offset being either positive (the model forecasted the low to arrive too late) or negative (the model forecasted the low to arrive too early). We studied the distribution of these offset values and the median offset on a station by station analysis over North America (Section 3.5).

2.6 Wind Speed and Direction Error Criteria

Unlike scalar variables like temperature and dewpoint, wind speed and wind direction combine into a vector that is more difficult to compare between observations and model output. As observed wind speed and wind direction values can vary greatly over the course of an hour, there is the potential for more noise in biases resulting from the direct comparison of forecast wind speed/direction values to observed wind speed/direction values. This noise could mask the overall wind speed/direction biases and make it difficult to assess how well the models are forecasting in windy and non-windy conditions. An additional factor is that observed wind direction is more reliable at higher wind speeds and becomes unreliable for low wind speeds (approx. below 3 m/s) (Lavdas 1997) which can also create noise when forecast and observed wind direction values are compared directly. It is important to accurately study errors in wind speed and direction as they can be contributing factors to temperature and dew point errors particularly in locations where small-scale thermal circulations such as land/sea breezes and mountain/valley breezes occur.

Given that wind speed/direction errors can be noisy, we focus only on wind speed and direction errors that are large enough to have impacts on aviation as defined by Terminal Aerodrome Forecast (TAF) amendment criteria. A forecast wind speed error meets TAF amendment criteria if the difference between the forecast wind speed and the observed wind speed is greater than or equal to 5.14 m/s (10 knots) (Department of the Air Force 2020). A forecast wind direction error meets TAF amendment criteria if the difference between the

forecast wind direction and the observed wind direction is greater than 30 degrees when winds greater than 7.71 m/s (15 knots) are forecast to occur (Department of the Air Force 2020).

All stations will meet TAF amendment criteria for wind speed and direction for at least some portion of the overall winter/summer period. We then focus our study on stations that meet TAF amendment criteria at a more frequent rate by assessing the total percentage of time, over the winter or summer period, that TAF amendment criteria was met at a particular station. The higher the percent of time that TAF amendment criteria was met, the more commonly the wind speed or wind direction are problematic for that station. If a station meets TAF amendment criteria less than 2% of the time, the station is considered to have minimal wind speed/direction biases. If the percent of time is greater than or equal to 2%, the station is considered to have notable wind speed and/or direction biases (Section 3.3).

2.7 Precipitation Event Timing Analysis

We analyze forecast precipitation events start and end times in comparison to the observed start and end times following the methodology in Fritz et al. (2023) with modifications (personal communication, M. Miller). Precipitation data is taken from NCEPs Meteorological Assimilation Data Ingest System (MADIS) for land ASOS stations across the CONUS. To identify when precipitation is occurring, we use a threshold > 0.01 in/hr (.254 mm/hr) of liquid equivalent precipitation. We use a threshold of 0.01 in/hour, the smallest measurable amount by a tipping-bucket rain gauge, so as to have similar definitions of start and end of precipitation events between the observations and the model. Models occasionally output very tiny rain rates of 0.005 in/hr which is much less than can be measured by a rain gauge.

We are interested in the "envelope" of the precipitating period. Precipitation at a particular location can have breaks even though it is part of the same large storm or set of related storm cells. To address these gaps in continuity, hours with precipitation that are separated by less than 5 hours are grouped as the same event. For example, grouping precipitating hours together with gaps will combine isolated and scattered precipitation into the same event (Fig. 2.14). After identifying each event, the function finds the event start time, end time, and duration of the event.

The time series of observed and forecast precipitation events are compared to deter-

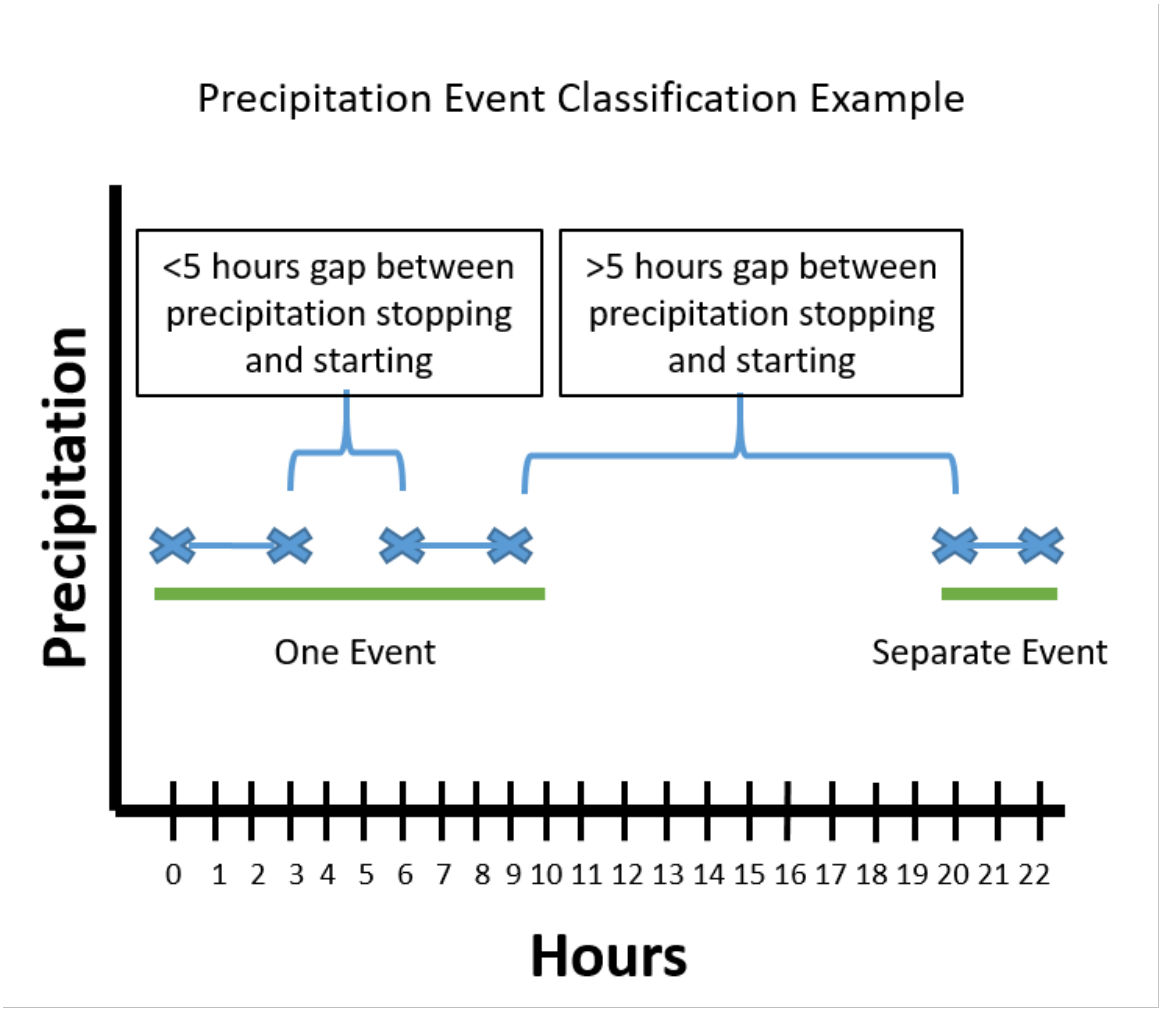

Figure 2.14: Example of envelope classification for precipitation events used in precipitation event timing analysis. Individual precipitation start and end times are denoted as a blue x with individual event duration marked as a solid blue line. Solid green line represents the classification of a larger precipitation event by combining individual events with short gaps in the "envelope" of precipitation into one larger event.

mine which events are paired (personal communication, M. Miller). Specific observed precipitation events are paired with the one closest in time forecast event in time within +/- 18 hours. If two, or more, forecast events occur within +/- 18 hours, only the forecast event closest to the observed event will be paired together and the remaining forecast event will be paired with the next closest observed event. This system pairs only the observed and forecast events that are closest in time to each other to prevent observed and forecast events from being paired incorrectly. Some observed precipitation events have no matching forecast and some forecasted precipitation events have no matching observation. These unpaired events are analyzed separately.

We analyze precipitation event timing biases in Section 3.6 by comparing the model start time to the observed start time, model end time to the observed end time, based on only the end times associated with paired start times, and model event duration to the observed event duration. We additionally calculate how many non-paired, or missed, observed precipitation events occur and how many non-paired, or incorrectly forecast, model precipitation events occur. In the case of stations that had precipitation events that occurred on the last day of the summer/winter periods (30 September or 28 February) and did not have an associated end time event (precipitation event continued after the requested period ended), the associated precipitation event start time was removed from the event start time list and not considered in this study.

CHAPTER

3

RESULTS 1 - ASSESSMENT OF ERRORS IN COAMPS AND GFS ACROSS NORTH AMERICA

3.1 Temperature

We examined the diurnal variation of median temperature biases for COAMPS and GFS in summer 2022 and winter 2022-2023 in the morning and afternoon and assessed to what degree temperature biases are tied to the amount of observed cloud cover. For weather stations in CONUS, Patel et al. (2021) found that at a 36-hour leadtime, GFS over the winter period of November 2019 to March 2020 was consistently too cold in the afternoon and frequently too warm in the morning. The work done in this study reaffirms the tendency for GFS to be too warm at 7AM and too cold at 3PM for the 48-hour lead time over the November 2022 to February 2023 season within CONUS. We also confirmed another result from Patel et al. (2021) that larger magnitude temperature biases in winter occurred with $< 25\%$ cloud cover as compared to all cloud cover conditions. Building on Patel et al. (2021),

this study extends the geographic region of the analysis into Canada and also examines the summer season.

Winter Temperature Analysis: **COAMPS 3PM**	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.86	-2.60	+0.85
<25% Cloud Cover	-1.49	-3.07	-0.03
<50% Cloud Cover	-1.33	-3.07	+0.17
COAMPS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.24	-2.07	+1.77
<25% Cloud Cover	+0.80	-1.19	+2.75
<50% Cloud Cover	+0.43	-1.66	+2.28
GFS 3PM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.73	-1.95	+0.52
<25% Cloud Cover	-1.16	-2.29	+0.05
<50% Cloud Cover	-1.05	-2.20	+0.16
GFS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.15	-1.67	+1.54
<25% Cloud Cover	+0.91	-0.72	+2.52
<50% Cloud Cover	+0.58	-1.14	+2.23

Table 3.1: COAMPS and GFS median overall biases across North America and 25th/75th percentiles for temperature biases in the winter (11/22 - 2/23).

3.1.1 Morning Low and Afternoon High Diurnal Variation

North America - Winter

The overall winter-time median biases for 7AM (COAMPS -0.24K, GFS -0.15K) in Table 3.1 are the result of distinct regional patterns of cold and warm biases nearly canceling out (Fig. 3.1A, B). For all cloud cover conditions, south of about 35° latitude, morning temperatures in both GFS and COAMPS tend to be too warm by 1K or more, while north of 40° latitude forecast temperatures are often too cold by > 1K. Sensitivity to cloud cover amount is illustrated by comparing Figure 3.1 for all cloud cover to Figure 3.2 for cloud cover < 25% station by station. A shift to more stations with larger warm biases for < 25%

Summer Temperature Analysis:**COAMPS 3PM**	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.49	-2.12	+1.19
<25% Cloud Cover	-1.03	-2.57	+0.37
<50% Cloud Cover	-0.74	-2.34	+0.93
COAMPS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.43	-1.68	+0.84
<25% Cloud Cover	-0.69	-1.92	+0.56
<50% Cloud Cover	-0.65	-1.85	+0.62
GFS 3PM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.5	-1.98	+0.89
<25% Cloud Cover	-0.52	-1.87	+0.57
<50% Cloud Cover	-0.51	-1.90	+0.65
GFS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.39	-1.48	+0.76
<25% Cloud Cover	-0.28	-1.37	+0.83
<50% Cloud Cover	-0.36	-1.48	+0.78

Table 3.2: COAMPS and GFS median overall biases across North America and 25th/75th percentiles for temperature biases in the summer (5/22 - 9/22).

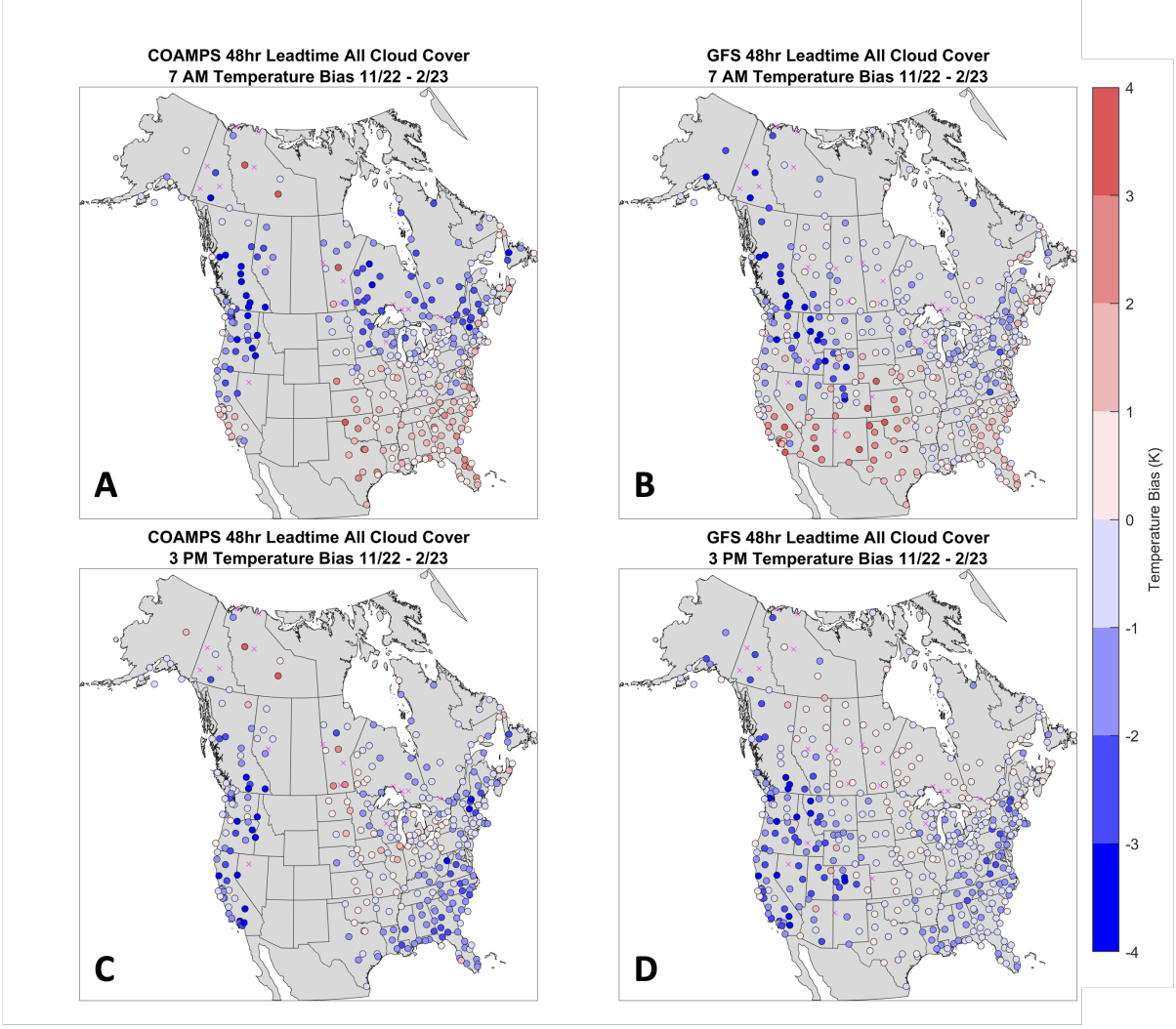

Figure 3.1: Diurnal temperature biases for COAMPS (A, C; left) and GFS (B, D; right) for November 2022 - February 2023 under all cloud conditions for morning (A, B; top panels) and afternoon (C, D; bottom panels). Stations marked with a pink 'X' denote stations with insufficient sample sizes (>30% missing data) to calculate a representative bias.

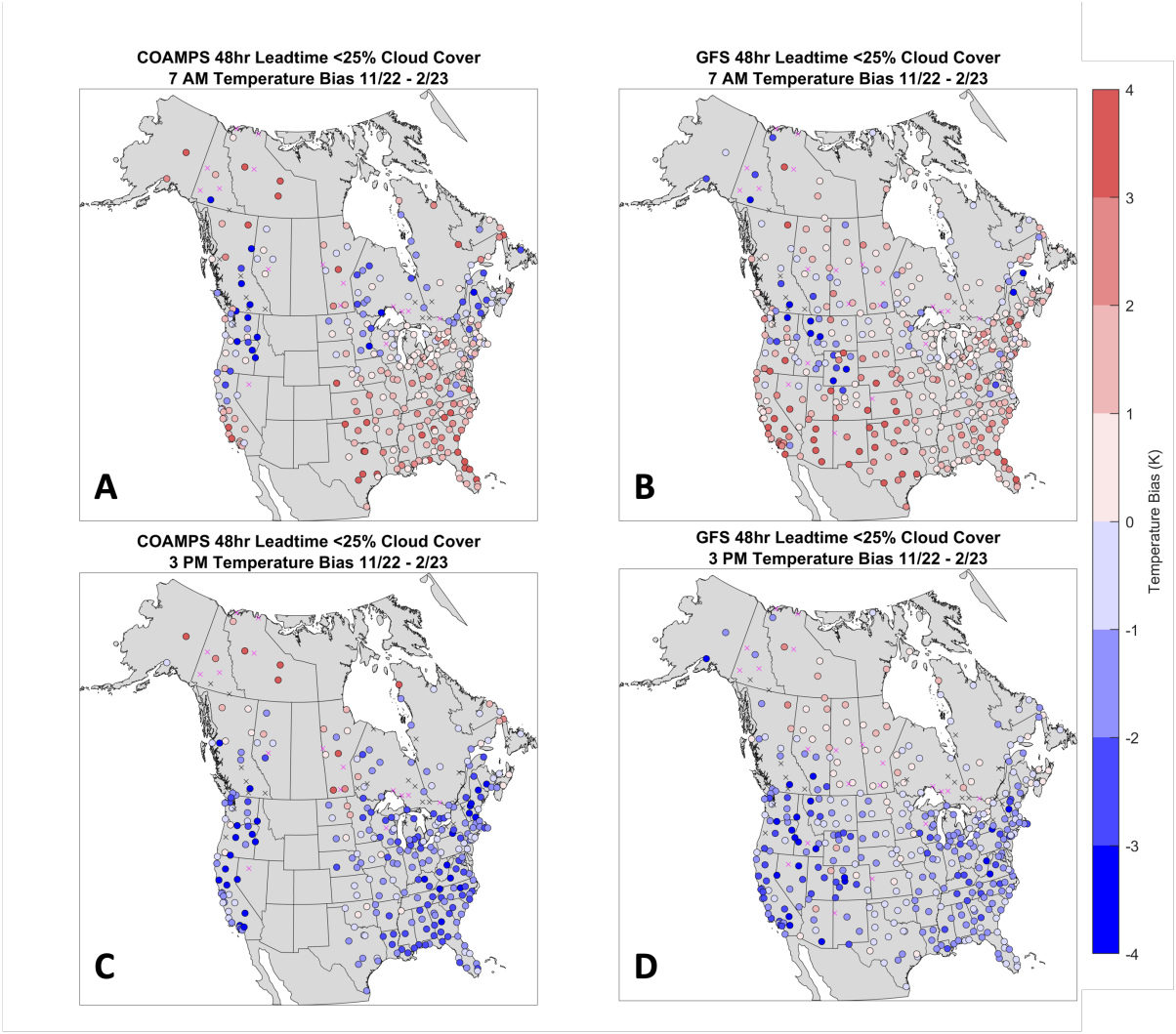

Figure 3.2: Diurnal temperature biases for COAMPS (left) and GFS (right) for November 2022 - February 2023 under <25% observed cloud cover (CLR, FEW). Stations marked with a pink 'X' denote stations with not enough non-missing observations (<30%) over the entire requested period to calculate a bias from. Stations marked with a black 'X' denote stations with enough observations over the entire period but that do not have enough observations when <25% cloud cover is present to calculate a reliable temperature bias from.

cloud cover is evident for 7 AM in winter. These station by station changes translate to a change in North America region median bias in winter mornings (Table 3.1) from a negative bias (cold) for all cloud cover conditions to a positive bias (warm) for < 25% cloud cover conditions.

All other conditions being equal compared to overcast conditions, when few clouds are present overnight (e.g. clear or scattered cloud cover), radiative cooling near the surface is larger, radiation inversions can form, and near surface overnight temperatures are lower. Morning warm biases that are larger in winter for < 25% cloud cover as compared to all cloud cover conditions may be associated with inadequate representation of overnight near surface radiation inversions in the models. Several factors in the boundary parameterization could contribute to whether a radiation inversion forms and its strength. If the model underestimates the amount of outgoing long wave radiation, cooling at the surface will be reduced which weakens the near-surface radiation inversion or prevents it from forming altogether. If low-level winds are too strong, there would be more mixing of warmer boundary layer above the surface with air near the surface. In conditions of calm winds, overnight lows are constrained by the dewpoint temperature as fog forms once the air temperature is reduced to the dewpoint. Morning warm biases in COAMPS may be influenced by some combination of these different factors within the boundary layer parameterization.

Winter afternoon (3PM) temperatures tend to be too cold in most areas of the US with the exception of stations in the Plains which have small errors (Table 3.1 and Fig. 3.1C, D). Examination of temperature biases for < 25% cloud cover shows the regional pattern of warm biases at 7AM extending further north, including stations in Alaska and portions of the northern tier of Canadian provinces (Fig. 3.2A, B). In the afternoon, station by station cold biases for cloud cover < 25% tend to be larger for most of CONUS (Fig. 3.2C, D). North America region median biases for winter afternoons do not change much among the cloud cover categories (Table 3.1) indicating compensating biases among the stations.

In both models, at 7AM and 3PM median temperature biases tend to be larger in mountainous terrain as compared to flatter areas including the Intermountain West, along the coastal range from California through the Pacific Northwest, and along the Appalachians in all cloud cover conditions (Fig. 3.1 B, D). COAMPS has both large cold and warm temperature biases in California which are discussed further in Chapter 4.

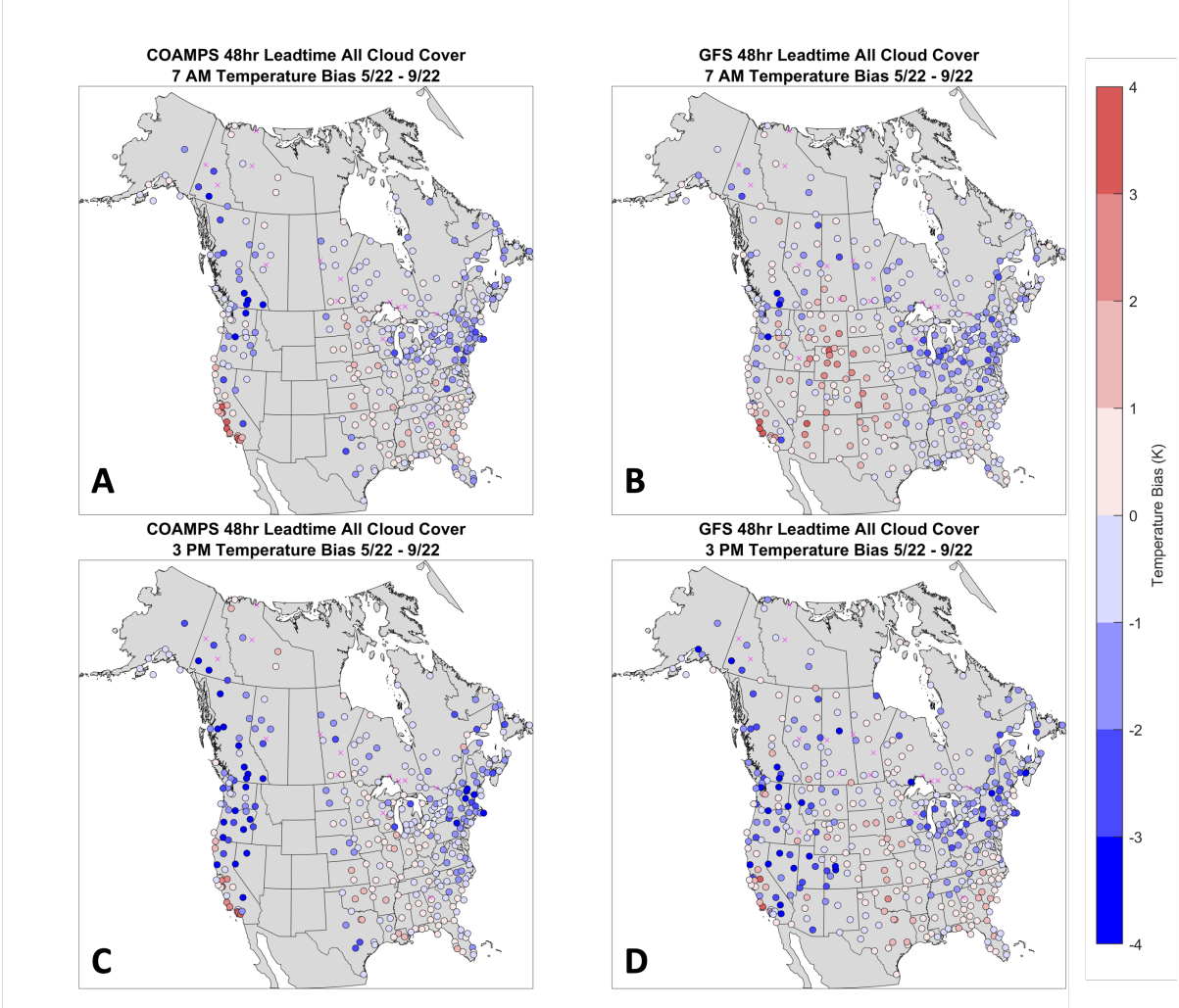

Figure 3.3: Same as Fig. 3.1 but for biases observed under all cloud cover conditions from May - September 2022.

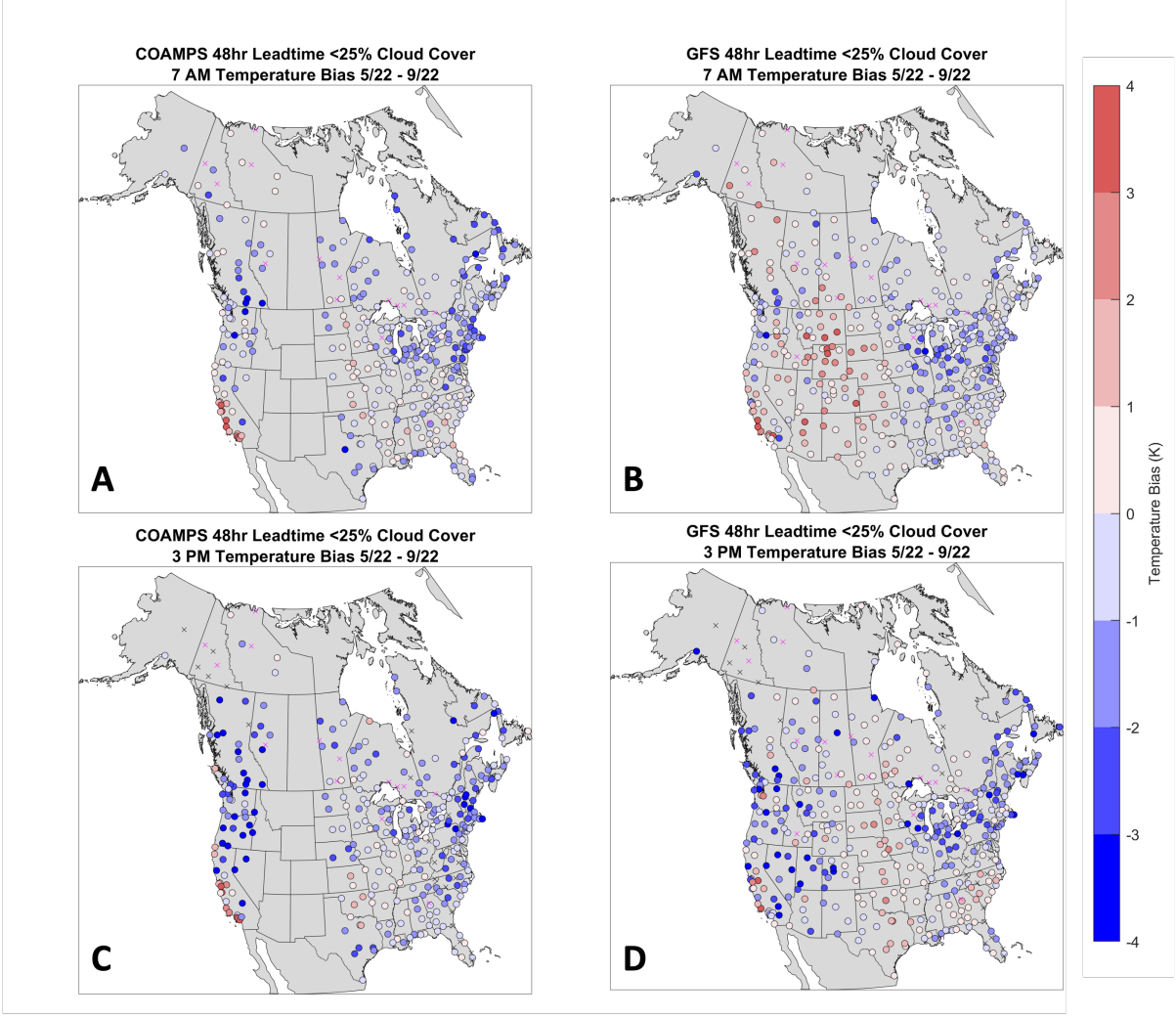

Figure 3.4: Same as Fig. 3.2 but for biases observed under <25% cloud cover from May - September 2022.

North America - Summer

Compared to winter, station by station temperature biases in summer are often smaller in magnitude and only weakly a function of cloud cover amount (Table 3.2 and Figs. 3.3 and 3.4). North America regional summer temperature biases at 7AM and 3PM are negative (cold bias) under all cloud cover conditions and <25% cloud cover with small differences in bias values among the cloud cover conditions with the exception of COAMPS 3PM (Table 3.2). Similar to the winter season, the geographic pattern of larger errors associated with mountainous terrain is present along the west coast and Intermountain region but is not as distinct along the Appalachians as it was in winter.

Sensitivity to differences in model grid and observation station elevations

Figure 3.5: Terrain elevation differences for COAMPS and GFS model grids and observation station elevation. A positive value indicates model terrain is too high. A negative value indicates model terrain is too low. A) COAMPS B) GFS.

COAMPS and GFS model grid elevation differences from observation station elevations have similar geographic patterns and are generally within +/- 200 m (Figure 3.5). Too high

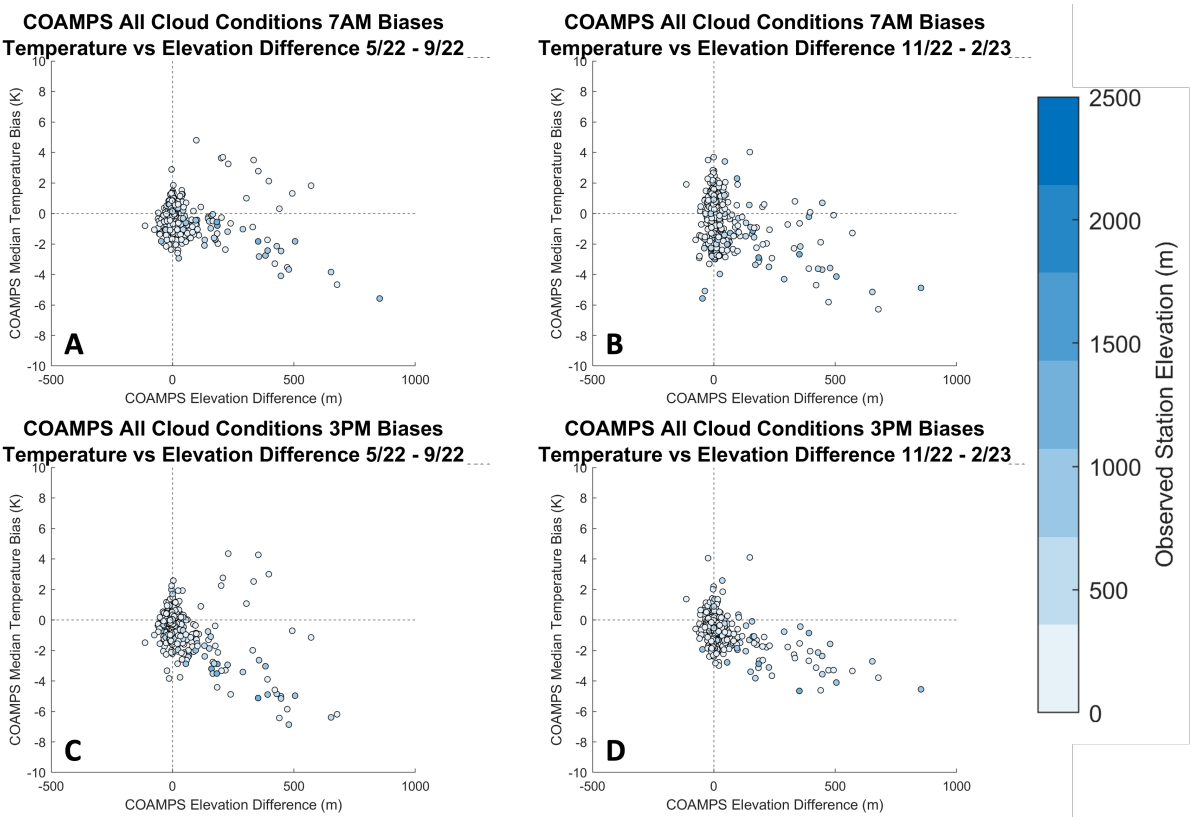

Figure 3.6: COAMPS 48-hour lead time temperature biases under all cloud conditions with COAMPS elevation differences from observations. Marker color represents the observation station elevation. A) 7AM 5/22 - 9/22 B) 7AM 11/22 - 2/23 C) 3PM 5/22 - 9/22 D) 3PM 11/22 - 2/23

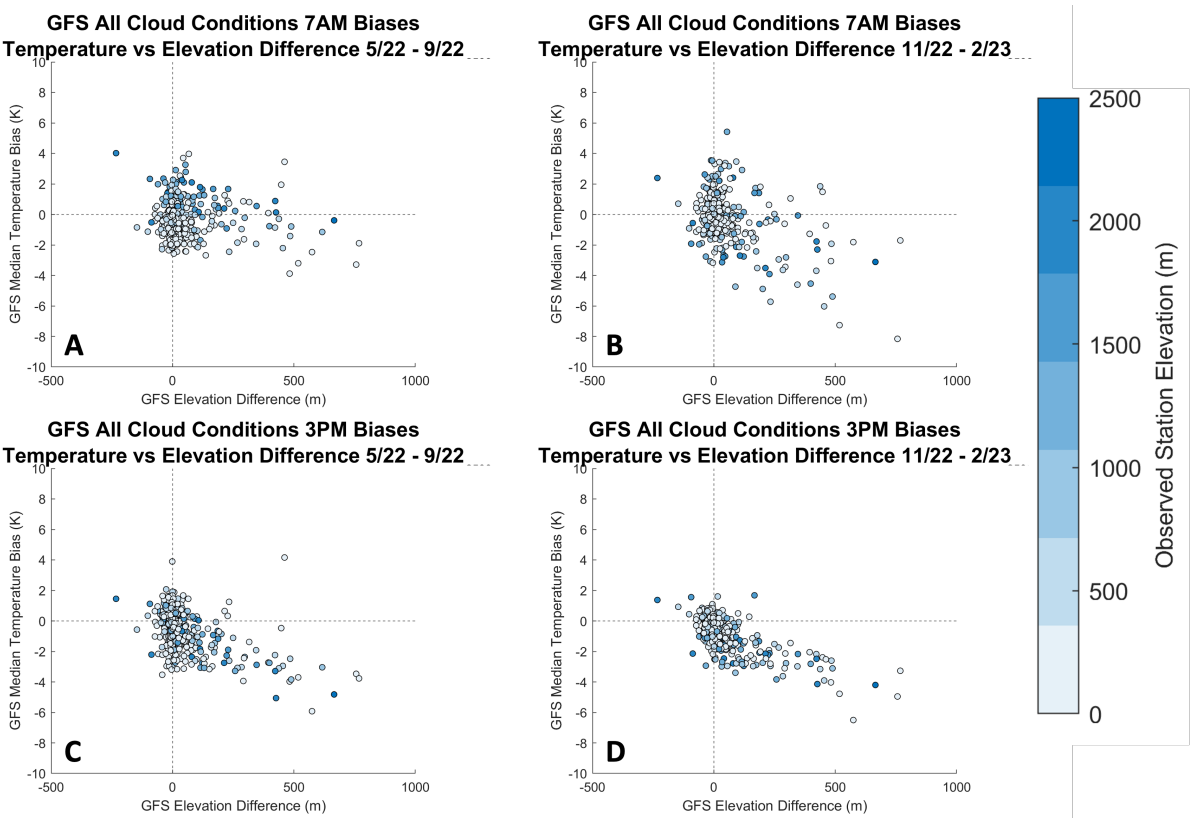

Figure 3.7: GFS 48-hour lead time temperature biases under all cloud conditions with GFS elevation differences from observations. Marker color represents the observation station elevation. A) 7AM 5/22 - 9/22 B) 7AM 11/22 - 2/23 C) 3PM 5/22 - 9/22 D) 3PM 11/22 - 2/23

heights tend to be in the complex terrain of western North America and slightly too low heights (< 100 m on average) occur sporadically in the central and eastern portions (Fig. 3.5). For both COAMPS and GFS, temperature biases are only loosely tied to model to observation elevation differences. Larger biases associated with elevation differences > 200 m occur both in higher terrain as indicated by darker colored dots and and lower terrain indicated by light colored dots in Figures 3.6 and 3.7. While differences between model grid and observation station elevations are a contributing factor for some stations, these differences alone do not account for temperature biases as there are many stations with very small elevation differences and several degree positive and negative temperature biases.

Model diurnal cycle implications

Comparison of temperature biases under all cloud cover at 7AM and 3PM (Fig. 3.8) for summer and for winter does not show clear evidence that the diurnal cycle is either too weak or too strong in the model overall. Many stations are too cool (bottom left quadrant) at both 7AM and 3PM in summer. In winter, the joint distributions of 7AM and 3 PM biases shows that the cold biases at 3 PM more evenly spread between cold and warm biases at 7 AM.

3.1.2 Bulk Seasonal Distributions of Temperatures

To remove forecast timing issues associated with computing biases at the same valid time between model and observations, we used a bulk analysis method (Sec. 2.4) to look at the seasonal distributions of observed temperatures and forecast COAMPS and GFS temperatures. This excludes data from stations in the Intermountain West where there are only GFS forecasts. We analyze this data for a 48-hour leadtime and a 72-hour leadtime to assess similarities and differences in the bulk analysis between these leadtimes.

Examination of the 48-hour and 72-hour leadtime temperature seasonal temperature distributions in Fig. 3.9 shows reasonably close correspondence between the observed and forecast temperatures except for slight offset (few deg C) to the left (too cold) for GFS during summer (Fig. 3.9 A, C)

In general, one would expect temperature forecasts to have larger errors at longer leadtimes. In terms of seasonal temperature distributions shown in Figure 3.9, differences between the 48-hour leadtime and 72-hour leadtime bulk analyses are minimal. Given that both COAMPS and GFS follow the overall observed temperature distributions fairly closely

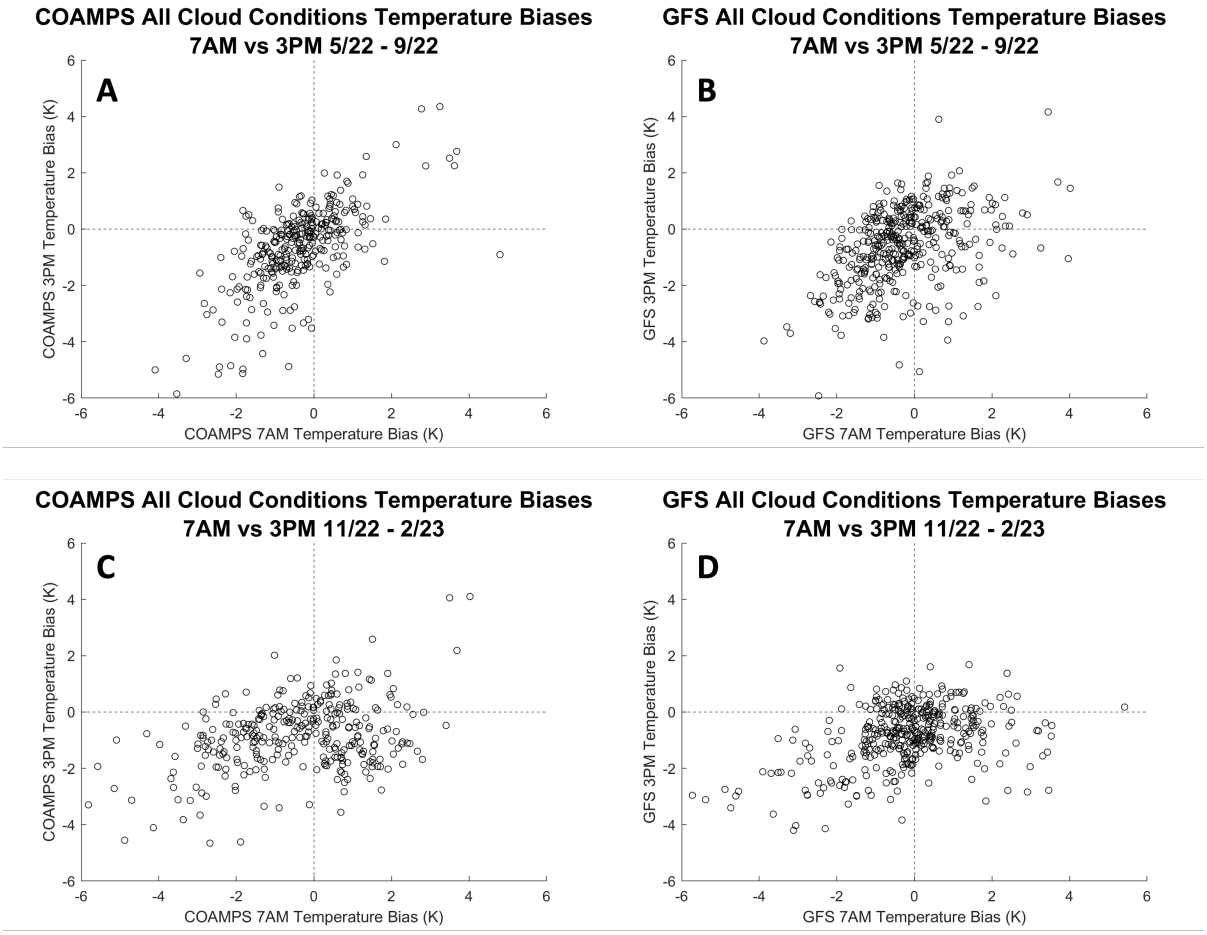

Figure 3.8: Scatter plots of temperature biases under all cloud cover conditions at 7AM (x-axis) and 3PM (y-axis) set against each other. A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

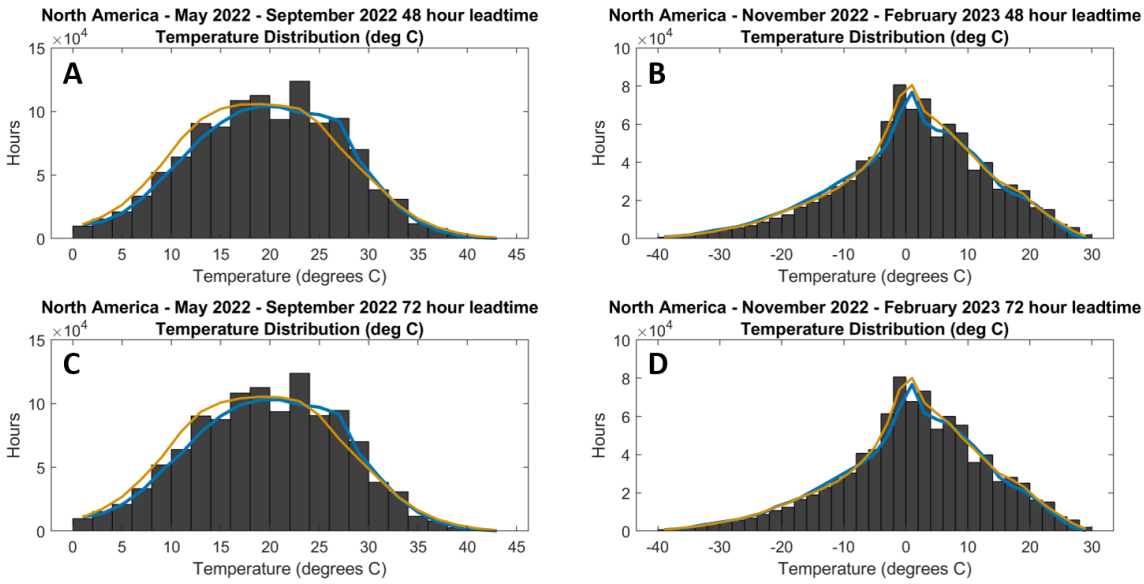

Figure 3.9: Bulk temperature distributions from May 2022-September 2022 (A, C; left panels) and November 2022-February 2023 (B, D; right panels) at 48-hour (A, B; top panels) and 72-hour (C, D; bottoms panels) leadtimes Black histogram represented observed temperature distribution, blue line represents COAMPS temperature forecast distribution, and orange line represents GFS temperature distribution for North America (all stations in COAMPS NEPAC and NWATL domains).

when matched timing is removed, event timing is likely a large contributor to some of the temperature biases in the matched model-observation bias analysis (Section 3.1.1).

3.1.3 Hourly events outside the 10th and 90th hourly temperature climatological percentiles

Observed Extreme Event Bias	Median Observed >90th Percentile Warm Event Bias (K)	Median Observed <10th Percentile Cold Event Bias (K)
COAMPS (summer)	-1.59	+2.28
GFS (summer)	-1.58	+0.28
COAMPS (winter)	-1.97	+1.16
GFS (winter)	-1.60	+0.83

Table 3.3: North America region median temperature biases for observed >90th percentile warm and <10th percentile cold event biases in COAMPS and GFS for summer and in winter.

We examined temperature biases for both COAMPS and GFS when the observed hourly temperature was greater than the 90th percentile and less than the 10th percentile. This allows us to assess how well numerical weather prediction models are forecasting hourly temperature events that fall within the tails of the climatological distribution. In the winter months of 11/22-2/23, warm spells are more common in the eastern region and cold spells more common in the west and central regions (right column of Fig. 3.10). In contrast in summer months of 5/22- 9/22, the geographic pattern is more diffuse with warm spells slightly more common in the west, central region, and Florida compared to other regions and cold spells widely distributed (left column of Fig. 3.10). Both models tend to underestimate the intensity of events outside of the 10th and 90th percentiles, with forecast temperatures too cool for warm events (observed temperature >90th percentile) and too warm for cold events (observed temperature <10th percentile) (Table 3.3).

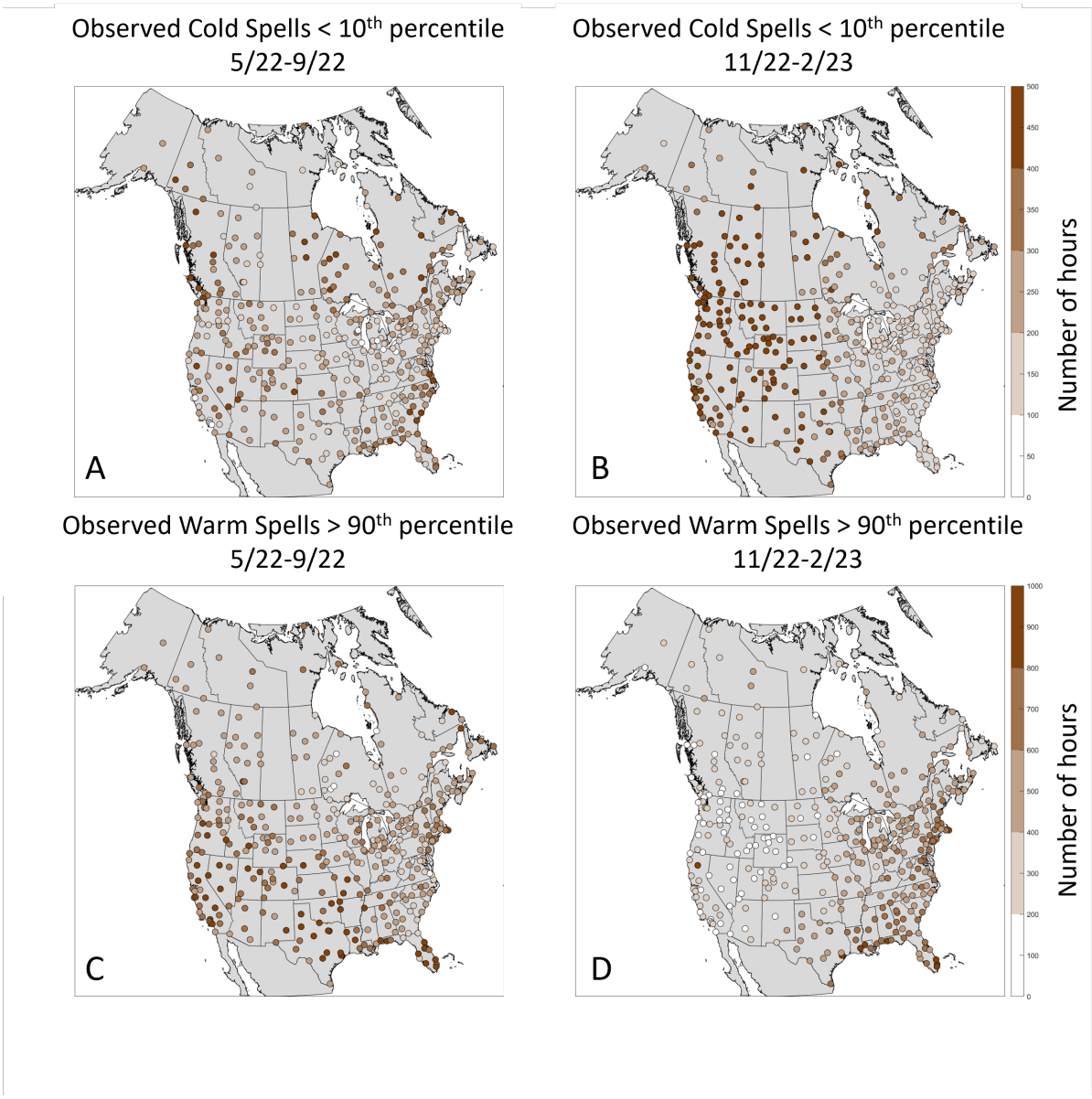

Figure 3.10: Number of observed hours corresponding to the tails of the temperature climatology. A) < 10th percentile for 5/22 - 9/22,B) < 10th percentile for 11/22-2/23, C) > 90th percentile for 5/22 - 9/22 and D) > 90th percentile for 11/22 - 2/ 23. Note that maximum value in the color scale for cold spells (top row) is 500 hours as compared to 1000 hours for warm spells (bottom row).

North America - Observed Temperature < 10th Percentile Cold Events

The geographic patterns of forecast temperature biases during observed <10th percentile cold events are more distinct for COAMPS than GFS (Fig. 3.11). For the eastern portion of CONUS, COAMPS is consistently several degrees too warm for cold events with larger errors in summer than winter. In comparison, GFS has a more mixed spatial pattern of weak, warm, and cold biases north of about 36 ° latitude with a tendency to forecast too warm south of that latitude. Both COAMPS and GFS have high spatial variability of temperature biases along the west coast and for GFS in the Intermountain West.

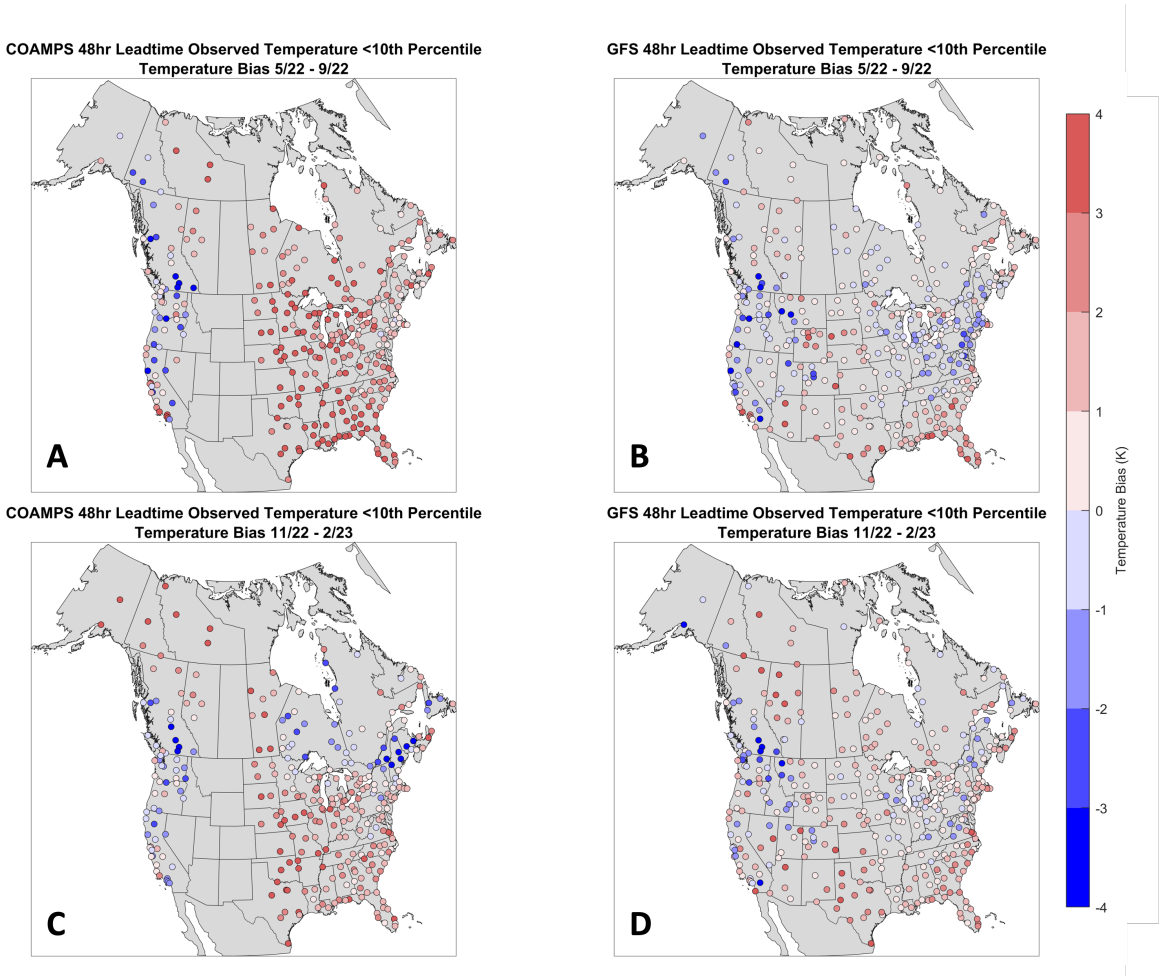

Figure 3.11: Temperature bias for observed cold events (observed temperatures < 10th percentile temperatures based on climatology data) during the summer (top row) and the winter (bottom row). A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

Figure 3.12: Number of hours where temperatures < 10th percentile were forecast by either COAMPS or the GFS but observed temperatures were not below the 10th percentile (false alarms). A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

The number of hours of cold event "false alarms" when a model forecasted temperatures < 10% of long term climatology did not verify in the observations has distinct geographic patterns by season (Fig. 3.12). Both COAMPS and GFS have very few cold event false alarms (< 100 hours per station out of 2,880 total hours) in winter for the eastern portion of the US and Canada. A few hundred hours of cold event false alarms occur for southern Plains, Intermountain West, and west coast. The geographic patterns of "missed" cold events in winter are similar to that for false alarms (Fig. 3.13) with a low number of hours for the eastern portion of the US and Canada and higher numbers of hours elsewhere. In summer,

Figure 3.13: Number of hours where temperatures < 10th percentile were not forecast (missed events) by either COAMPS or the GFS but observed temperatures were below the 10th percentile. A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

overall GFS has few missed cold events (typically < 100 per station out of 3,672 total hours) across North America whereas COAMPS tends to miss cold events in the eastern portion of the US (typically between 100-300 missed hours per station out of 3,672 total hours).

North America - Observed Temperature > 90th Percentile Warm Events

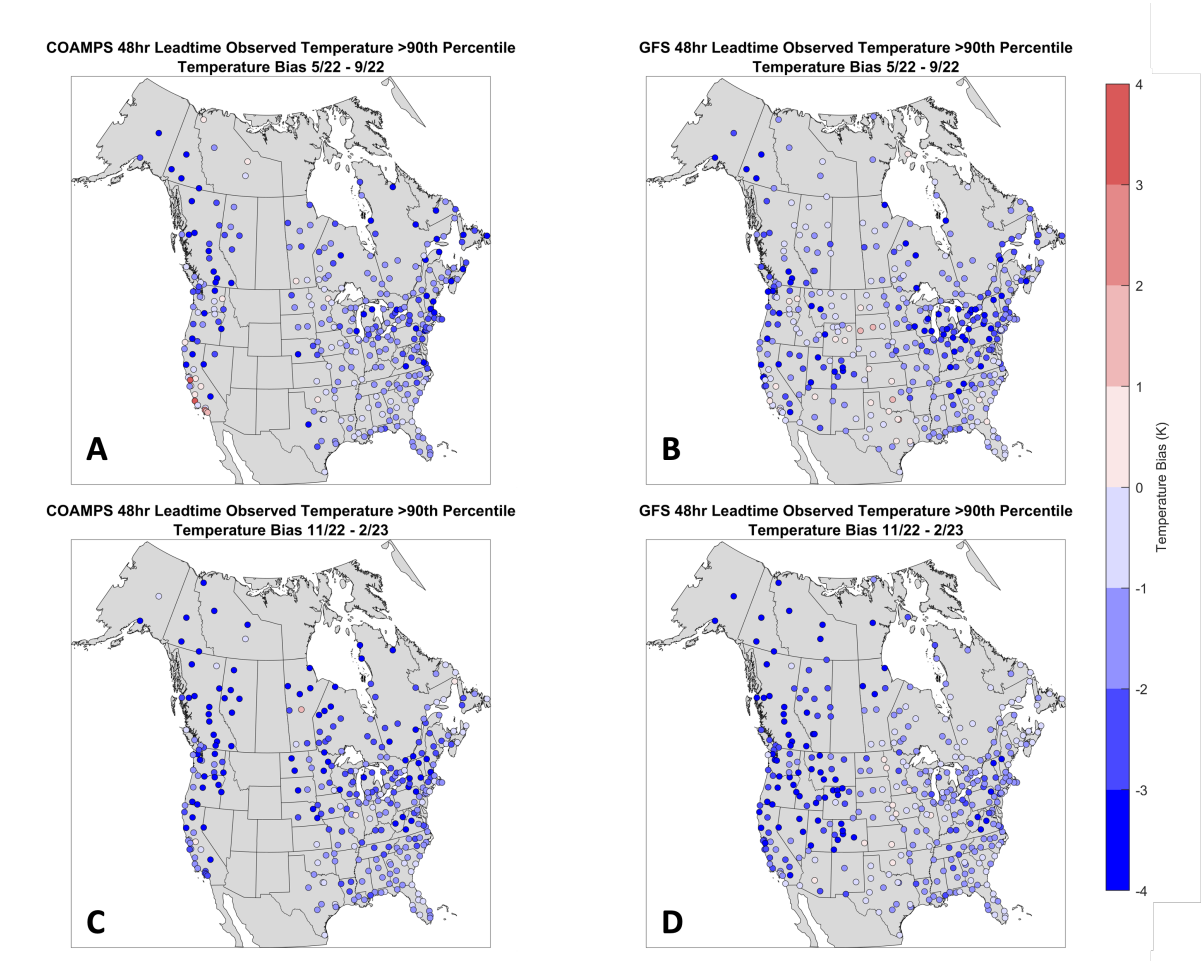

Figure 3.14: Temperature bias for observed warm events (observed temperatures > 90th percentile temperatures based on climatology data) during the summer (top row) and the winter (bottom row). A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

The geographic temperature bias pattern for observed warm events is similar in both the summer and the winter for both COAMPS and the GFS. Both models consistently

Figure 3.15: Number of hours where temperatures > 90th percentile were forecast by either COAMPS or the GFS but observed temperatures were not above the 90th percentile (false alarms). A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

Figure 3.16: Number of hours where temperatures > 90th percentile were not forecasted by either COAMPS or the GFS but observed temperatures were above the 90th percentile (missed events). A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

underestimate the strength of warm events (i.e. have cold biases) throughout North America with limited exceptions (Southern California for COAMPS and the Plains for GFS) (Fig. 2.7). Some of the larger cold biases in summer are in the northern tier of US states where air conditioning is less common than other areas of the US.

False alarms for warm event hours are more common in summer than winter in both models with stations averaging between 200-500 false alarms in the summer and 100-300 false alarms in the winter (Fig. 3.16). Geographic patterns of the number of false alarm hours are mixed overall except for a tendency for more frequent errors in many coastal locations along the Florida and Gulf coasts in summer. Missed forecasts for observed warm event hours are also more common in summer than in winter along a similar pattern as the number of false alarms at each station (Fig. 3.16).

Based on examination of many time series at numerous stations, the model's maximum daily temperature usually coincides with the observed maximum and the model's minimum daily temperature usually coincides with the observed minimum. Hence, the underforecasts of tail warm events and cold events are unlikely to be timing related. The types of metrics such as regional monthly mean biases and root mean square errors at a given lead time that are often used by modeling centers such as NCEP and NRL to evaluate model improvements between versions have a tendency to prioritize doing well for average conditions rather than tail conditions. This is in effect a model design philosophy issue and is beyond the scope of this study.

3.2 Dewpoint

Knowing whether or not the model is biased towards a moister or drier atmosphere can help to diagnose underlying issues that may be influencing model forecast temperatures and precipitation event timing. For example, small changes in dewpoint can yield large changes in Convective Available Potential Energy (CAPE) which impacts a model's ability to forecast severe weather. Similar to the model minus observation calculation of biases we used for temperature, if dewpoint value is forecast to be higher than the observed dewpoint temperature, the model indicates that the air is moister than it actually was. As in the discussion of temperature biases by cloud cover amount in Section 3.1, we focus discussion on all cloudiness conditions and conditions with $< 25\%$ cloud cover.

In winter and in summer, COAMPS has a regionally varying geographic pattern of dewpoint biases that tend to be too high in eastern North America and too low in western

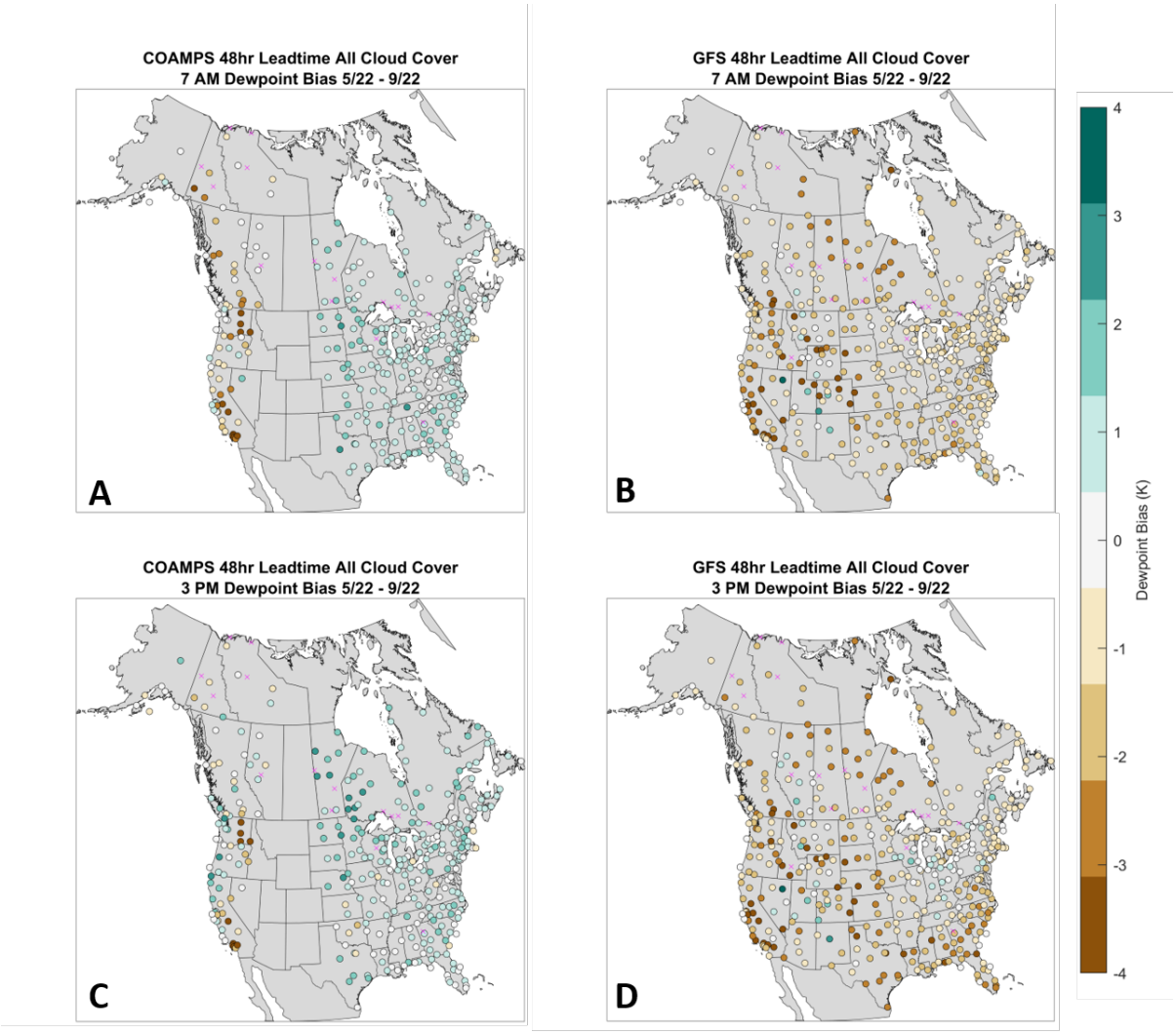

Figure 3.17: Diurnal dewpoint biases for COAMPS (left) and GFS (right) for May 2022 - September 2022 under all cloud conditions. Stations marked with a pink 'X' denote stations with insufficient sample sizes (>30% missing data) to calculate a representative bias. A) COAMPS morning dewpoint biases B) GFS morning dewpoint biases C) COAMPS afternoon dewpoint biases D) GFS afternoon dewpoint biases.

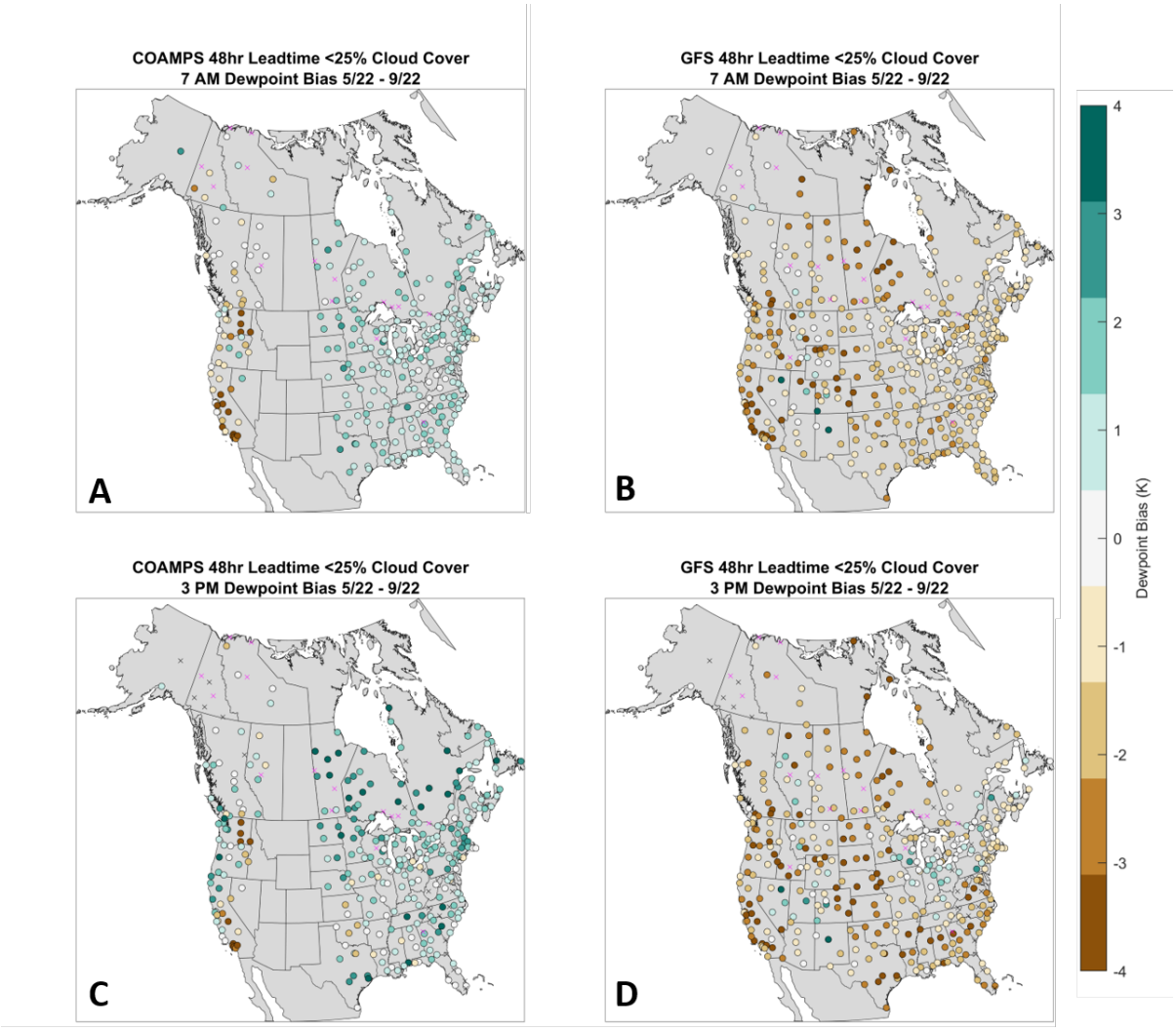

Figure 3.18: Diurnal dewpoint biases for COAMPS (left) and GFS (right) for May 2022 - September 2022 under <25% observed cloud cover (CLR, FEW). Stations marked with a pink 'X' denote stations with insufficient sample sizes (>30% missing data) to calculate a representative bias. Stations marked with a black 'X' denote stations with enough observations over the entire period but that do not have enough observations when <25% cloud cover is present to calculate a reliable dewpoint bias from. A) COAMPS morning dewpoint biases B) GFS morning dewpoint biases C) COAMPS afternoon dewpoint biases D) GFS afternoon dewpoint biases.

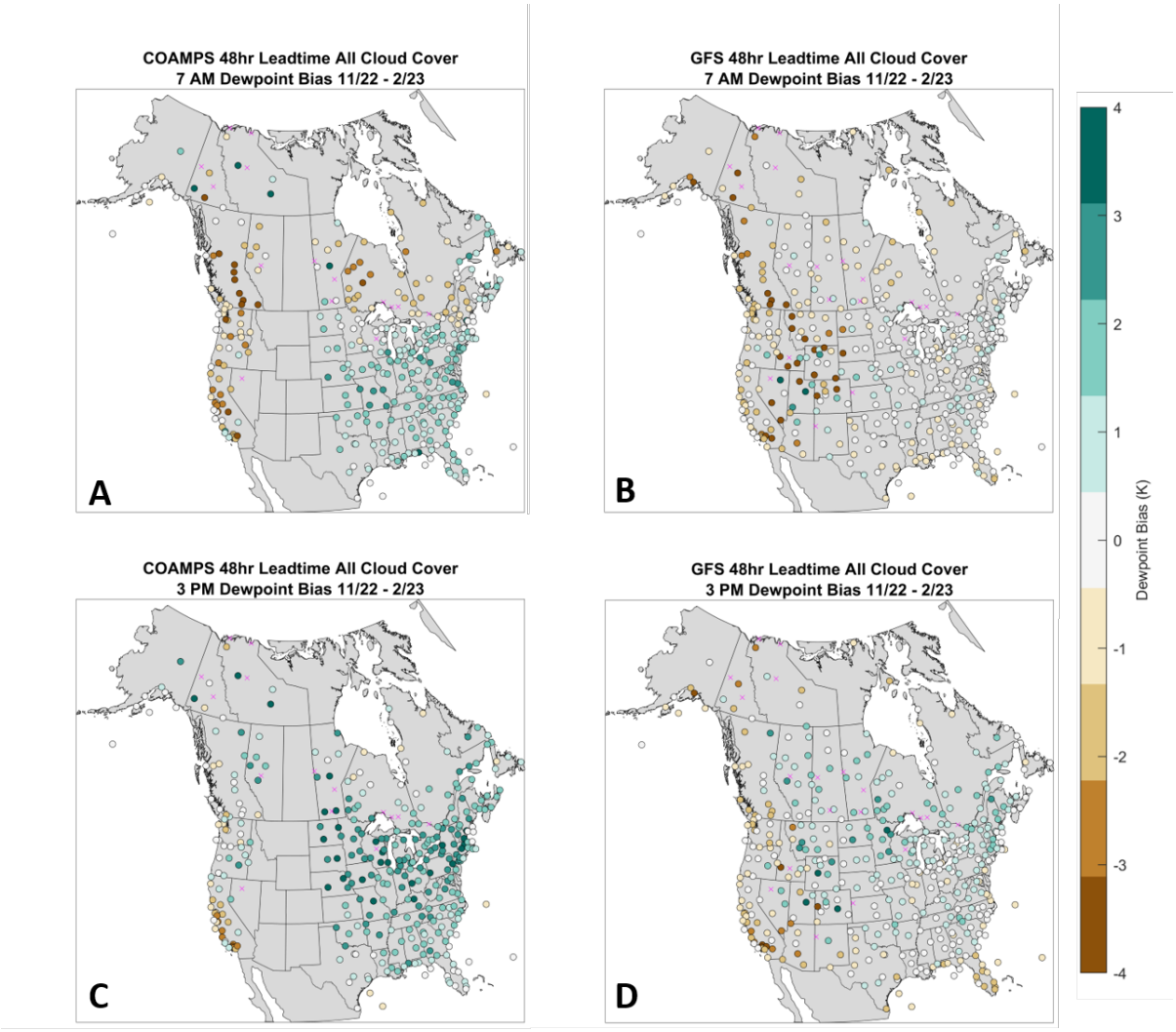

Figure 3.19: Same as Fig. 3.17 but for dewpoint biases under all cloud conditions from November 2022 - February 2023.

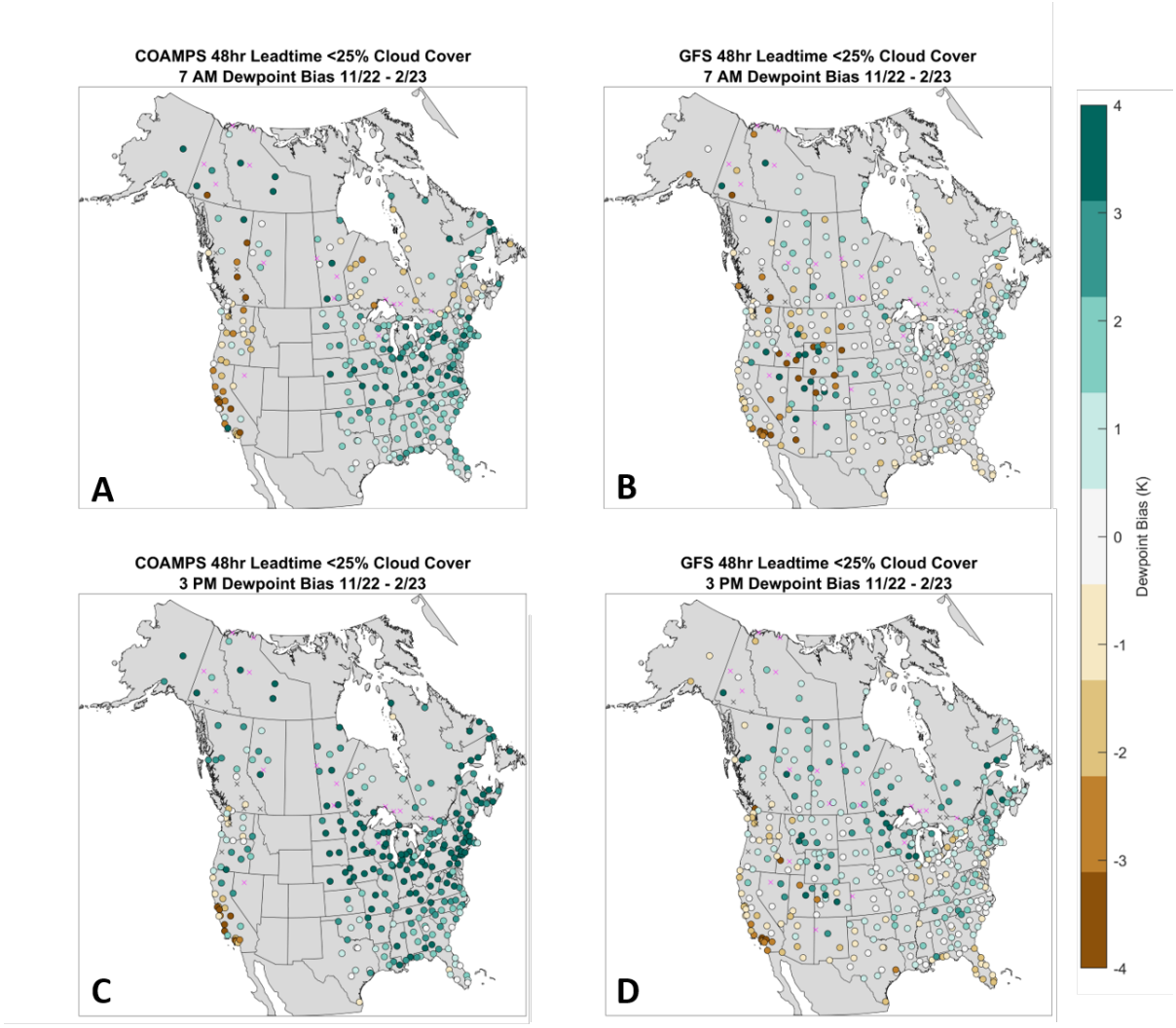

Figure 3.20: Same as Fig. 3.18 but for dewpoint biases under <25% observed cloud cover from November 2022 - February 2023.

North America (Fig. 3.19 and 3.17). Winter season dewpoint biases in COAMPS are often slightly higher than those in summer. There is a similar geographic pattern for < 25% cloud cover with station by station dewpoint biases often indicating higher magnitudes compared to all cloud cover (Fig. 3.20 and 3.18). In summer for COAMPS, the strongest dry dewpoint biases tend to be located within the elevated terrain of the Pacific Northwest and central California.

In contrast, for GFS in both the winter and the summer, dewpoints in the eastern US have very small median biases at 7 AM and slightly higher values at 3 PM (Fig. 3.19). In the west, the picture is mixed with stations with small errors, too moist and too dry in close proximity. For < 25% cloud cover (Fig. 3.18), dewpoint median biases in the eastern US become more variable compared to for all cloud cover.

Dew point values are jointly influenced by soil moisture (an initialization field), moisture advection, and the parameterization of evapotranspiration which is more important in the summer growing season than the winter dormant season. The mix of moisture sources may influence dewpoints differently in the eastern and Western United States. For example, too moist dewpoints in the East may be associated with overestimation of moisture advection coming from the warmer Atlantic Ocean. With only the output of the operational models, it is not possible to separate the relative roles of the different moisture sources in contributing to the overall biases. Sensitivity tests would be needed to assess the varied regional and seasonal contributions.

3.3 Winds

We examined the percent of time the model forecast wind speeds and model forecast wind directions meet or exceeded TAF amendment criteria (Section 2.6) for COAMPS and GFS during the summer (5/22 - 9/22) and the winter (11/22 - 2/23) over all times of day (not focusing on the diurnal cycle). As discussed in section 2.6, a forecasted wind speed value meets TAF amendment criteria if the difference between the forecast and observed wind speeds is greater than or equal to 10 knots. A forecasted wind direction value meets TAF amendment criteria if the difference between the forecast and observed wind directions is greater than 30 degrees when winds greater than 15 knots are forecast to occur (Department of the Air Force 2020). Meeting or exceeding TAF amendment criteria represents an error in winds that would impact aviation and requires the TAF to be amended from its original forecast. At most stations, wind speeds and wind directions were well forecast and exceeded

TAF Amendment criteria less than 2% of the time. For COAMPS (Fig. A.3), in summer, 233 out of 321 stations (73%) exceeded TAF amendment criteria for both wind speed and wind direction <2% of the time while in winter, 182 out of 321 stations (57%) exceeded TAF amendment criteria <2% of the time. For GFS (Fig. A.4), in summer, 268 out of 403 stations, (67%) exceeded TAF amendment criteria for both wind speed and wind direction <2% of the time and in winter, 234 out of 397 stations (59%) meet TAF amendment criteria <2% of time. We consider stations that exceeded TAF amendment criteria <2% of the time to not have substantial wind errors as aviation activities would be infrequently impacted by wind speed/direction errors at these stations. Stations that meet TAF amendment criteria >2% of the time are considered to have more notable errors and require further analysis to assess what may be contributing to the higher frequency of occurrence with these errors.

3.3.1 Geographic Patterns of Wind Speed and Wind Direction Errors

The predominant geographic regions with higher frequency wind speed and wind direction errors in both models are regions with elevated terrain along the West Coast and the Intermountain West, and coastal regions in northeast Canada (Fig. 3.21 and 3.22). In comparison to wind speed, both GFS and COAMPS had fewer wind direction errors exceeding the TAF amendment criteria. Compared to GFS, COAMPS has a higher frequency of wind speed errors (locations that just exceed the 2% threshold) in the Great Plains states (Texas, Oklahoma, Kansas, Nebraska, North Dakota, and South Dakota). Model elevation differences between the COAMPS grids and the weather stations shows that stations in the South and Plains states are usually within +/- 100 meters indicating the model has minimal issues representing the terrain in these areas (Fig. 3.5 A). Regionally consistent wind speed/direction errors in a particular region may be caused in part by issues related to surface roughness (an initialization field). For example, if roughness is over smoothed it can lead to an overestimate of wind speeds.

Wind speed and direction errors along the coasts may be related to representation of the sea breeze or issues with resolving mountainous terrain along the coastline along the west coast. Wind errors in mountainous terrain are likely related to smoothing of terrain within the model compared to actual terrain (discussed further in Section 4.2). Another potential error source comes from the track associated with cyclones as they move through a region. If a cyclone is forecast to go north of a station but it actually goes south this could lead to a large number of wind direction errors in a given forecast which will result in TAF amendment criteria being met more frequently.

Figure 3.21: Percent of time that wind speed meets TAF amendment criteria for COAMPS (A, C; left) and GFS (B, D; right) from May 2022 - September 2022 (A, B; top row) and November 2022 - February 2023 (C, D; bottom row). The subset of stations plotted meet TAF amendment criteria >2% of the time which is considered notable.

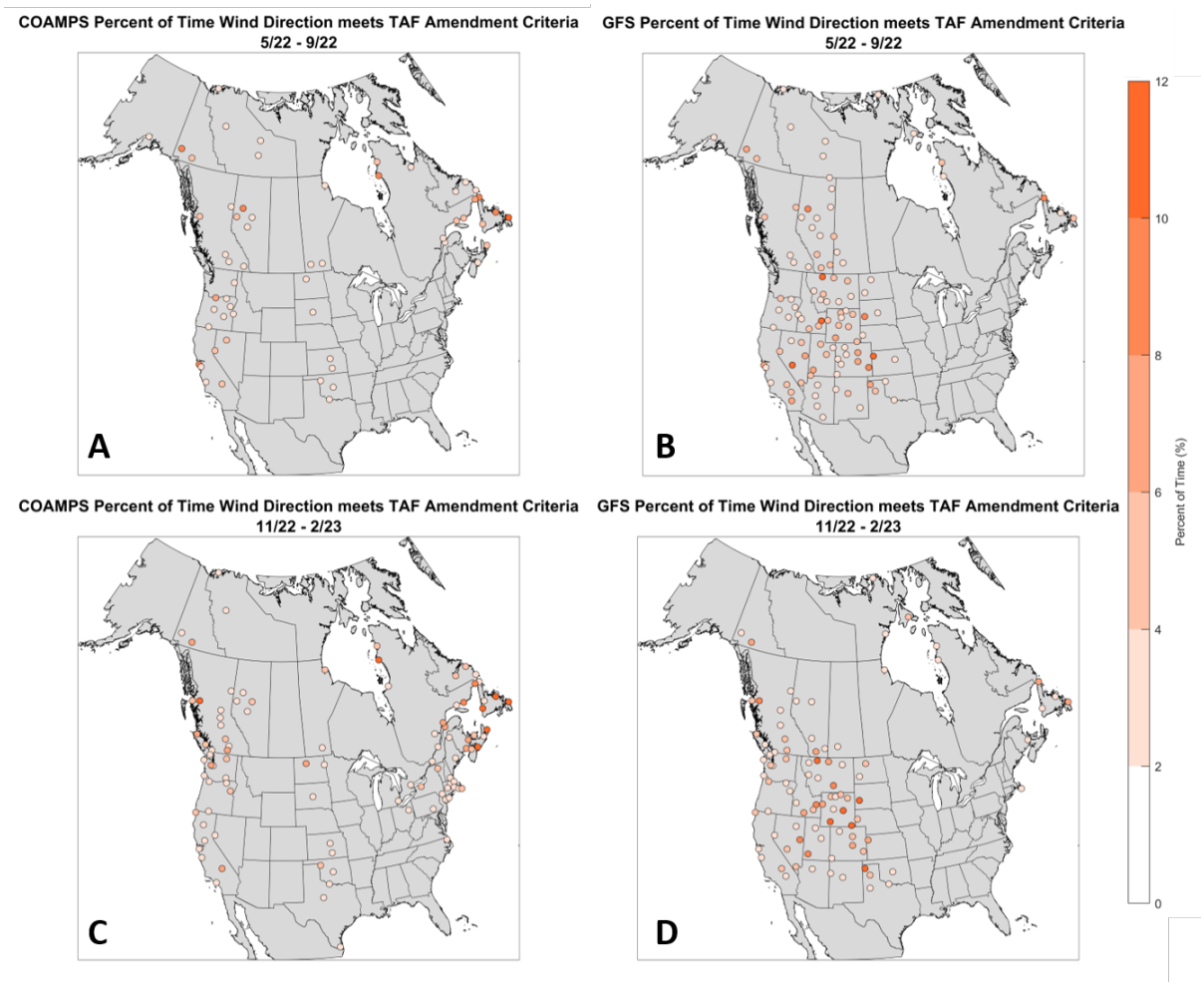

Figure 3.22: Same as Fig. 3.21 but for percent of time that wind direction meets TAF amendment criteria.

3.4 Relationships Between Temperature Biases and Dewpoint/Wind Direction Biases

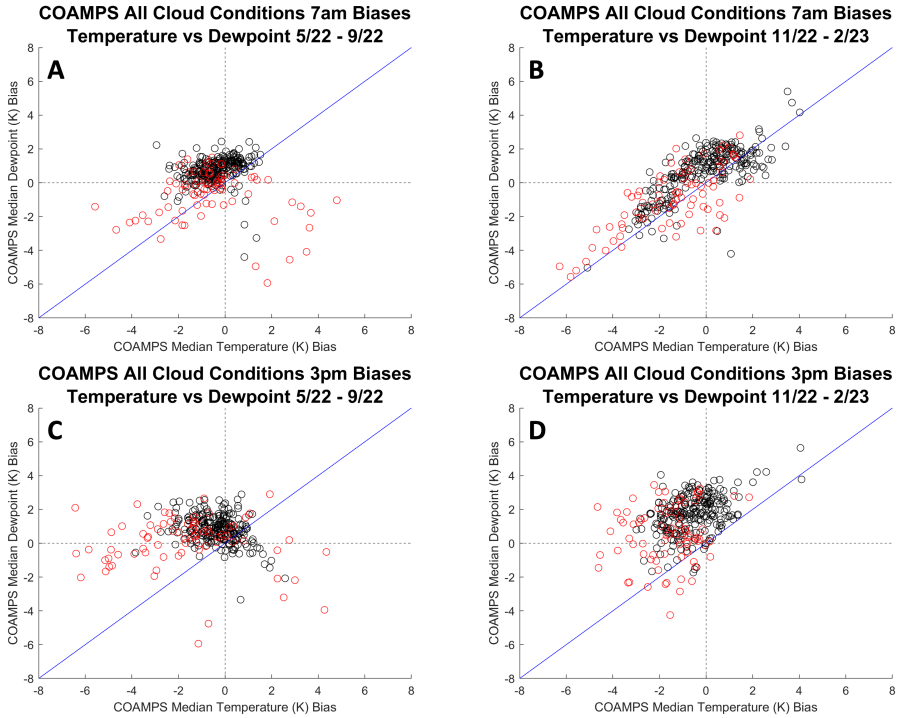

Figure 3.23: Scatter plot of COAMPS temperature biases (x-axis) and dewpoint biases (y-axis) set against a one-to-one line at 7AM/3PM in the summer and winter. Stations marked in red are mountain stations. Stations marked in black are non-mountain stations. A) 5/22 - 9/22 7AM B) 11/22 - 2/23 7AM C) 5/22 - 9/22 3PM D) 11/22 - 2/23 3PM.

Examination of scatter plots of joint temperature and dewpoint biases in COAMPS indicates little to no correlation except for a weak relationship in winter at 7AM which indicates increasing dewpoint bias magnitude with increasing temperature bias magnitude (Fig. 3.23). For GFS, there is little to no correlation between temperature and dewpoint biases at any time or season (Fig. 3.24). The subset of non-mountain stations (black dots in Figures 3.23 and 3.24) clump together compared to the mountain stations (red dots) emphasizing the lack of correlation even in flatter terrain conditions. Comparison of COAMPS and GFS temperature biases versus percent of time wind direction met TAF amendment criteria showed no clear relationship between the frequency of the wind direction errors and

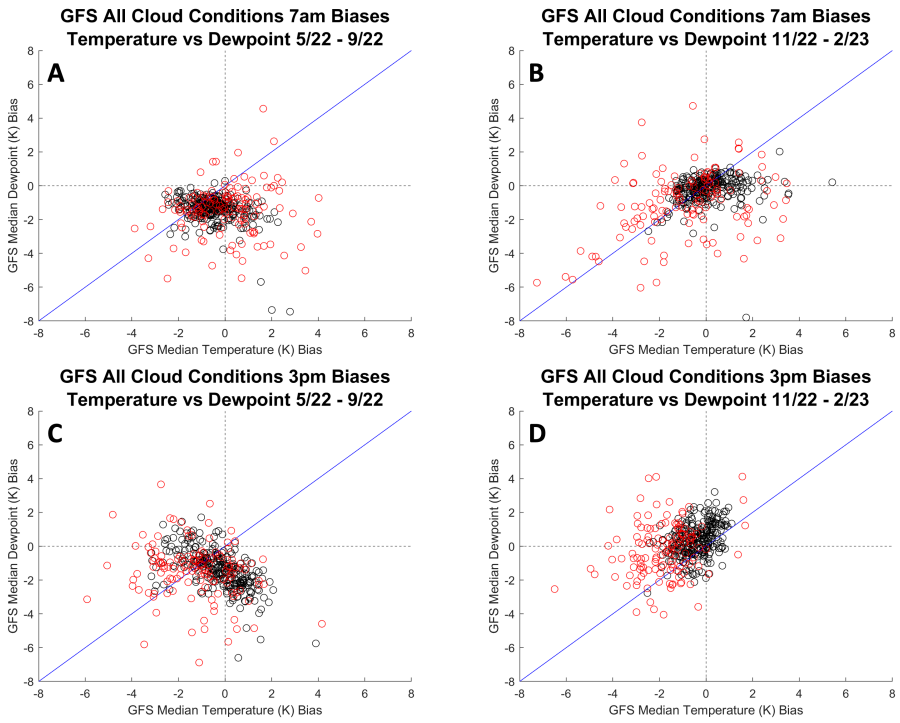

Figure 3.24: Same as Fig. 3.23 but for GFS. A) 5/22 - 9/22 7AM B) 11/22 - 2/23 7AM C) 5/22 - 9/22 3PM D) 11/22 - 2/23 3PM.

the magnitude of the temperature errors. The lack of strong correlations shown in these comparisons rules out simple relationships between temperature errors and dewpoint errors and between temperature errors and wind direction errors. Examination of weather-conditioned subsets of biases beyond observed cloud cover may aid in determining further bias relationships but such analysis is beyond the scope of this study.

3.5 Timing of Low Pressure Events

To assess model forecast skill for the timing of low pressure system passages, we examined the 9 hour pressure tendency to see how often COAMPS and GFS forecast low pressure system passages and whether they were consistently forecast to arrive too early or too late (Section 2.5.2).

Examination of the number of detected observed low pressure center passages shows an unexpected result (Fig. 3.27). There are more low pressure passages in the Southwest US in summer than any other location or season. Many stations in the Southwest summer

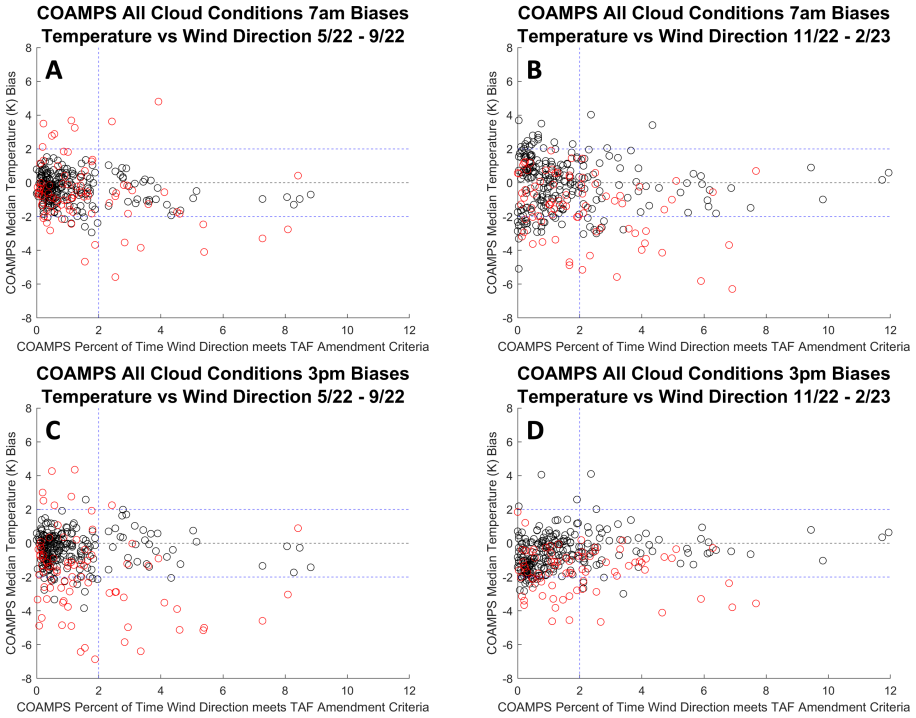

Figure 3.25: Scatter plot of COAMPS temperature biases (y-axis) and percent of time wind direction met TAF amendment criteria (x-axis) at 7AM/3PM in the summer and winter. Stations marked in red are mountain stations. Stations marked in black are non-mountain stations. Dashed horizontal lines are located at +/- 2 and 0 on the y-axis and dashed vertical line is located at 2% on the x-axis. A) 5/22 - 9/22 7AM B) 11/22 - 2/23 7AM C) 5/22 - 9/22 3PM D) 11/22 - 2/23 3PM.

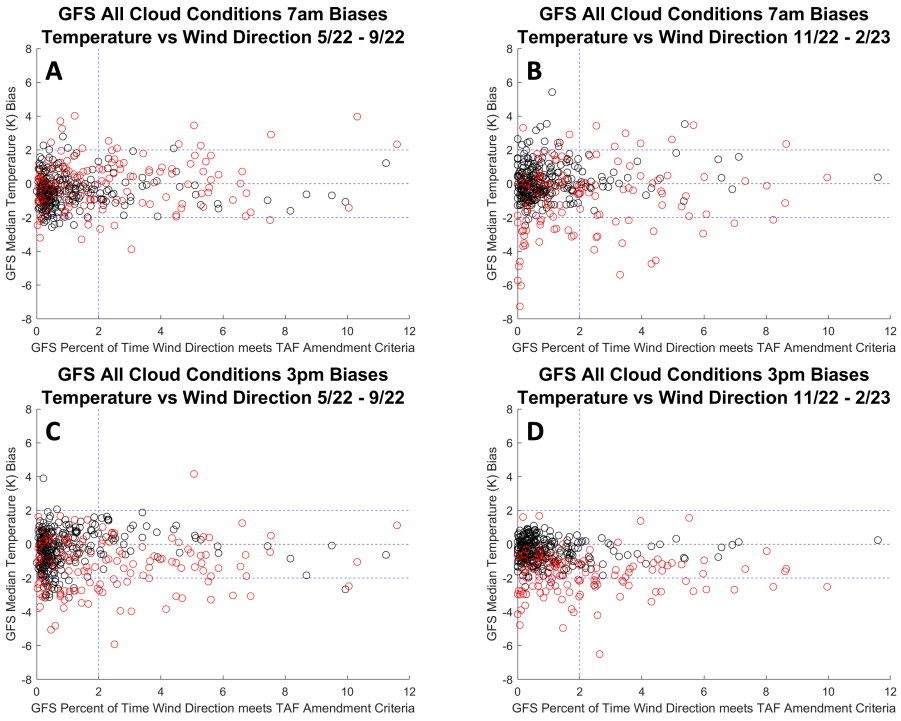

Figure 3.26: Same as Fig. 3.25 but for GFS. A) 5/22 - 9/22 7AM B) 11/22 - 2/23 7AM C) 5/22 - 9/22 3PM D) 11/22 - 2/23 3PM.

have 4+ low pressure events per week, whereas the number of winter low pressure events remains between 1-2 low pressure events per week. If a extratropical cyclone occurred every 6 days over the 4 month winter (119 days) it would yield 20 events so the winter season numbers of approximately 1-2 low pressure events per week are within realistic bounds. As expected, in winter the largest number of observed low pressure passages occur in the upper Midwest and into the Northeast. In the winter, stations with less than twelve observed low pressure passages are primarily located in Florida and California.

We are skeptical that the 4+ observed low pressure passages per week at stations across the US southwest in summer (Fig. 3.27) represent synoptic scale storms as there are few if any extratropical cyclones in the summer in this region. We suspect that this large number of low pressure events is associated with variations in thermal lows that are detected by the pressure tendency method. High summertime temperatures can yield a thermal low and may result in a diurnal pressure trace which can trigger our method of detecting low pressure passages (Fig. 3.28). Thermal lows are defined as a stationary closed off region of low pressure that is not associated with any type of cold front or warm front passage (Rowson and Colucci 1992). The region with the highest number of days with a thermal low recorded

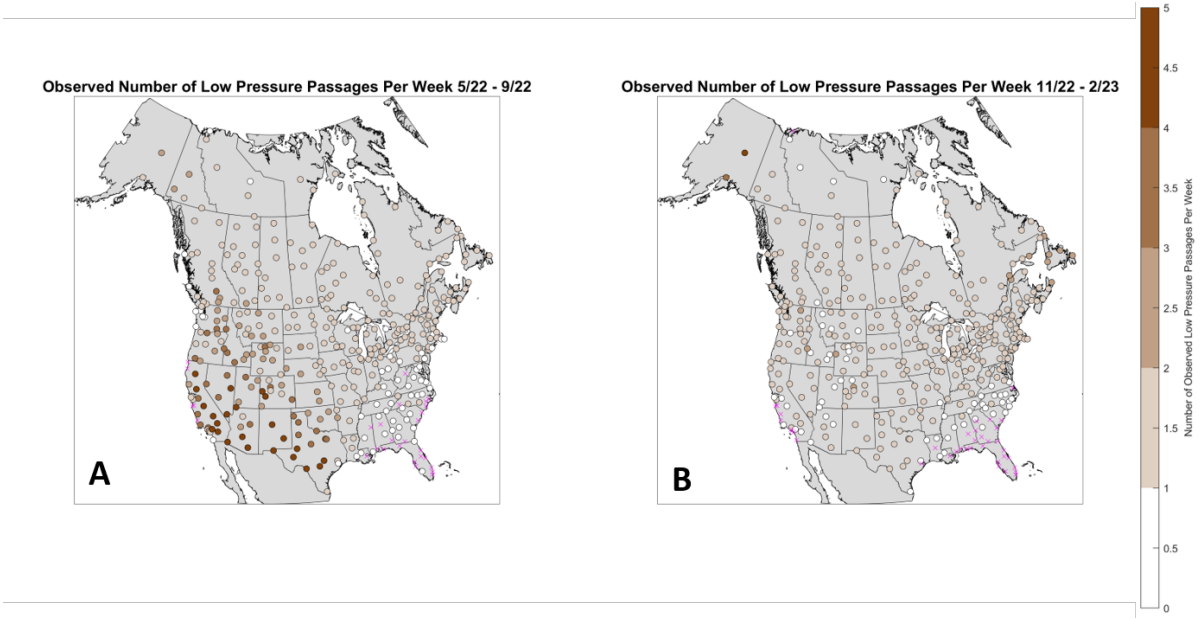

Figure 3.27: Observed number of low pressure passages per week across North America from A) May - September 2022 and B) November 2022 - February 2023. Stations marked with a pink 'X' indicate stations where fewer than 12 forecast low pressure passages were paired with observed low pressure passages. Biases at these stations are considered non-representative and were excluded.

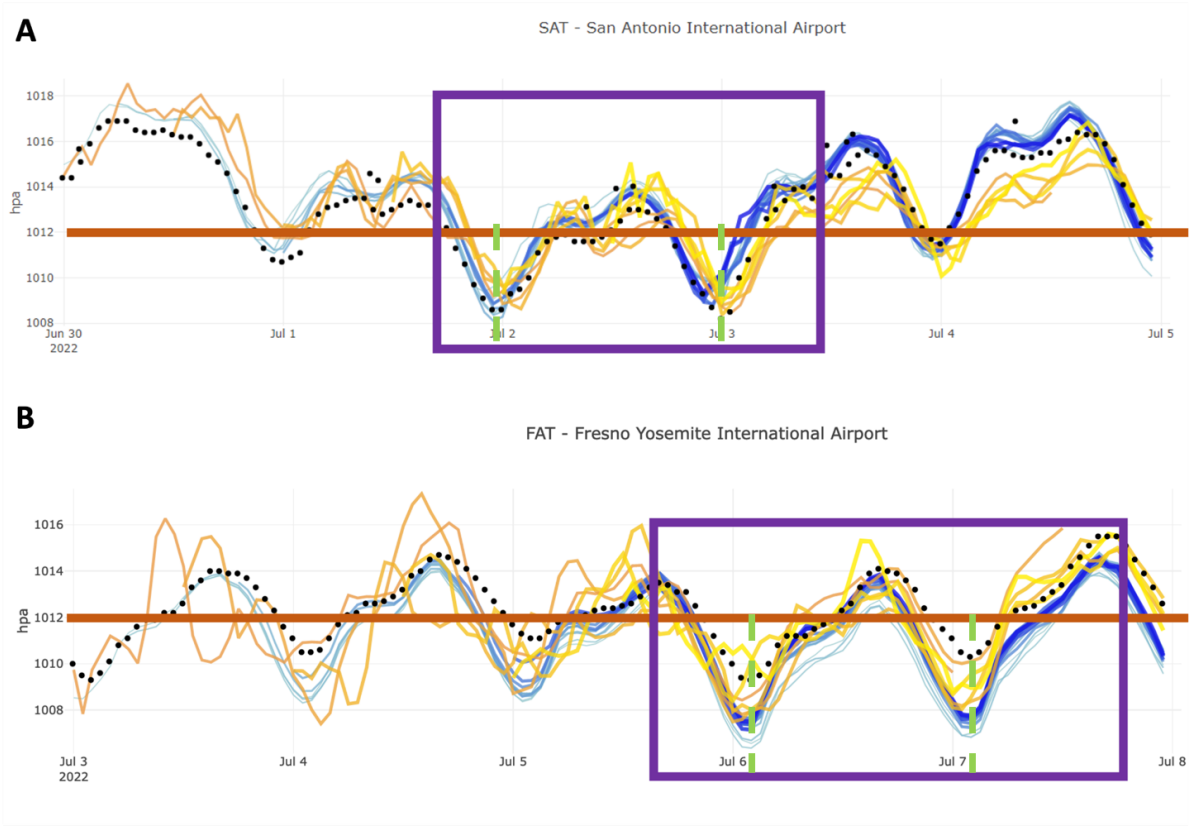

Figure 3.28: Observed (black) and forecast (COAMPS - yellow, GFS - blue) pressure trace values for A) KSAT - San Antonio, Texas from June 30 - July 5, 2022 and B) KFAT - Fresno, California from July 3 - July 8, 2022. Pressure trace values in both plots cycle up and down each day (similar to diurnal temperature cycles) instead of remaining relatively constant with no major dips or rises when a low pressure system is not present. Pressure traces marked in purple boxes indicate observed and forecast pressure traces that indicate a false low pressure passage. Dashed green lines indicate approximate time of low pressure passage.

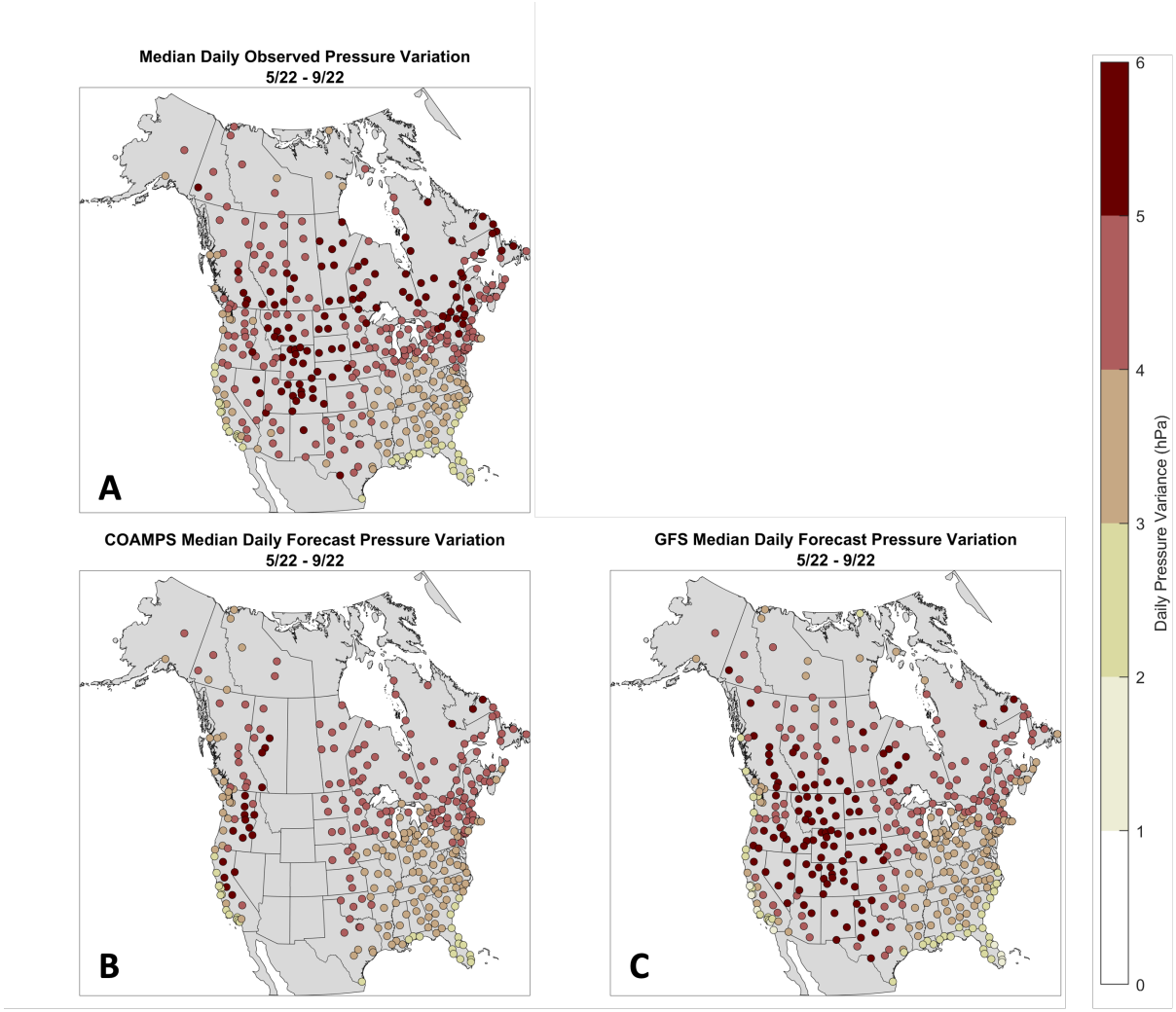

Figure 3.29: Median daily observed and forecast pressure variation (hPa) for May 2022 - September 2022 for A) Observations B) COAMPS C) GFS

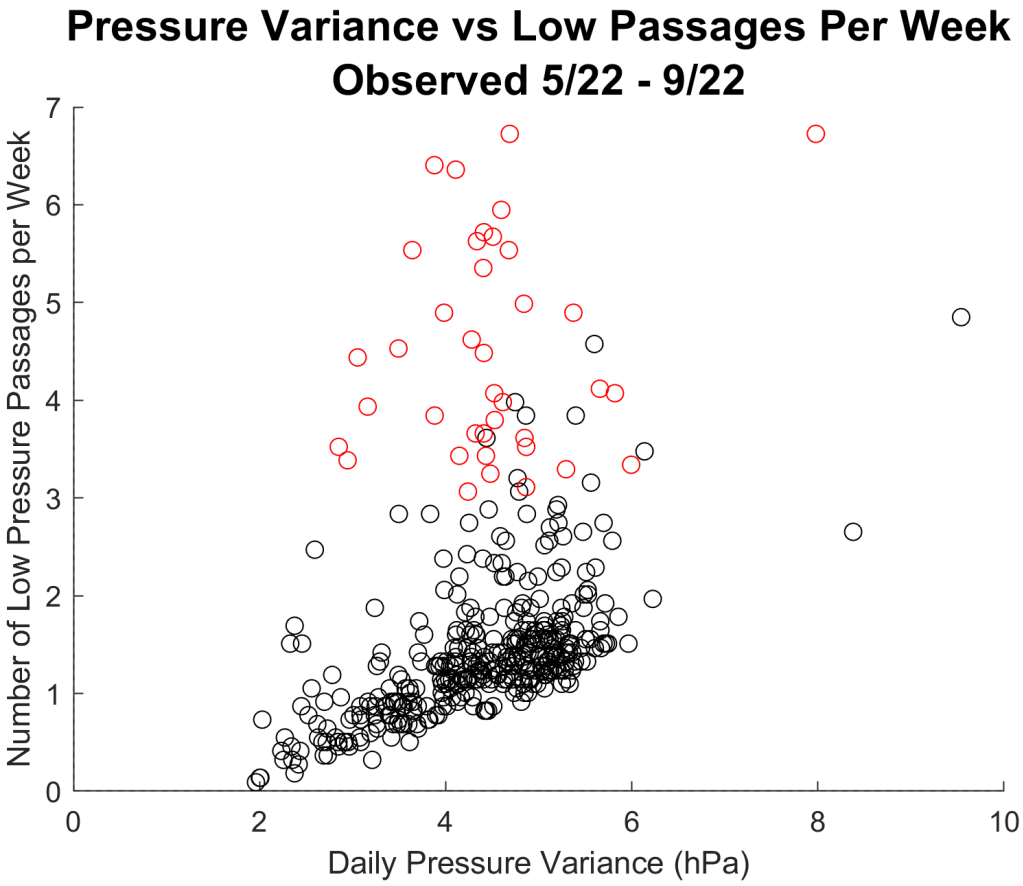

Figure 3.30: Scatter plot of observed daily pressure variation (hPa) and the number of observed low pressure passages per week at each station from May 2022 - September 2022. Stations in the southwest are marked with a red circle, stations outside of the southwest are marked with a black circle.

(from June to October) is southeastern California and southwestern Arizona (Rowson and Colucci 1992). Thermal lows are most common during the hottest summer months. As surface temperatures warm, surface air parcels become more positively buoyant and begin to rise, eventually leading to the formation of a thermal low (Johnson 2003). Thermal low strength fluctuates diurnally, strengthening during the day as temperatures increase and weakening during the night as temperatures decrease (Johnson 2003).

Time series of surface pressure (Fig. 3.28) indicate that our low pressure passage detection method can be triggered by diurnally varying pressures consistent with thermal lows. In order to further investigate the occurrence of thermal lows, we examined the distribution of median daily pressure variation from 5/22 - 9/22 for observations, COAMPS, and GFS (Fig. 3.29), where daily variation is calculated by finding the difference between the maximum and minimum pressure over a 24 hour period. For observations, COAMPS, and GFS the largest daily pressure variations were located in the Intermountain West and in the Southwestern United States. Most stations in the Eastern and Southeast US have smaller daily pressure variations (3-4 hPa) compared to the Southwest (4-6 hPa).

Analysis of station by station median daily pressure variance and the number of observed low pressure passages per week (Fig. 3.30) shows that 39 out of 50 stations with more than 3 low pressure passages per week occur in the Southwest. While there is an association between increasing daily pressure variance and increasing number of low pressure passages per week for the data set as a whole, this relationship is not apparent for the subset of southwest stations (red dots in Fig. 3.30).

In future work, further refinements to low pressure passage criteria to filter out thermal lows are needed. This is particularly an issue for the summer season. Key additional criteria could include information on the frequency and timing of lows, for example a filter that removes pressure lows that coincide in timing with diurnal heating.

Pressure Tendency Offset Bias	Median Offset Bias (hours)	25th Percentile (hours)	75th Percentile (hours)
COAMPS (summer)	+0.0	+3.00	-3.00
GFS (summer)	-1.0	+1.00	-2.75
COAMPS (winter)	+0.5	+3.00	-2.00
GFS (winter)	+0.0	+2.00	-2.00

Table 3.4: Median biases for pressure tendency offsets in hours for COAMPS and GFS in summer and in winter. Excludes stations with fewer than 12 low pressure passage events.

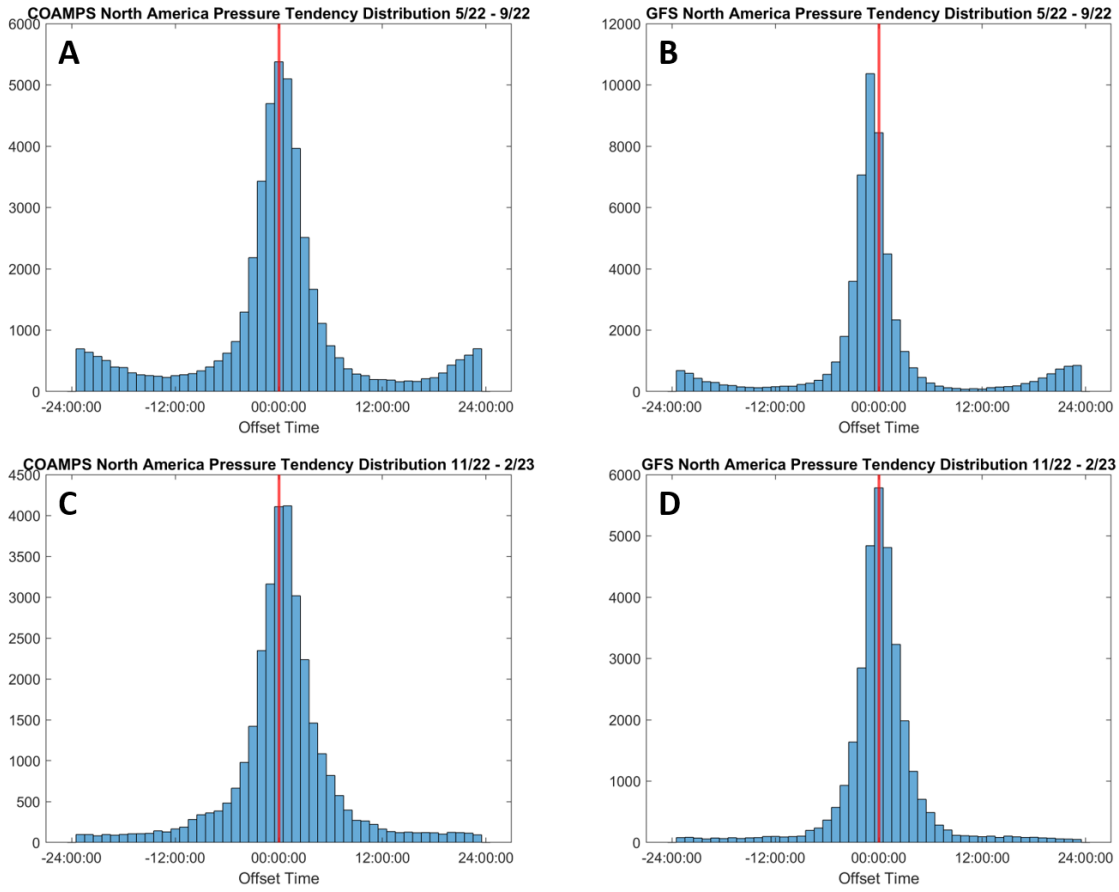

Figure 3.31: Distributions of timing offsets from all matched observed and forecast low pressure passage offsets. The y-axis shows the number of low pressure passages that occurred at each offset time, from A) COAMPS in 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23. Each valid time is evaluated for 8 forecast initializations.

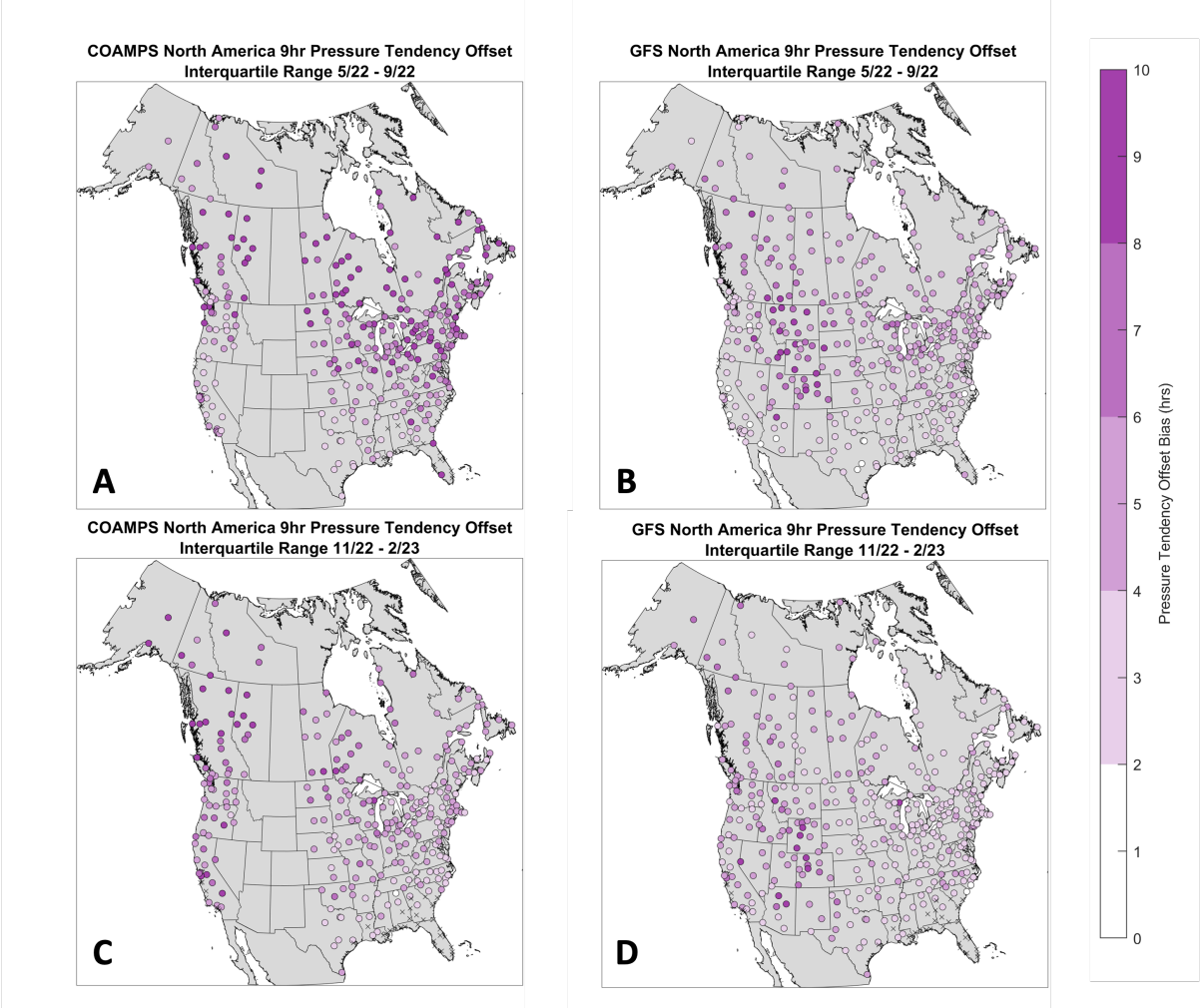

Figure 3.32: Interquartile range between the 25th and 75th percentiles for low pressure passage offset timing biases at each station from COAMPS (A, C; left panel) and GFS (C, D; right panel) valid for May - September 2022 (A, B; top row) and November 2022 - February 2023 (C, D; bottom row). The larger the interquartile range, the more error prone a station is.Stations marked with a black 'X' have too few samples to yield representative statistics.

In terms of interpreting our results, we infer that the vast majority of low pressure passages in winter detected with the current method are associated with synoptic-scale storms while those in summer are a mix of storms and thermal lows. The distribution of timing errors for matched forecast and observed low pressure patterns for all stations combined is centered near zero (Fig. 3.31), and median values are less than one hour (Table 3.4) indicating that on average both models have reasonable skill with low pressure center passage timing. Each observed pressure event is matched to the closest forecast event in

time within +/- 24 hours for a given initialization. Since each valid time is examined with respect to 8 model runs (initializations), one needs to divide the sample sizes by 8 to get sample sizes equivalent to those for a specific lead time.

There are regional and seasonal differences in the interquartile ranges of low pressure center timing errors (Fig. 3.32). In summer, COAMPS has the largest interquartile ranges in Northeast and in the Great Lakes region with values over 8 hours (Fig. 3.32 A). Winter interquartile ranges tend to be smaller than those in summer. In both summer and winter, GFS has the largest interquartile ranges in the Intermountain West which is outside of the COAMPS domains.

3.6 Timing of Precipitation Events

Similar to assessing how well numerical weather prediction models forecast the timing of low pressure passages, we assessed numerical weather prediction model forecast skill with regards to precipitation event timing. We restrict our analysis of precipitation event time to CONUS. Precipitation amounts from the Canadian stations are not part of the MADIS data feed used in this analysis and only the Alaska stations in Fairbanks, Anchorage, and Juneau reliably report hourly precipitation amounts. Precipitation event timing biases for both the summer and winter are examined for matched events (Section 2.7) over all times of day (Fig. 3.33). There is a regional pattern in the number of matched precipitation events corresponding to seasonal precipitation climatologies (NCEI 2023) with more events in the southeast US in summer compared to other locations and more events across northern tier as compared to southern tier of CONUS in winter.

3.6.1 Event Start Time and End Time

Overall, the tendency is for both COAMPS and GFS forecasts for events matched to observations to start precipitation events a bit too early and end them substantively too late (Fig. 3.34 and Fig. 3.35). The error distributions in Fig. 3.34 show that COAMPS and GFS are both strongly skewed to the left, or, towards starting precipitation events too early. The error distributions for event end time biases for COAMPS and GFS are both skewed to the right, ending precipitation events too late, with COAMPS having a stronger late bias in the summer and GFS having a stronger late bias in the winter (Fig. 3.35).

Figures 3.36 and 3.37 show the geographic distribution of start time and end time biases. Stations marked with black X's are those that did not have at least ten paired precipitation

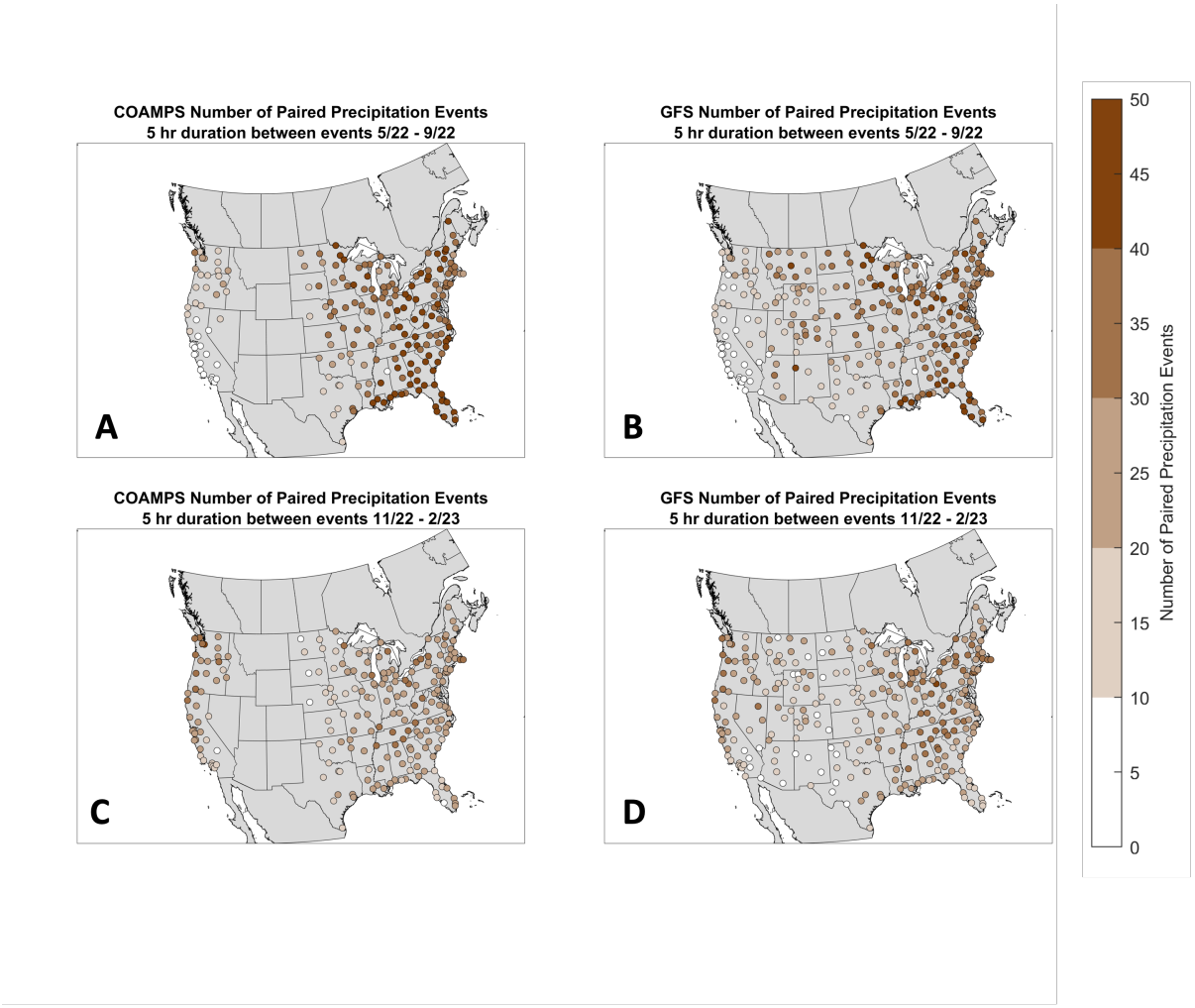

Figure 3.33: Geographic maps of the number of precipitation events matched between observations and 48-hour lead time forecast for A) May 2022- September 2022 COAMPS, B) May 2022- September 2022 GFS, C) November 2022 - February 2023 COAMPS, and D) November 2022 - February 2023 GFS.

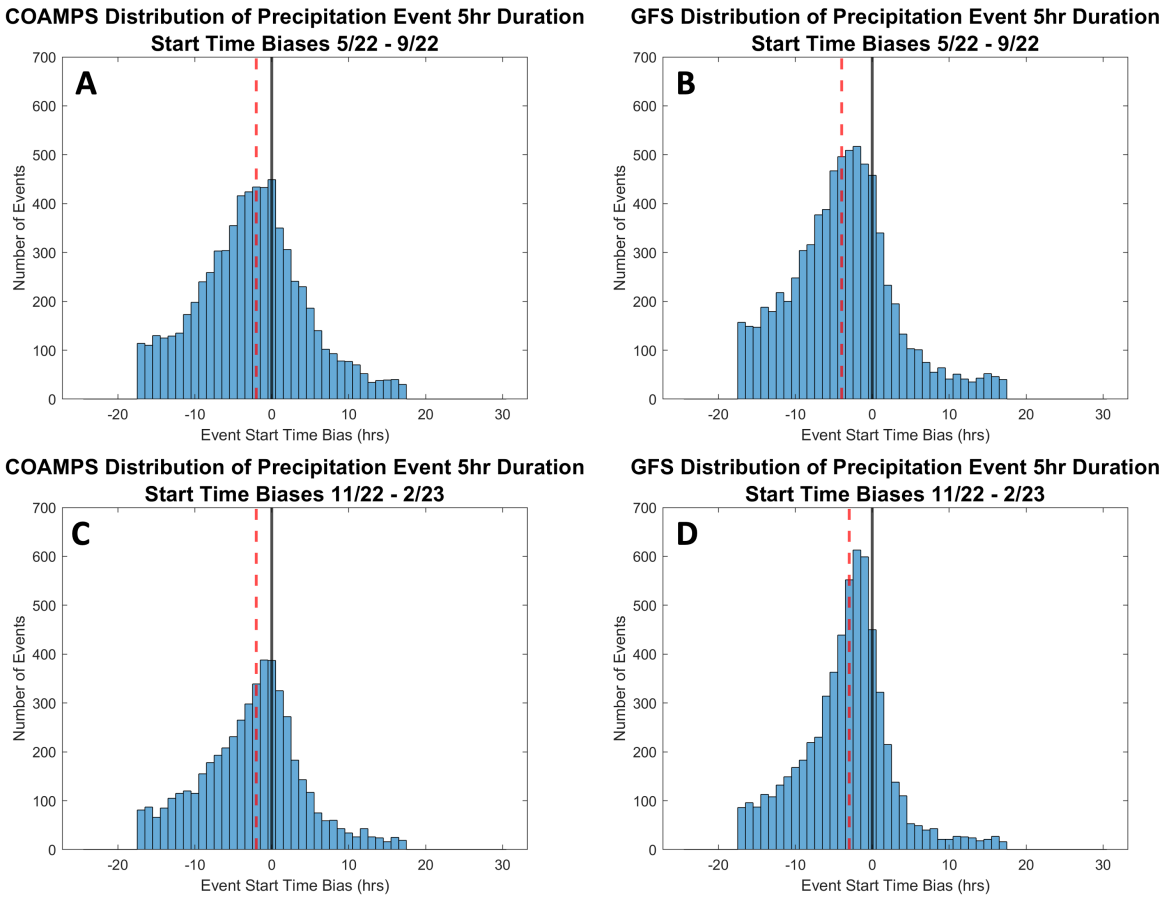

Figure 3.34: Histograms of COAMPS and GFS precipitation event start time errors for A) COAMPS May 2022- September 2022, B) GFS May 2022- September 2022, C) COAMPS November 2022 - February 2023, and D) GFS November 2022 - February 2023 for 48-hour lead times. Dashed red line represents the median start time bias. Solid black line represents a start time bias of 0 hours.

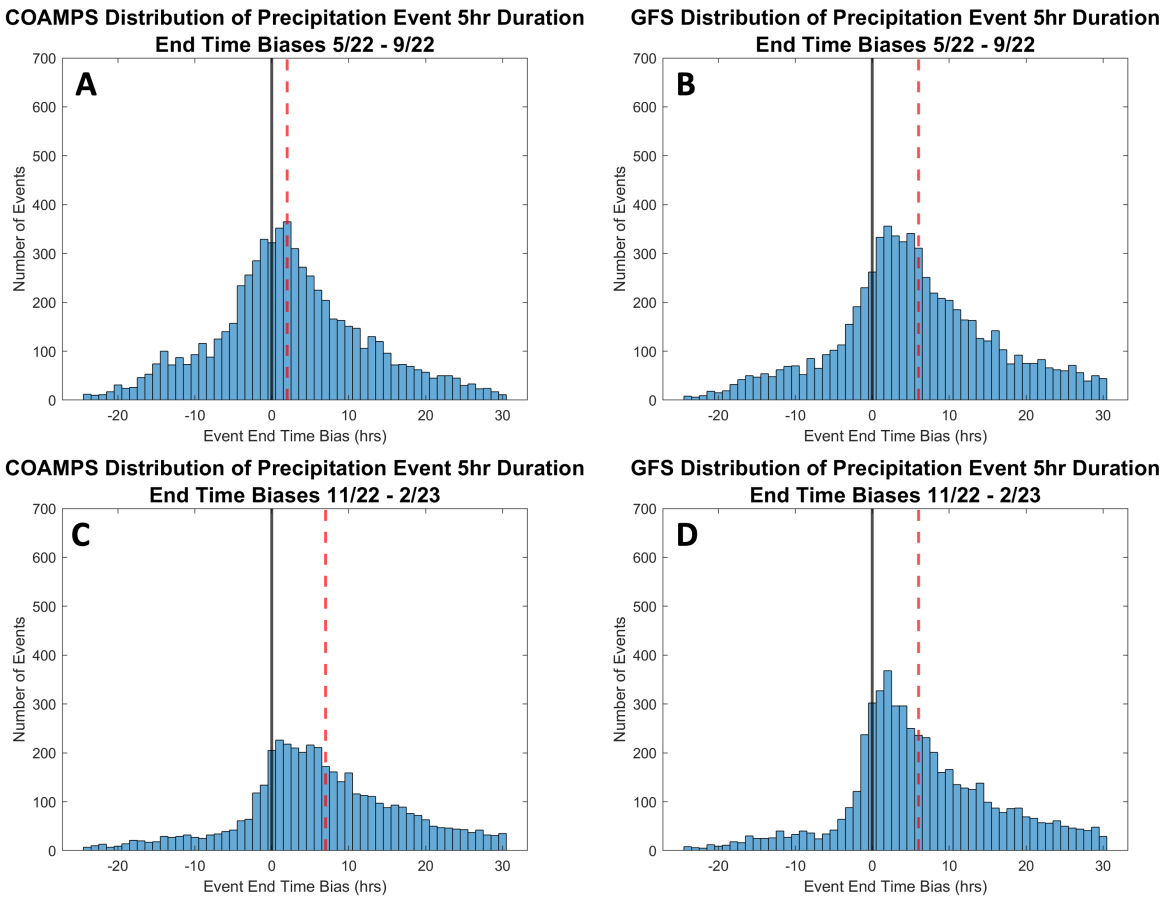

Figure 3.35: Same as Fig. 3.34 but for precipitation event end time errors for 48-hour lead times.

events between the model and observations. Nov 2022 - Feb 2023 was fairly dry across most of the Midwest. Additionally, several smaller airports in the Midwest and Intermountain West experienced instrument issues which artificially yielded only a few precipitation events. One example is KMOT (Minot, ND) where the precipitation sensor stopped reporting in October 2022. The nearby station, KBIS (Bismark, ND) had an uninterrupted record and did meet the minimum ten paired events criteria. Since stations with missing precipitation marked those times as 0 in/hr instead of NaN, stations with less then 10 paired precipitation events or missing precipitation data are marked with a black 'X' on precipitation plots.

With the exception of a handful of stations, precipitation start biases across the United States were predominantly too early (Fig. 3.36). There is no clear geographic pattern to locations of the larger start time biases for either COAMPS or GFS.

Looking at station by station analysis, nearly all the precipitation end time biases are too late for both COAMPS and GFS (Fig. 3.37). Precipitation event end time biases were typically larger in the winter for COAMPS and GFS than in the summer. There is a slight geographic pattern to end time biases, with regions in the Northeast typically having the largest end time biases in both COAMPS and the GFS in summer and with. Regions in the Plains states tended to have the smallest end time biases for the GFS.

3.6.2 Event Duration

To better understand the relationships between model precipitation event duration and observed precipitation event duration, we compared the joint distributions of paired model and observed precipitation events duration in terms of observed duration versus the difference between forecast and observed duration (Fig. 3.38). The forecast duration errors have a long tail out to a few events lasting > 48 hours. These very long duration events tend to occur in winter and in the Pacific Northwest and Northeast (not shown). In Figure (Fig. 3.38) correctly forecast durations would fall along a horizontal line at zero difference between forecast and observed durations in the joint distributions. Focusing on the winter season precipitation timing data, which we consider as more reliable, the joint frequency pattern indicates that the magnitude error in duration is not proportional to observed duration.

The sharp edge of the distribution in the lower left corners of the subpanels in Figure 3.38 is an outcome of the way the joint distributions are plotted and that the smallest possible forecast duration is 1 hr. For example, for an observed duration of 6 hours, the minimum difference between forecast and observed durations is 1 hr - 6 hr = -5 hr.

Figure 3.36: Median precipitation start time biases at individual stations for A) COAMPS 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23, and D) GFS 11/22 - 2/23. Stations with less than 10 paired precipitation events are marked using a black 'X'.

Figure 3.37: Median precipitation end time biases at individual stations for A) COAMPS 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23, and D) GFS 11/22 - 2/23. Stations with less than 10 paired precipitation events are marked using a black 'X'.

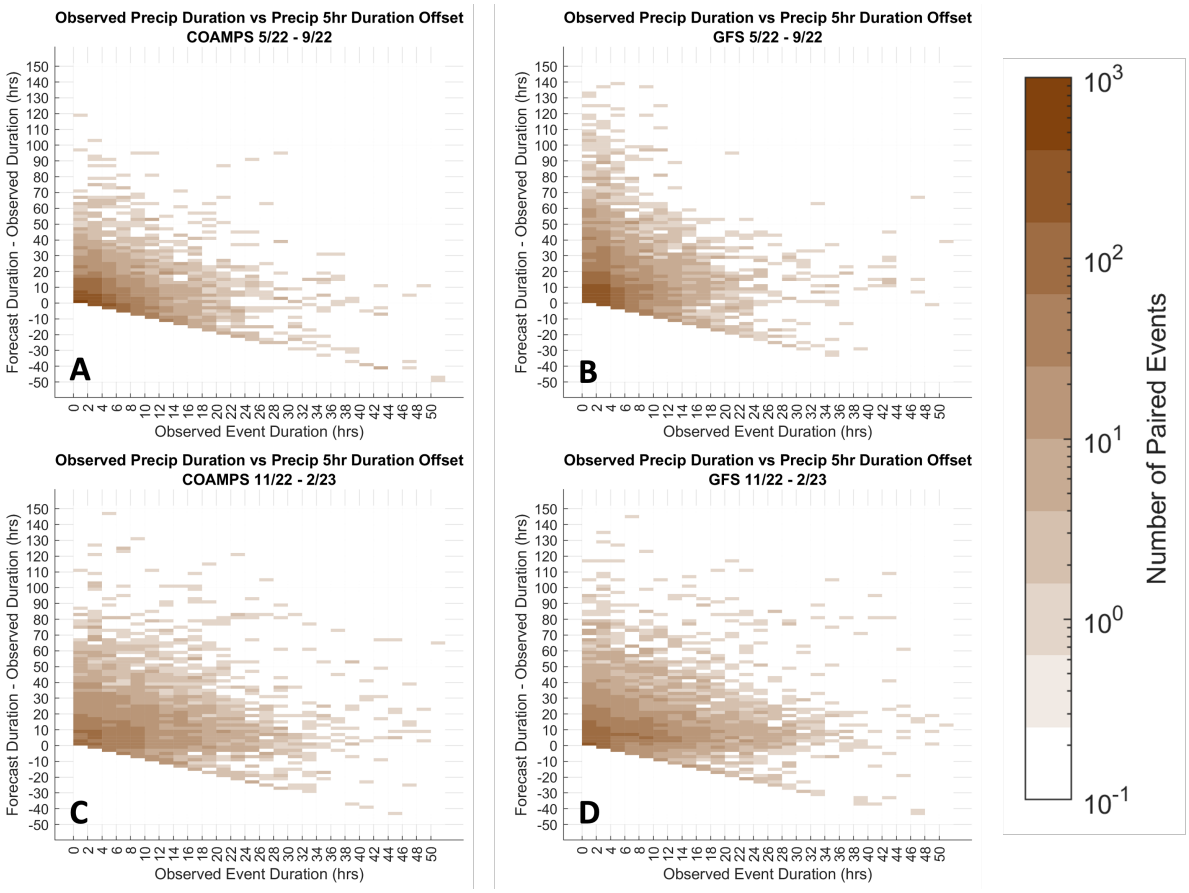

Figure 3.38: Paired model and observed precipitation event duration compared against each other for A) COAMPS 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23, and D) GFS 11/22 - 2/23.

3.6.3 Missed Precipitation Events

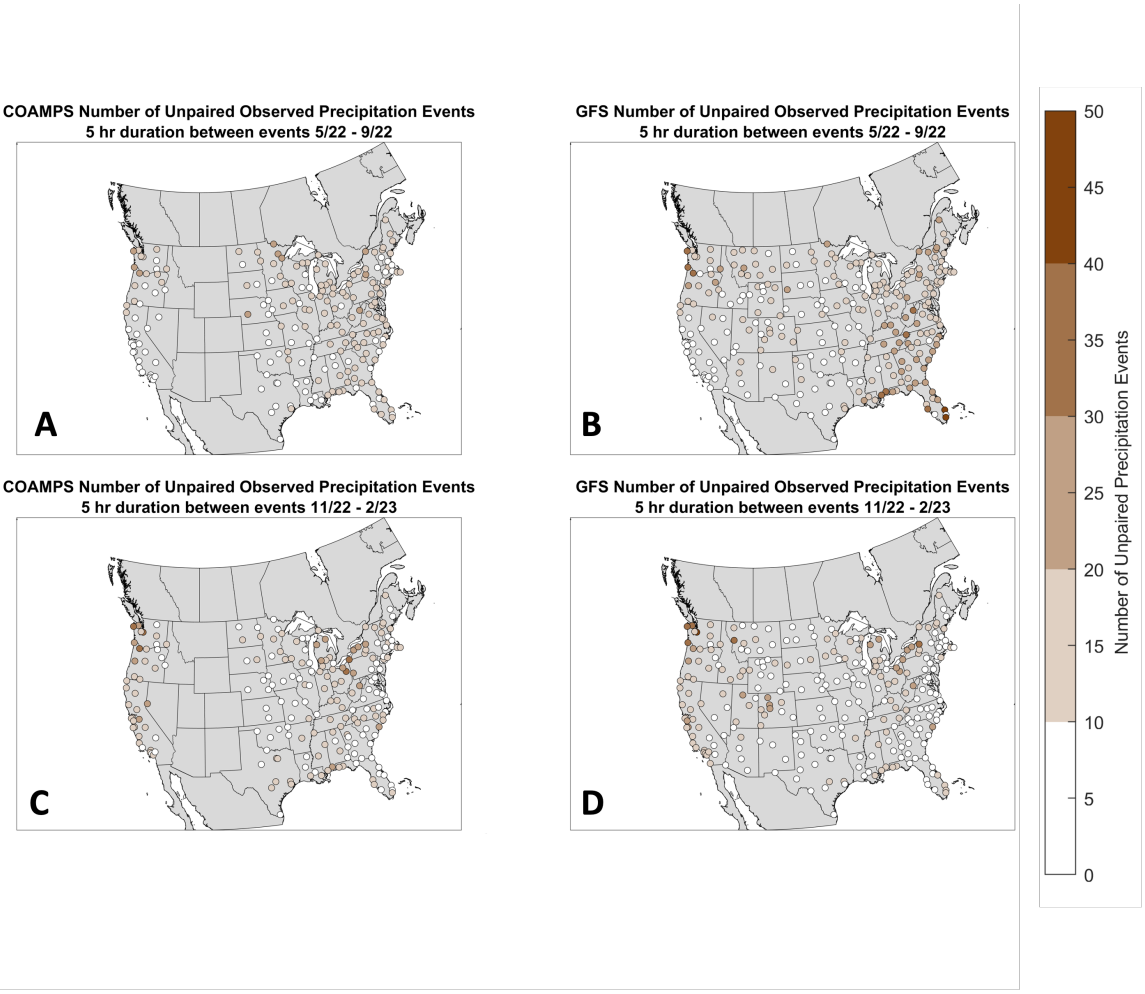

Figure 3.39: Number of observed precipitation events at each station that were not paired with a forecast model precipitation event for A) COAMPS 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23, and D) GFS 11/22 - 2/23.

Overall, there are more unpaired events (observed events without a corresponding forecast and forecast events without a corresponding observation) in summer when deep convective precipitation events are usually smaller in size and shorter duration as compared to winter storms (Figs. 3.39 and 3.40). In the Florida panhandle in summer, COAMPS misses fewer actual precipitation events than GFS (Fig. 3.39).

The station by station numbers for forecast precipitation events with no matching

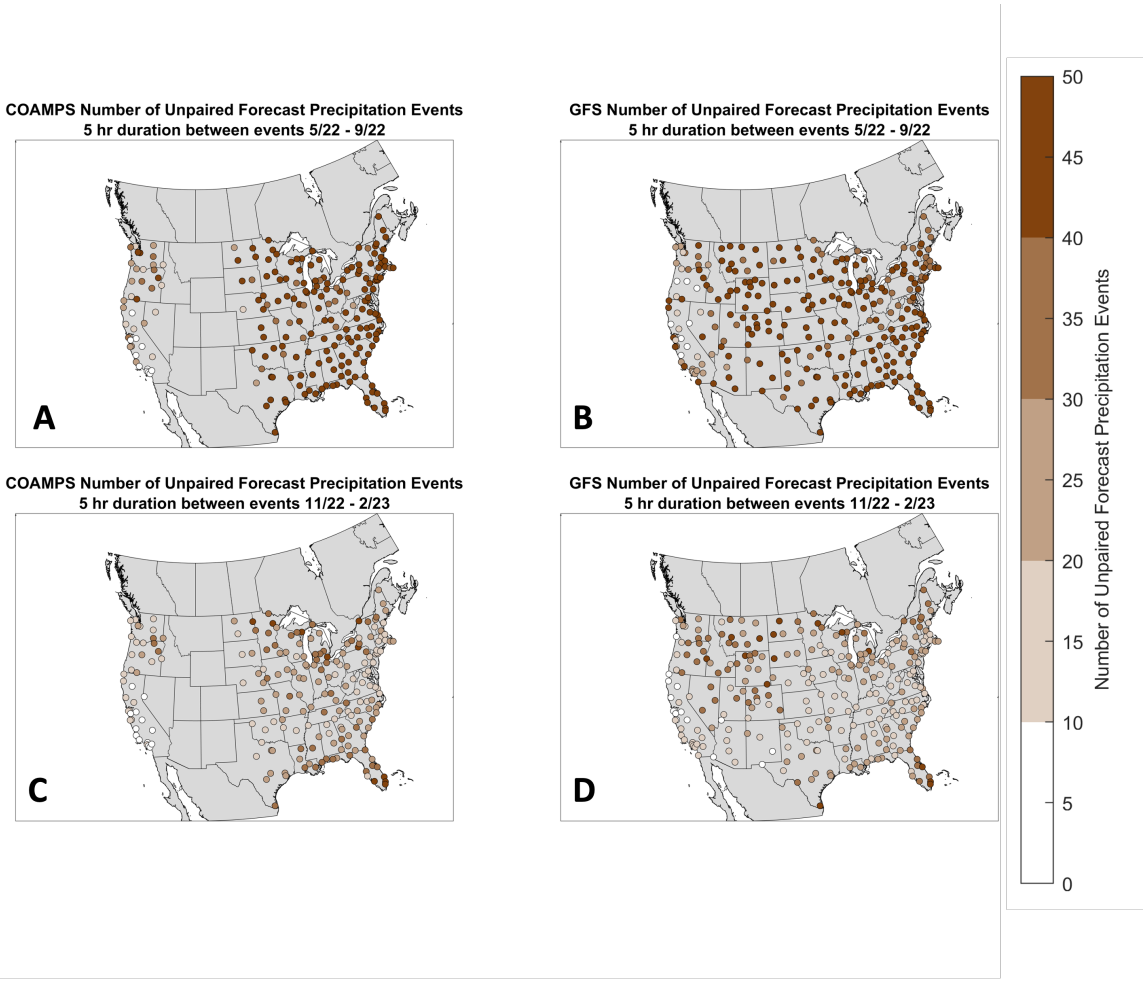

Figure 3.40: Number of forecast precipitation events at each station that were not paired with an observed precipitation event for A) COAMPS 5/22 - 9/22, B) GFS 5/22 - 9/22, C) COAMPS 11/22 - 2/23, and D) GFS 11/22 - 2/23.

observation are larger than those of observed precipitation events without a forecast. More forecast events that do not have a matching observed event occur in summer than in winter. In the summer, the largest number of unpaired forecast events for COAMPS occurs in the Southeast and for GFS in the Southeast and Intermountain West (Fig. 3.40).

3.6.4 Sensitivity Tests

Sensitivity tests were conducted to assess how precipitation start and end time biases shifted depending on the length of the precipitation start time pairing window and the allowed number of hours between precipitation events. Precipitation start times were paired within windows of +/- 21 hours, +/- 18 hours, +/- 15 hours, +/- 12 hours, +/- 9 hours, +/- 6 hours, or +/- 3 hours. The number of allowed hours between precipitation events (gap length) tested was 3 hours, 5 hours, 7 hours, and 9 hours. Sensitivity test results by a pairing window of +/- 18 hours across all event gap durations (Fig. A.9, A.10) showed no large changes for start/end time biases depending on event gap length. Separate examinations of the start and end time biases associated with a <5 hour duration between precipitation events across all pairing windows also showed little sensitivity to window length (Figs. A.11, A.12). While not shown in this paper, sensitivity tests for all other pairing windows and event durations yielded no significant changes in calculated biases. There is minimal sensitivity in precipitation biases to the window of time that precipitation events are paired within and the duration of time that can pass between precipitation events.

3.7 Summary and Implications

COAMPS and GFS both have stronger temperature biases in the winter and weaker temperature biases in the summer. Temperature biases in the winter are more strongly influenced by the amount of observed cloud cover than in summer. In summer, other factors, such as soil moisture errors or issues adequately resolving terrain, may contribute to the subset of larger temperature biases. Dewpoint biases do not appear to be a result of variation in the amount of cloud cover. There is a distinct regional pattern in dewpoint errors with the western United States being biased towards too dry dewpoints and the eastern United States being biased towards too moist dewpoints. Previous work by Lin et al. (2017), suggested that certain model biases (soil moisture, dewpoint) can combine and create a scenario that leads to enhanced temperature biases. They argue that if soil moisture is forecast to be too dry, this will lead to less evapotranspiration, less moisture

in the air, and less precipitation forecast in the model which then results in evaporational cooling being poorly represented, thereby generating a warm temperature bias. Currently, in situ soil moisture observations are not taken consistently across the United States so there is a dearth of data to validate satellite retrievals which are the basis of soil moisture inputs to models.

Additionally, this study reaffirmed the findings of Patel et al. (2021) by demonstrating that in the winter the GFS had a cold bias during the afternoon and a warm bias overnight. There is not much variation between the temperature bulk analysis results for a 48 hour and 72 hour leadtimes. This lack of variation indicates that, ignoring the timing of certain events, the models remain consistent in their distributions of forecast temperatures up to three days out.

The underlying reasons for the underforecasts of temperatures outside of the $<$ 10th percentile and $>$ 90th percentile are not readily apparent but may be related to the tendency to evaluate model output and target model improvements to reduce errors in the most common occurring (average) conditions rather than the tails of the distribution.

Several biases tend to be larger in regions of elevated terrain, such as wind speed and direction, temperature, and dewpoint (Sections 3.3, 3.1.1, 3.2). This is most evident for the GFS in the Intermountain West where the strongest temperature, dewpoint, wind speed, and wind direction biases in the winter and summer are located. Smaller model grid resolution will resolve terrain details better but is not a cure in itself as the next chapter that compares the COAMPS model at two grid resolutions will illustrate.

The timing of forecast low pressure passages for COAMPS and GFS occurs within a reasonable timing boundary of within +/- 2 hours of observed low pressure passages in winter. Our intent in developing a metric to detect low pressure passages was to focus on synoptic systems. Currently this metric is picking up stronger diurnal thermal lows as well and may need to be refined to remove those. With the current metric, we found that the model had many more thermal lows than observed for the Southwest US in summer which suggests that the models may be overpredicting the magnitudes of thermal lows. Further examination of the daily pressure variance associated with thermal lows and synoptic low-pressure passages will help to further filter and understand the nature of thermal lows in the Southwest.

CHAPTER

4

RESULTS 2 - CALIFORNIA - COAMPS AT TWO GRID RESOLUTIONS

When model parameterizations and initialization are identical, conventional wisdom indicates that a numerical weather prediction model with a finer grid should perform better than a coarser grid. Finer grids are able to better resolve relevant land surface details such as coastlines and gaps through mountain ranges (Parker 2015). This should lead to improvements in model forecast skill as the absence or presence of mountain gaps can make the difference between moist flow that is blocked by a mountain range or goes through the gap to the neighboring watershed which in turn will impact temperature, dewpoint, wind speed/direction, and potentially precipitation forecasts for a region. Previous model studies (e.g. Mass et al. 2002; Hoadley et al. 2004) have noted that a higher resolution model does not always guarantee a more accurate forecast. Mass et al. (2002) found that reducing modeling grid spacing from 30 km to 12 km in a mesoscale model improved overall forecast skill. In contrast, overall forecast skill degraded between a 12 km grid and a 4 km grid but some aspects such as precipitation amounts on windward slopes of a mountain range improved with the smaller grid size.

The Navy runs two operational versions of their regional COAMPS model for the Califor-

ICAO ID	Station
KACV	Humboldt County, CA*
KBUR	Burbank, CA
KCEC	Crescent City, CA*
KFAT	Fresno, CA
KLAX	Los Angeles, CA
KLGB	Long Beach, CA
KMOD	Modesto, CA
KMRY	Monterey, CA*
KNID	China Lake Naval Station, CA
KOAK	Oakland, CA*
KONT	Ontario, CA
KPRB	Point Robles, CA
KRDD	Redding, CA
KSBA	Santa Barbara, CA*
KSFO	San Francisco, CA
KSJC	San Jose, CA
KSMF	Sacramento, CA
KSMX	Santa Maria, CA
KSNA	Orange County, CA*
KSNS	Salinas, CA
KUKI	Ukiah, CA

Table 4.1: ICAO ID and locations for 21 ASOS stations utilized in this study across California. Stations that are located in the wrong surface type (ocean surface type rather than on land surface type) are marked with an asterisk.

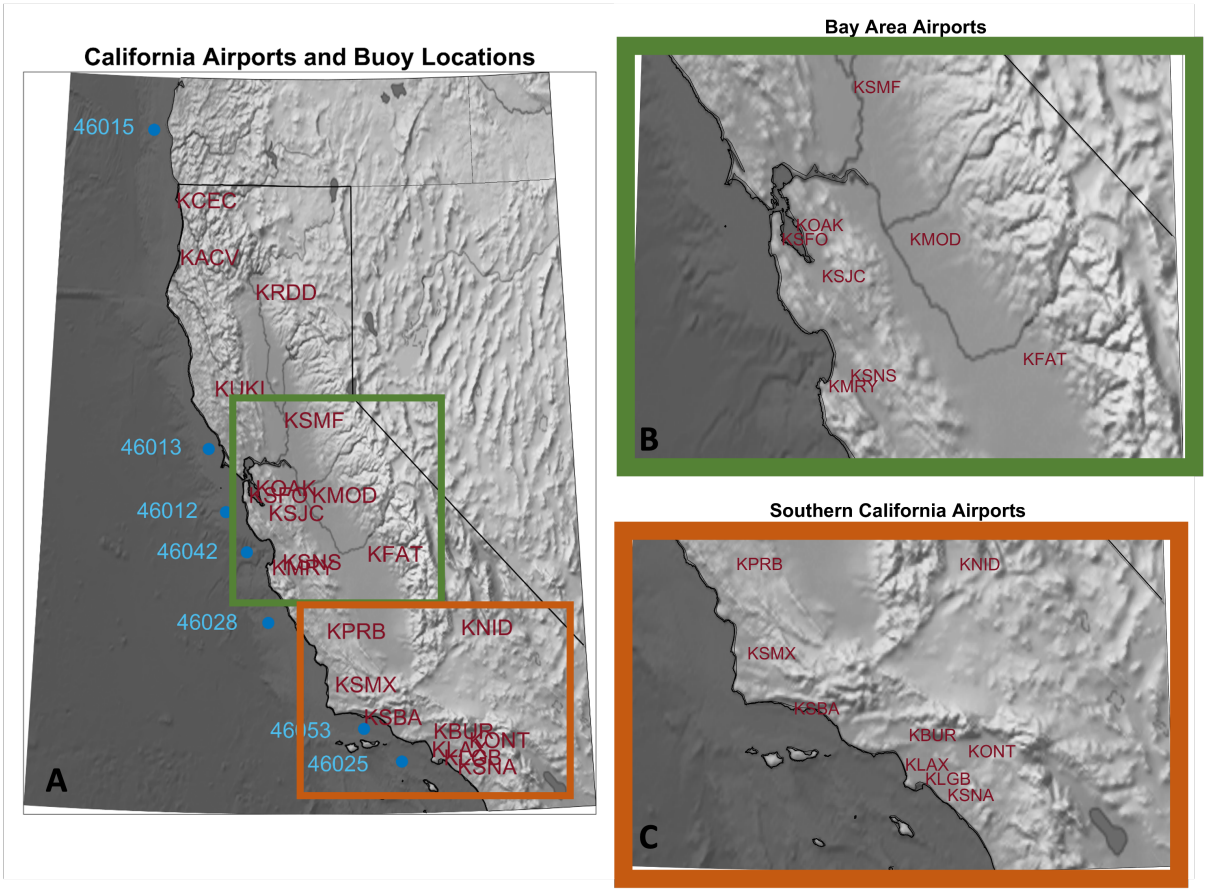

Figure 4.1: A) Map of all land airport ASOS stations and ocean buoy stations in the NEPAC and CENCOOS domains. B) Map of all airport ASOS stations in the Bay Area C) Map of all airport ASOS stations in Southern California.

nia region, CENCOOS (∼ 3.7 km grid spacing) and NEPAC (∼ 15.5 km grid spacing) (Fig. 2.3, and 2.2, and 4.1). These runs use nearly identical parameterizations and initializations and provide the opportunity to examine COAMPS model performance when the grid spacing is changed.

The land/sea mask used in a model can cause issues in coastal regions where locations that are actually located on land are indicated as located in the ocean in the model. This misclassification of surface type negatively impacts forecasts in coastal regions as land stations have a larger diurnal variation (particularly for temperature) than ocean stations. If a surface observation station at a coastal airport is mistakenly located in an ocean grid box, the diurnal cycle is less likely to be properly forecast. Examination of the NEPAC land-sea mask indicates that KCEC, KACV, KOAK, KMRY, KSBA and KSNA are placed in ocean grid boxes. As a result of smaller grid spacing in the CENCOOS model, the CENCOOS land-sea mask has only KCEC and KOAK in ocean grid boxes. Errors at these misclassified locations are likely mostly primarily related to use of the wrong land surface type. For NEPAC and CENCOOS, sea surface temperature (SST) initial conditions are updated using the Coupled Ocean Data Assimilation (CODA) before each model run (Chen et al. 2003).

We analyzed the matched observation and model temperatures and winds for both CENCOOS and NEPAC for the summer season from May - Sep 2022 and for the winter season from Nov 2022 to Feb 2023 for 48-hour lead time forecasts. The temperature at 7AM local time is used to estimate the overnight low and temperature at 3PM local time is used to estimate daily high. There are 21 ASOS stations and 3 buoys (5 buoys in the winter) with sufficiently complete records within the region where the CENCOOS and NEPAC domains overlap (Table 4.1).

4.1 Temperature Biases at 7AM and 3PM

Overall both NEPAC and CENCOOS tend to have a warm bias in the summer (Table 4.2, Fig. 4.2) but while CENCOOS maintains a warm bias in the winter, NEPAC transitions to an overall cold bias (Table 4.3, 4.5). Surprisingly, CENCOOS with finer grid resolution than NEPAC has the stronger temperature biases in the summer (3K at 7AM and 2.3K at 7AM) and overnight in the winter (1.7K at 7AM). The one exception is during the day in the winter where CENCOOS has very weak average median biases and NEPAC has large (> 3K at 3PM) cold biases.

4.1.1 Summer Season

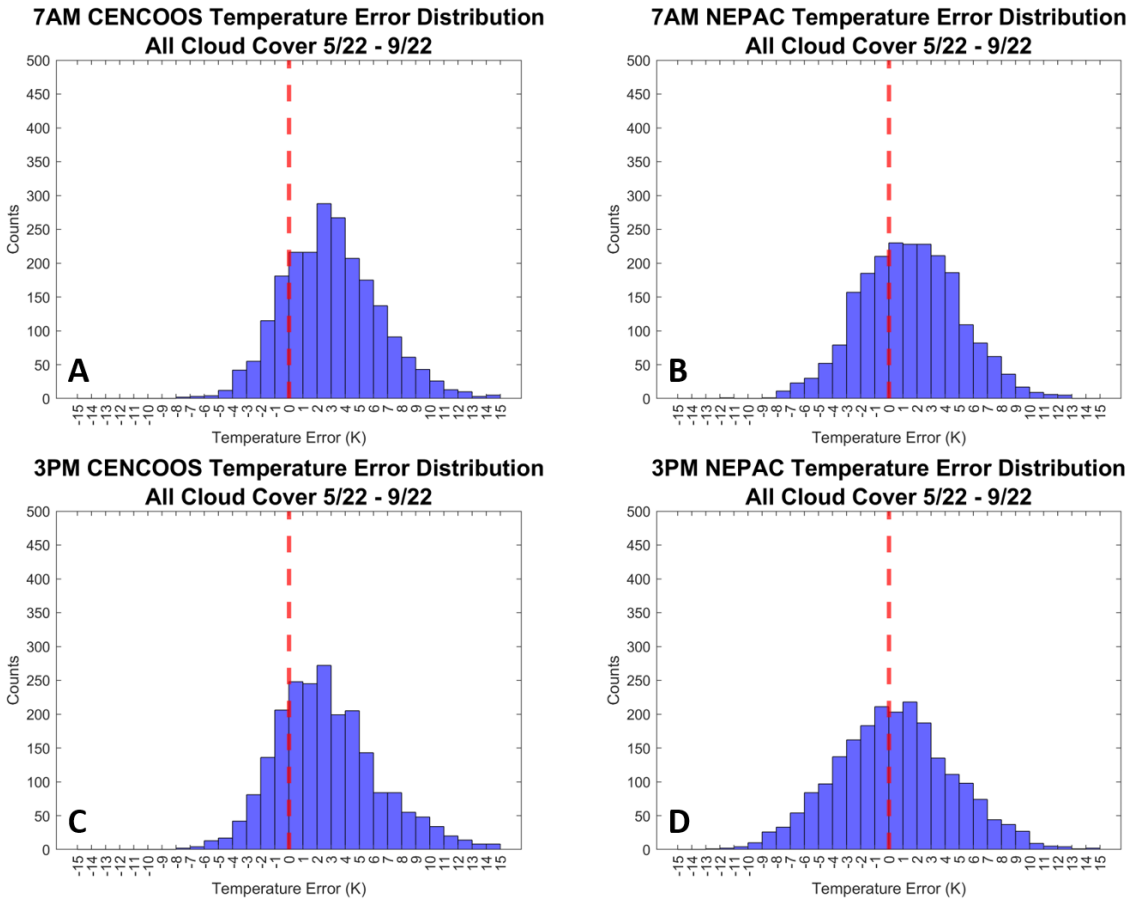

Figure 4.2: May - September 2022 (summer) distributions of temperature error of matched model forecast value minus observed value for 15 land based stations (stations in land grid boxes only, 6 land stations in ocean grid boxes excluded) COAMPS NEPAC and CENCOOS for all cloud cover conditions at a 48 hour lead time. A) COAMPS NEPAC 7AM B) CENCOOS 7AM C) COAMPS NEPAC 3PM D) CENCOOS 3PM.

For the summer season, the overnight lows have the largest errors with median bias at 7AM for CENCOOS of 2.9K, and NEPAC of 1.35K. CENCOOS also struggles with afternoon high at 3PM with median bias of 2.32K compared to 0.67K for NEPAC (Table 4.2) Both the CENCOOS 7AM and 3PM error distributions are skewed sufficiently toward higher temperatures to yield 75th percentile errors > 4K (Fig. 4.2). Unlike the findings in Section 3.1 for North America as a whole, in California, the summer magnitudes of temperature

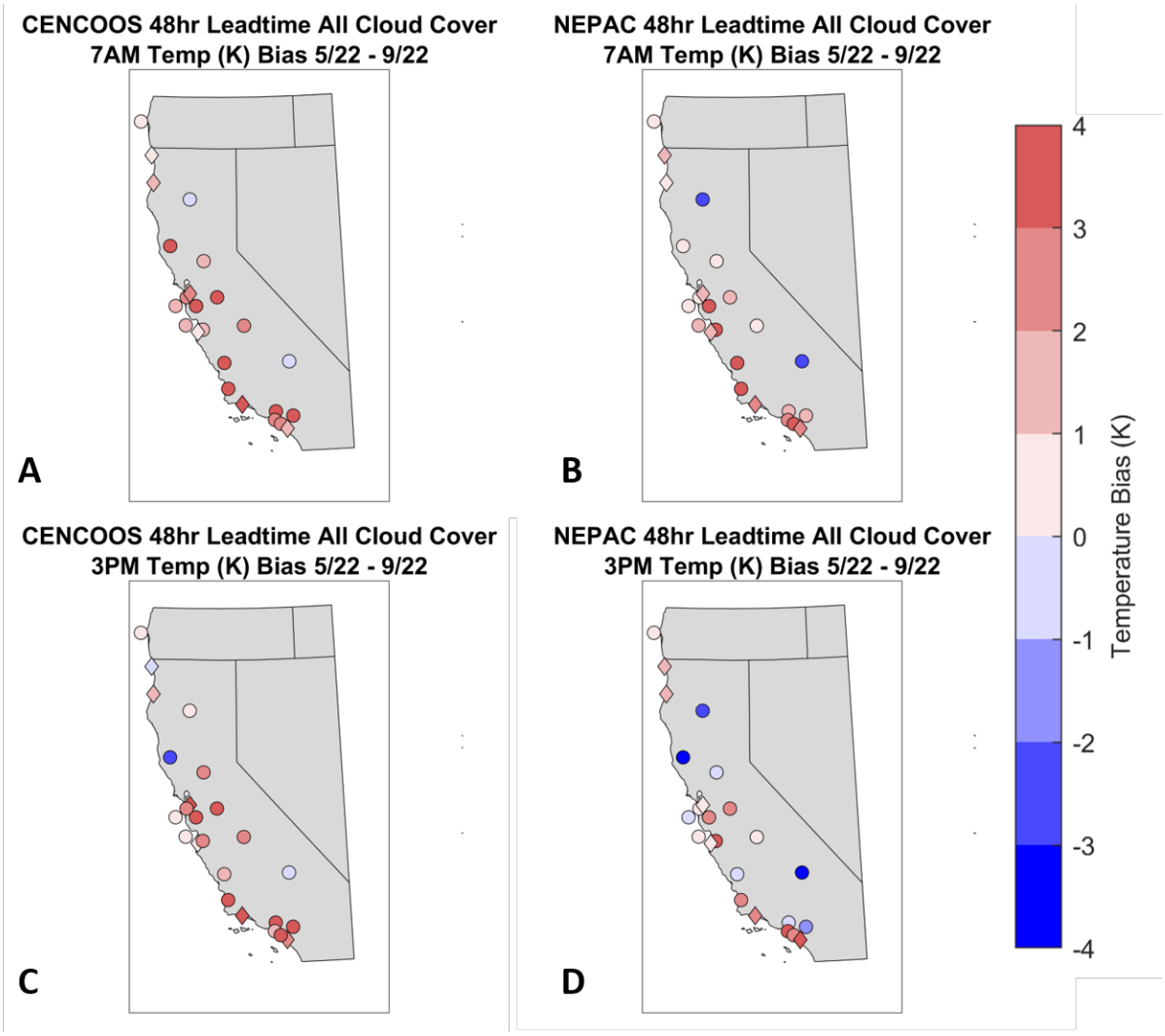

Figure 4.3: May - September 2022 map of temperature biases in degrees C at 7AM (A, B; top panel) and 3PM (C, D; bottom panel) for COAMPS NEPAC (B, D; right panel) and CENCOOS (A, C; left panel) models at a 48-hour lead time under all cloud conditions.

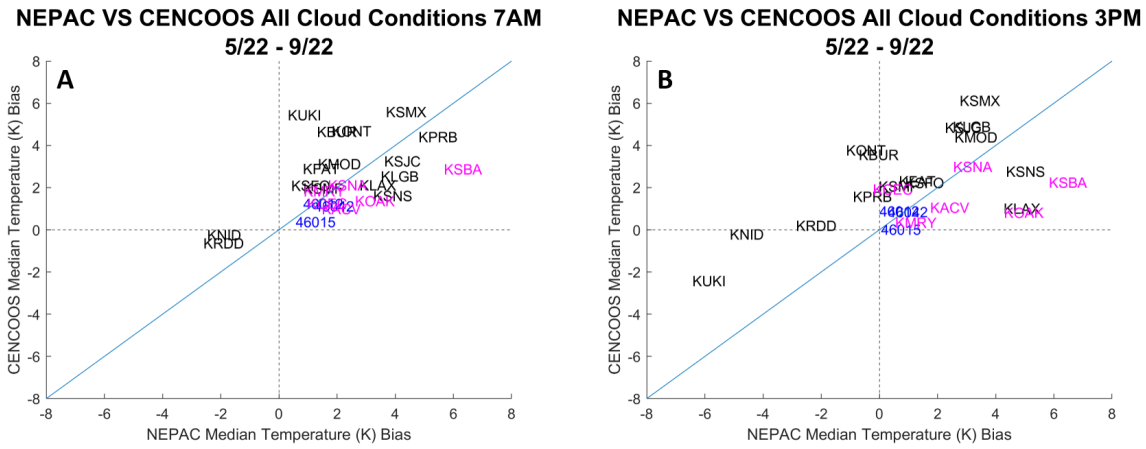

Figure 4.4: May - September 2022 scatterplot of COAMPS NEPAC temperature biases versus CENCOOS temperature biases by station for all cloud conditions at A) 7AM and B) 3PM. Blue line represents a 1-1 line where COAMPS NEPAC temperature biases equal CENCOOS temperature biases. In-land stations marked in black, buoys marked in blue, NEPAC stations placed in the ocean instead of on land marked in pink.

Summer Temperature Analysis: **NEPAC 3PM**	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+0.67	-1.09	+2.57
<25% Cloud Cover	+0.79	-1.10	+2.86
<50% Cloud Cover	+0.63	-1.09	+2.53
NEPAC 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+1.35	+0.52	+3.44
<25% Cloud Cover	+1.38	+0.86	+3.29
<50% Cloud Cover	+1.38	+0.72	+3.17
CENCOOS 3PM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+2.32	+1.16	+4.25
<25% Cloud Cover	+3.06	+1.21	+4.23
<50% Cloud Cover	+2.27	+0.98	+4.13
CENCOOS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+2.9	+2.00	+4.59
<25% Cloud Cover	+3.03	+2.27	+4.74
<50% Cloud Cover	+2.93	+2.18	+4.62

Table 4.2: Temperature bias analysis from May - September 2022 at 7AM and 3PM for COAMPS NEPAC and CENCOOS models. Median, 25th percentile, and 75th percentile values from the temperature bias distributions are further subset by observed cloud cover amounts.

bias errors are not strongly a function of observed cloudiness conditions (Table 4.2).

There is some variation in strength of temperature biases during the summer depending on geographical location within California (Fig. 4.3). Stations in Central and Southern California were dominated by warm biases at both 7AM and 3PM with stronger warm biases observed inland and weaker warm biases observed in coastal areas. Stations in Northern California were primarily dominated by weaker warm biases at 7AM and had weak cold and neutral biases at 3PM. The strongest warm biases are located in coastal Southern California (KSMX-Santa Monica Airport for 7AM and 3PM, KPRB-Paso Robles Municipal Airport for 7AM only), the Bay Area (KSJC-San Jose Airport for 7AM and 3PM and at KMOD-Modesto City Airport at 7AM and 3PM) (Fig. 4.4).

A few stations (KRDD, KNID, KUKI) have moderate to weak median cold biases in both models (Figs. 4.3 and 4.4). KRDD (Redding Regional Airport) is located near the base of mountainous terrain in Northern California which could contribute to the consistent cold bias here. KNID (China Lake Naval Station) is located close to Death Valley National Park and forecasts might be influenced by elevated terrain in this region. KUKI (Ukiah Municipal Airport) is a mixed bag. It displays a weak cold bias at 7AM and a strong cold bias at 3PM in the NEPAC model but displays a moderate warm bias at 7AM and a moderate cold bias at 3PM in CENCOOS. KUKI is located slightly inland (∼48 kilometers from the coast) at the base of mountainous terrain which might influence biases here.

4.1.2 Winter Season

For the winter season, CENCOOS again displays the larger errors overnight with a median bias of +1.71K at 7AM as compared to NEPAC's median bias of -0.12K. The picture at 3PM is different, with CENCOOS doing better on average for daytime high temperatures, bias of only +0.18K at 3PM, compared to NEPAC with a bias of -1.74K at 3PM(4.3). Distributions of all temperature errors for NEPAC and CENCOOS at 7AM shows that NEPAC is skewed significantly towards the left (cold errors) at both 7AM and 3PM while CENCOOS is strongly right skewed (warm errors) at 7AM and weakly right skewed at 3PM (Fig. 4.5). Table 4.3 indicates that temperature biases at 7AM in NEPAC are strongly influenced by observed cloud cover amount (-0.12 K for all and + 1.01 K for < 25%) in contrast to CENCOOS biases at 7AM which vary moderately with cloud cover (+1.71 K for all and + 2.27 K for < 25%). Radiation fog is more common in the Central Valley of California in winter than in summer and potentially this sensitivity to cloud cover could relate to representation of overnight radiation inversions and/or fog in both models.

Winter Temperature Analysis: **NEPAC 3PM**	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-1.74	-3.21	-0.37
<25% Cloud Cover	-1.87	-3.41	-0.89
<50% Cloud Cover	-1.74	-3.26	-0.45
NEPAC 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	-0.12	-2.09	1.75
<25% Cloud Cover	+1.01	-1.50	+1.98
<50% Cloud Cover	+0.70	-1.63	+1.98
CENCOOS 3PM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+0.18	-0.84	+1.43
<25% Cloud Cover	-0.01	-1.06	+0.96
<50% Cloud Cover	+0.14	-0.88	+1.05
CENCOOS 7AM	Median Bias (K)	25th Percentile (K)	75th Percentile (K)
All Cloud Cover	+1.71	-0.46	+3.27
<25% Cloud Cover	+2.27	+0.76	+3.63
<50% Cloud Cover	+1.93	+0.02	+3.70

Table 4.3: Temperature bias analysis from November 2022 to February 2023 at 7AM and 3PM for COAMPS NEPAC and COAMPS CENCOOS California model domains. Median, 25th percentile, and 75th percentile values from the temperature bias distributions are further subset by observed cloud cover amounts.

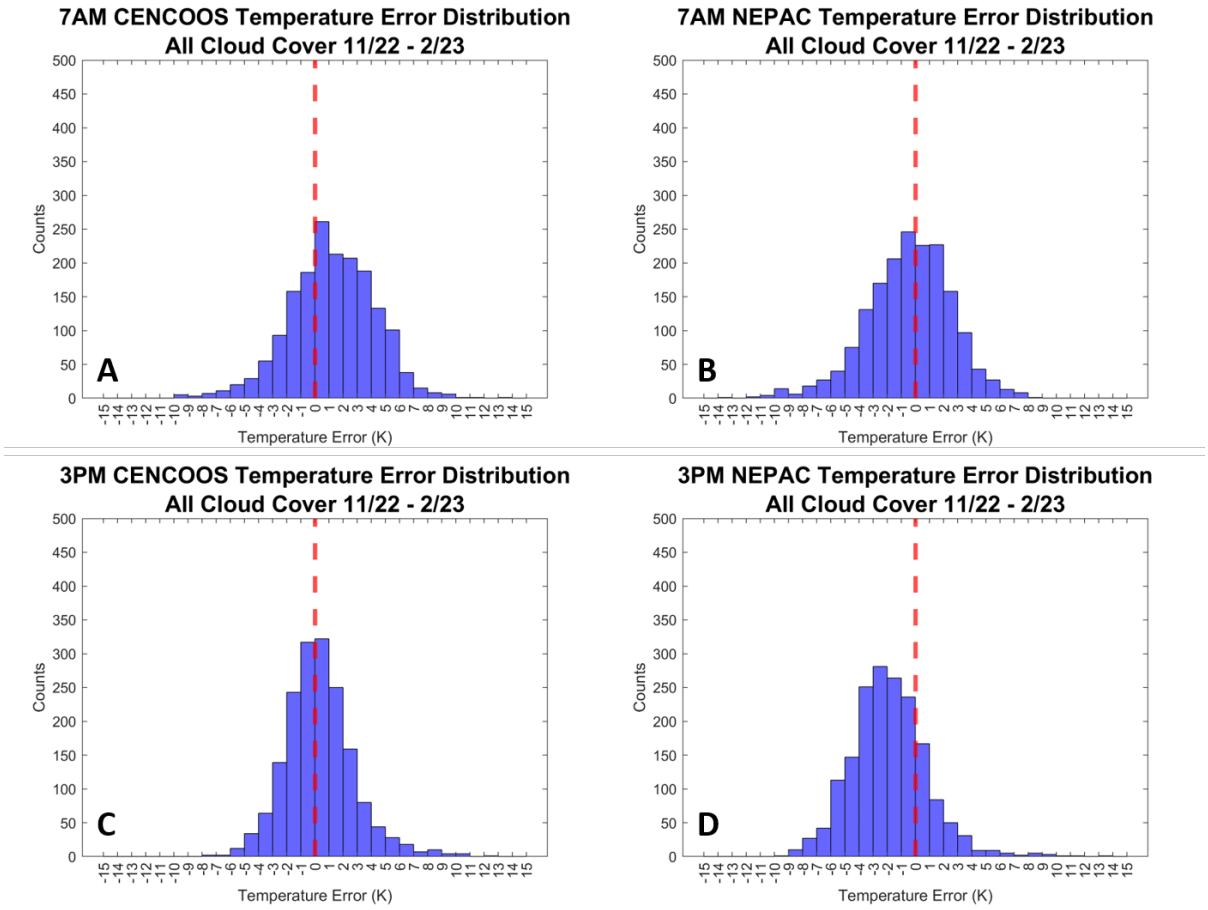

Figure 4.5: November 2022 - February 2023 distributions of temperature error at 7AM (A, B; top panel) and 3PM (C, D; bottom panel) for COAMPS NEPAC (B, D; right panel) and CENCOOS (A, C; left panel) models at a 48-hour lead time for all cloud cover conditions.

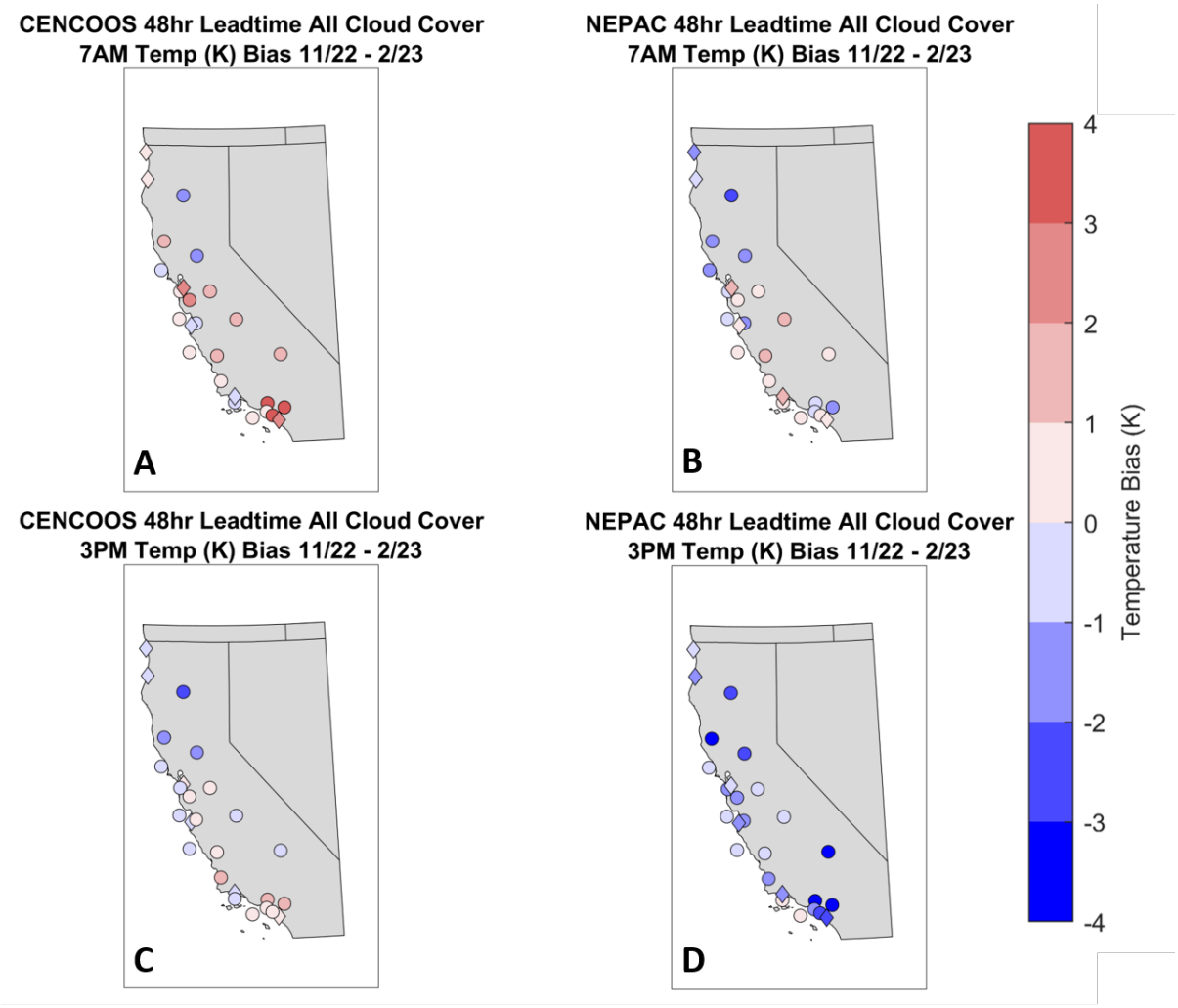

Figure 4.6: Same as Fig. 4.3 but for November 2022 - February 2023.

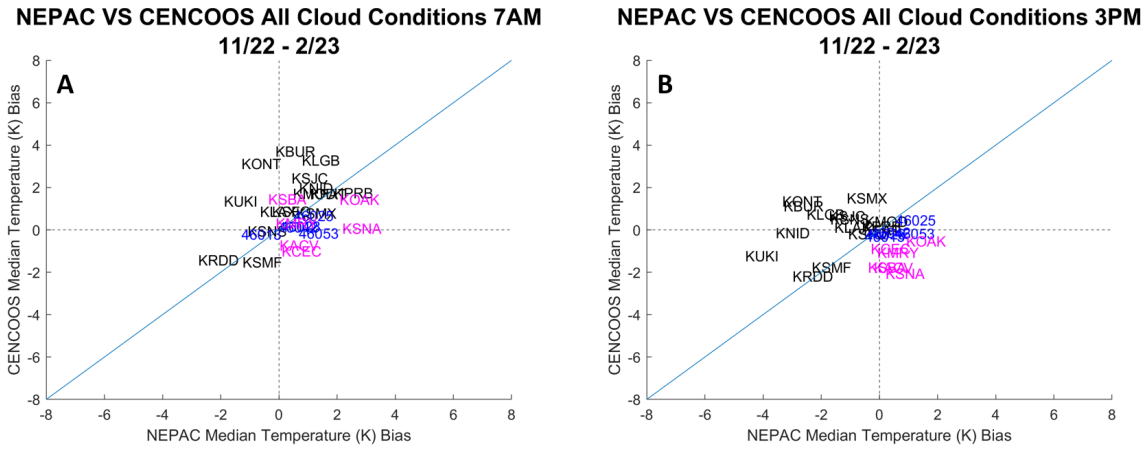

Figure 4.7: Same as Fig. 4.4 except for November 2022 - February 2023

The distribution of temperature biases across California is less consistently impacted by geographical location in the winter than in the summer (Fig. 4.6). Stations across southern, central, and coastal California were dominated by warm biases in CENCOOS at 7AM but transitioned to weak warm and cold biases across the state by 3PM. Southern California had the strongest warm biases at 7AM and maintained weaker warm biases into the afternoon which is consistent with biases noted in this region during the summer. NEPAC does not have a geographically consistent pattern for bias distribution at 7AM but transitions to cold dominated biases across the entire state by 3PM.

In comparison to summer, station median biases in winter are often smaller and tend to cluster around the 0 bias center point on the diagram (Fig. 4.7). Individual stations in southern California including KBUR (Burbank, CA), KONT (Ontario, CA) and KLGB (Long Beach, CA) have median biases >3K in CENCOOS and lower values in NEPAC. Buoy temperature biases seem to be fairly low at both 7AM and 3PM indicating a higher forecast skill for forecasting buoy temperatures in both NEPAC and CENCOOS than in forecasting inland temperatures. However, oceanic diurnal air temperature ranges are smaller than on land since sea surface temperature diurnal variations are small which can lead to more favorable forecasting conditions.

As expected, CENCOOS has a smaller range of terrain elevation differences between the model grid and observation station (-100m to +100m) compared to NEPAC (-100m to 500m) (Figs. 4.8 and 4.9). Neither CENCOOS or NEPAC show much correlation between temperature biases and elevation differences except for a weak tendency for increasing cold biases with increasing elevation difference for NEPAC 3 PM winter (Fig. 4.9D).

4.2 Wind speed and direction error frequency for California

Overall, both NEPAC and CENCOOS forecasts have infrequent large errors in wind speed and wind direction for California with the exception of a few stations. Wind speed forecast errors exceeding the TAF criteria are more common than wind direction errors (Fig. 4.10 and 4.11). Previous research by Irina Sandu (2020) suggests that errors in a model's ability to represent turbulence and friction in the boundary layer contributes to increased wind speed and wind direction forecast biases by either slowing down or shifting the forecast wind speeds and wind directions too much or too little. Another potential source of error, discussed further below, is the impact of terrain resolution on wind speed and direction forecasts.

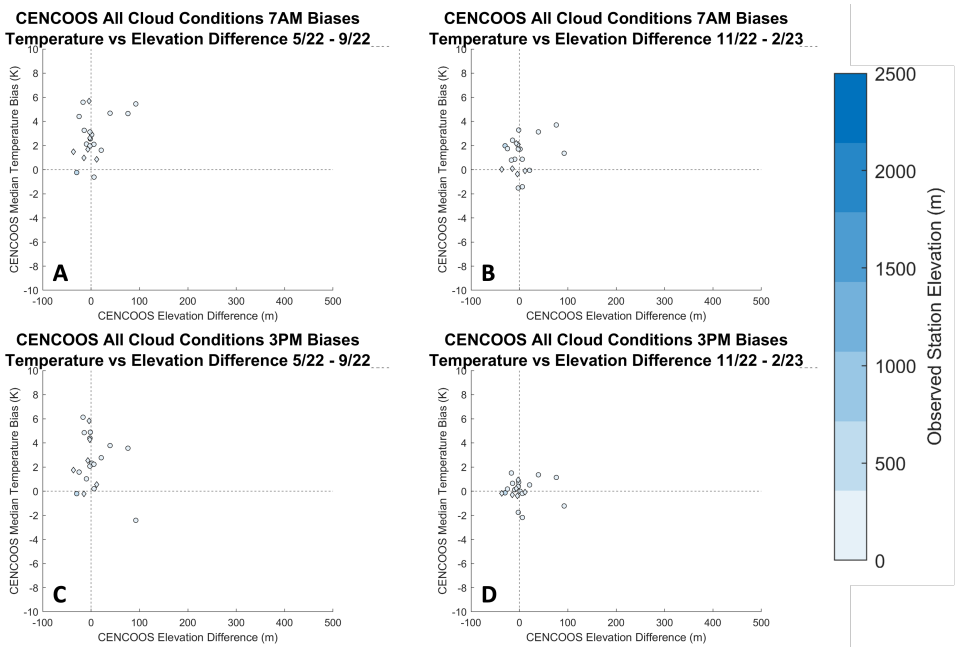

Figure 4.8: CENCOOS 48-hour lead time temperature biases under all cloud conditions with CENCOOS elevation differences from observations. Marker color represents the observation station elevation. A) 7AM 5/22 - 9/22 B) 7AM 11/22 - 2/23 C) 3PM 5/22 - 9/22 D) 3PM 11/22 - 2/23

There are more stations exceeding TAF amendment criteria more than 2% of the time in winter than in summer. For NEPAC, 10 out of 21 ASOS stations meet wind speed amendment criteria and 4 stations meet wind direction amendment criteria greater than 2% of the time from May 2022 to September 2022. For CENCOOS, 11 stations meet wind speed amendment criteria and 3 stations meet wind direction amendment criteria greater than 2 percent of the time from May 2022 to September 2022. From November 2022 to February 2023, NEPAC had 17 ASOS stations meet wind speed TAF amendment criteria and 7 stations meet wind direction TAF amendment criteria greater than 2 percent of the time. In the winter, CENCOOS had 16 ASOS stations met wind speed TAF amendment criteria and 7 stations met wind direction TAF amendment criteria greater than 2 percent of the time.

The few stations with a higher percentage of times exceeding TAF amendment criteria for wind speed and/or wind direction tend to be in coastal regions and in regions with mountainous terrain. For wind speed in the summer, the stations with the highest frequencies of times exceeding the TAF criteria are KSFO (San Francisco Airport, 8.69% in CENCOOS and 12.96% in NEPAC), KNID (China Lake Naval Station, 6.40% in CENCOOS and 5.50% in NEPAC), Buoy 46012 (Bay Area, 4.00% in CENCOOS and 7.87% in NEPAC), and

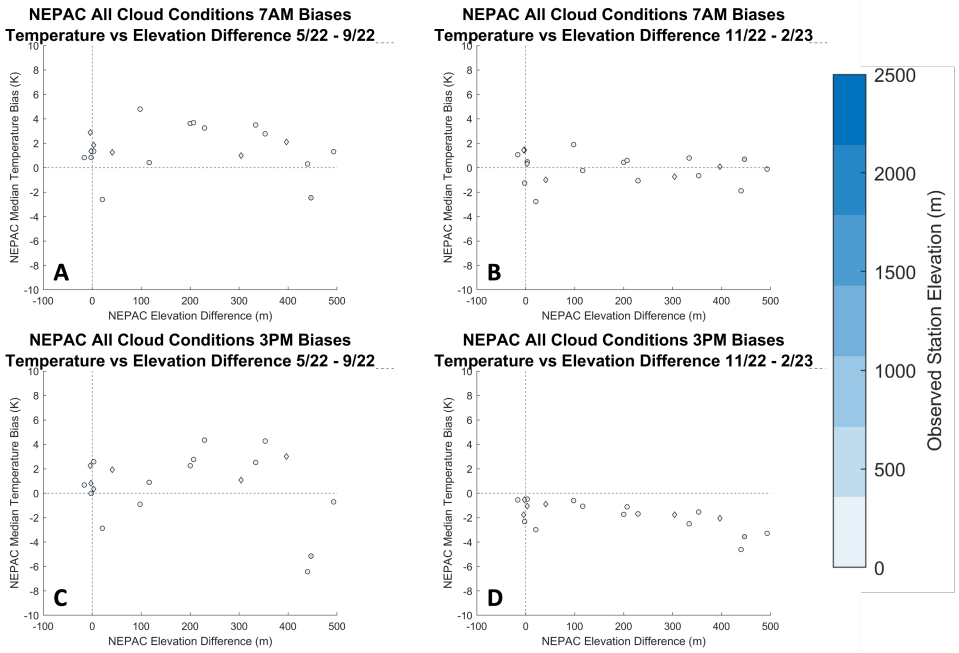

Figure 4.9: NEPAC 48-hour lead time temperature biases under all cloud conditions with NEPAC elevation differences from observations. Marker color represents the observation station elevation. A) 7AM 5/22 - 9/22 B) 7AM 11/22 - 2/23 C) 3PM 5/22 - 9/22 D) 3PM 11/22 - 2/23

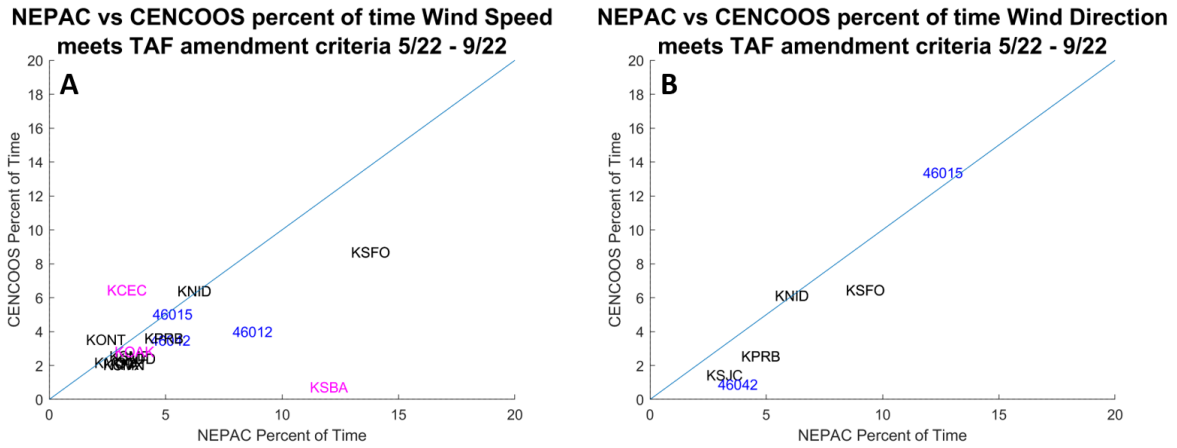

Figure 4.10: May 2022 - September 2022 scatterplot of frequency of exceeding TAF amendment criteria for wind speed (right) and wind direction (left) for NEPAC versus CENCOOS by station set against a 1-1 line. In-land stations marked in black, buoys marked in blue, NEPAC stations placed in the ocean instead of on land marked in pink. Stations with frequencies < 2% are not plotted

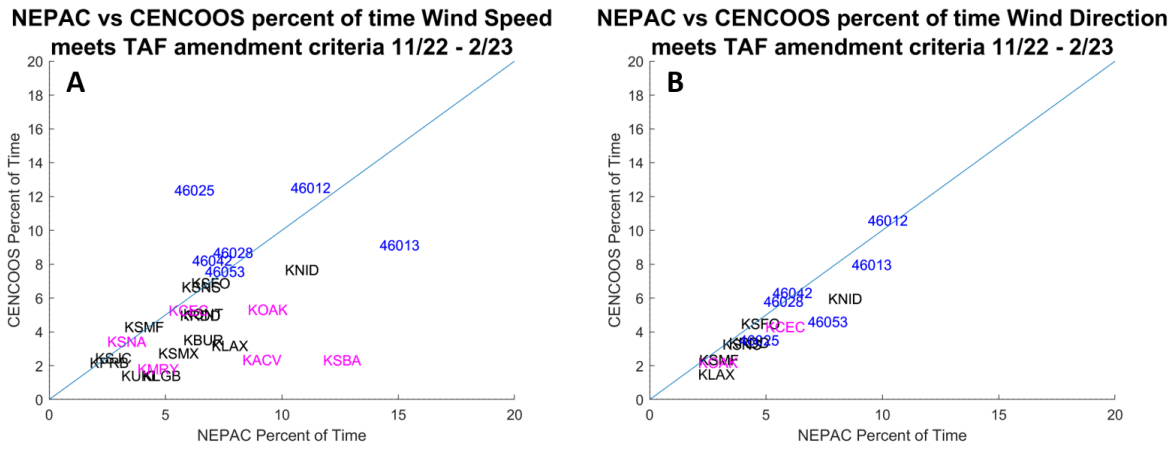

Figure 4.11: Same as Fig. 4.10 but for November 2022 - February 2023.

KPRB (Paso Robles Airport, 3.622% in CENCOOS and 4.09% in NEPAC). For inland stations, issues modeling small features in mountainous terrain might explain why model forecast wind speeds and wind directions are often incorrect. In order for a feature to be properly resolved in a model it must be at least 8 times the size of the model grid spacing (Parker (2015)). For example, stations around the San Francisco Bay (KOAK, KSFO) often have high wind speed and wind direction errors (Fig. 4.12 and Fig. 4.13). The entrance to the San Francisco Bay is approximately 5 km in length which is much smaller than NEPAC 15 km grid spacing and only slightly larger than CENCOOS 3.7 km grid spacing. In either case, the entrance to the bay is far too small to be resolved properly in either model which will negatively affect wind speed and wind direction forecasts in this region. In terms of inland biases, small changes in elevation within regions of mountainous terrain (specifically for KNID-China Lake, CA and KUKI-Ukiah, CA) could be too small to resolved properly and lead to incorrect wind speed and wind direction forecasts there.

Of the land stations with high frequencies of wind errors, only KPRB and KNID also had large biases in temperatures. Hence, while wind speed and direction forecast errors likely contribute to temperature errors in some locations, it does not appear that systematic wind speed or direction errors are major contributors to the strong temperature biases seen in many stations in California.

KCEC (Crescent City Airport) with moderately frequent wind speed biases, 6.45% in CENCOOS and 2.47% in NEPAC, is erroneously located in an ocean grid box in both NEPAC and CENCOOS. KSBA (Santa Barbara) with wind speed biases exceeding TAF criteria 0.74% of the time in CENCOOS and 11.19% of the time in NEPAC is located in an ocean grid box

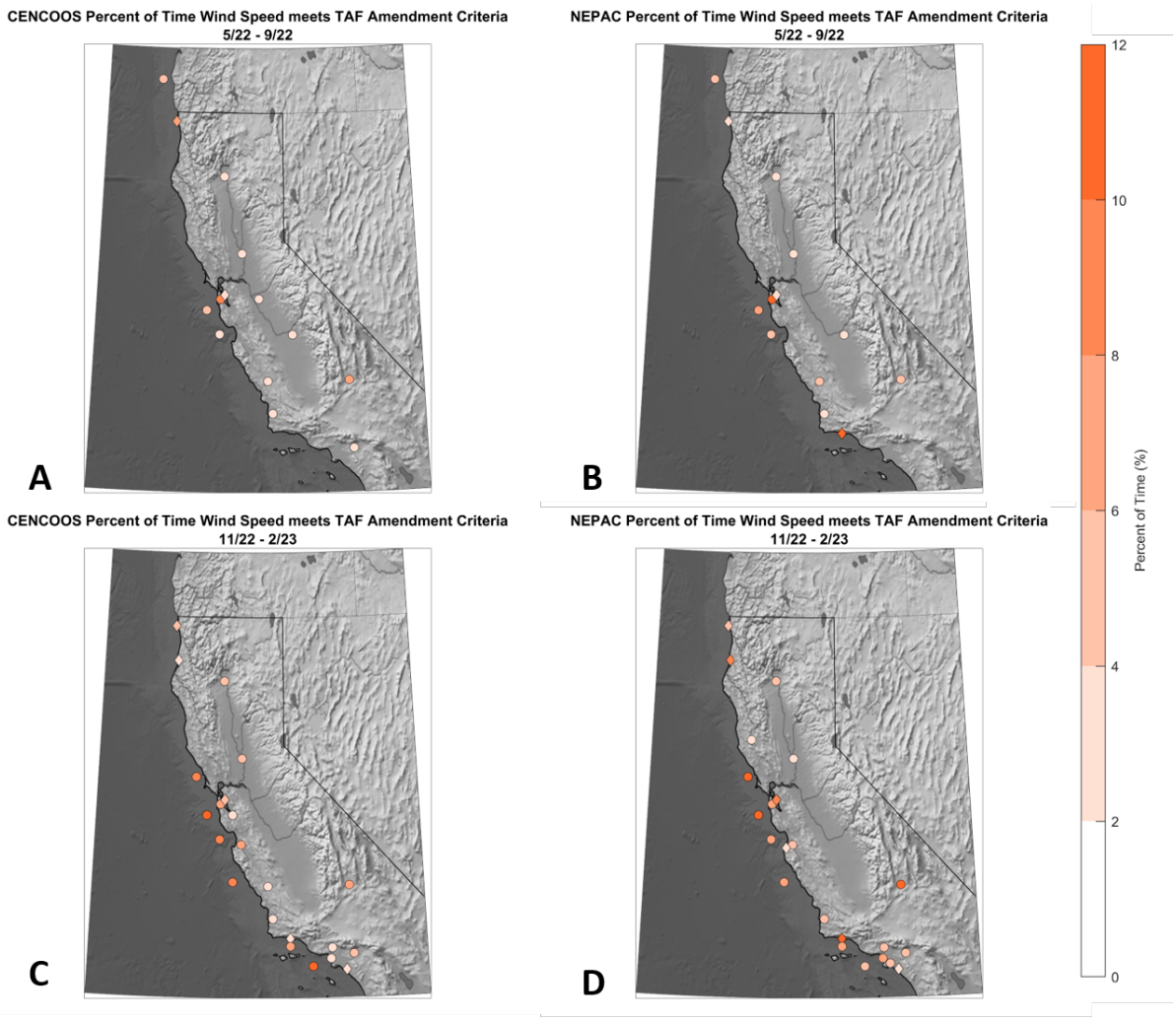

Figure 4.12: Stations plotted where forecast wind speeds meet TAF amendment criteria for wind speed (wind speed error > 5.14 m/s) more than two percent of the time. A) CENCOOS stations 5/22 - 9/22, B) NEPAC stations 5/22-9/22, C) CENCOOS stations 11/22 - 2/23, D) NEPAC stations 11/22 - 2/23. Land stations that are misclassified as ocean in land/sea mask indicated by diamond shapes.

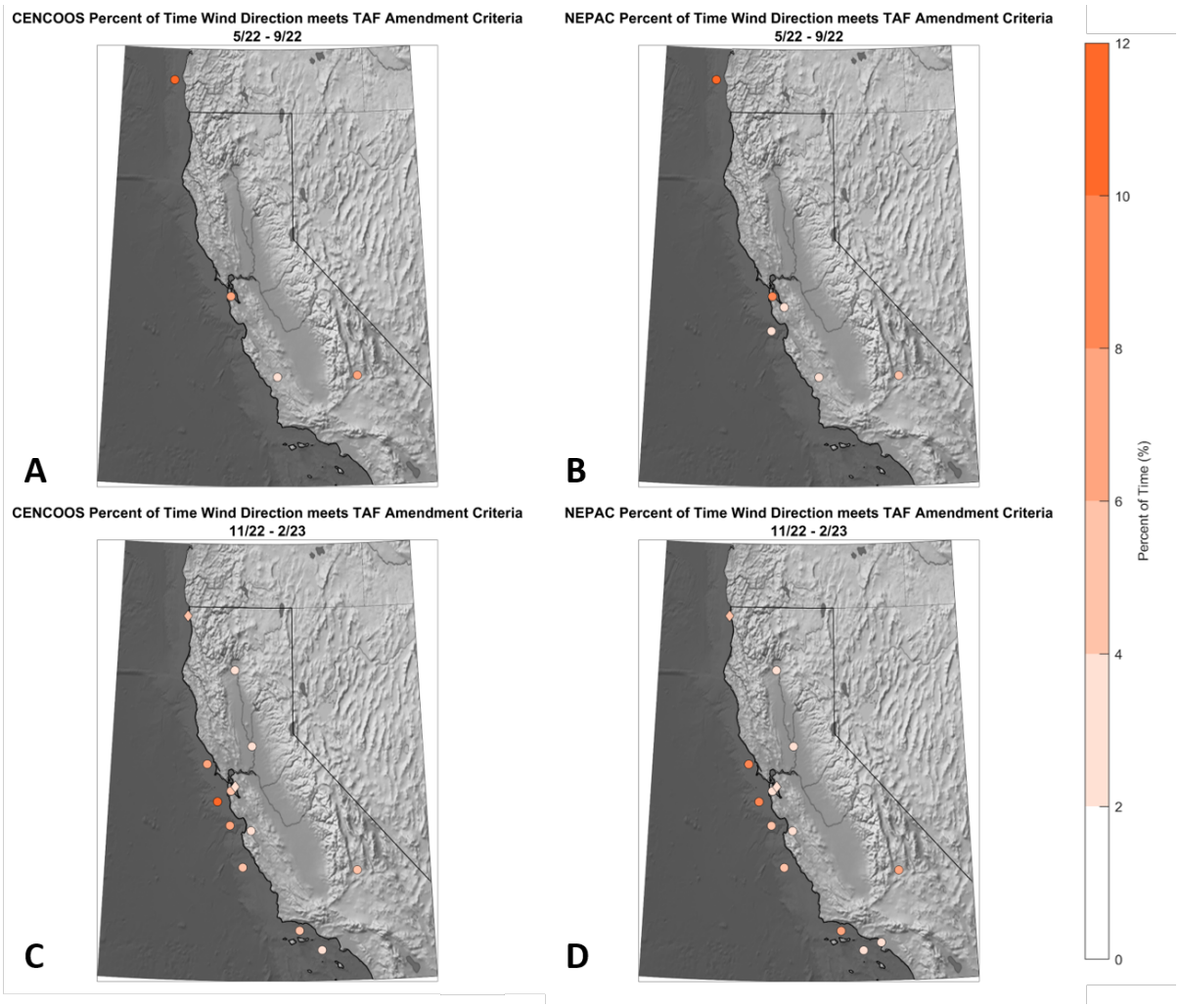

Figure 4.13: Same as Fig. 4.12 except for wind direction (direction error > 30 degrees azimuth)

for NEPAC but not for CENCOOS. These results illustrate that misclassification of surface type to ocean contributes to wind speed errors.

In the winter, buoy wind speed and wind direction errors were more frequent than in summer for Buoy 46012 (Bay Area, 12.53% in CENCOOS and 10.38% in NEPAC) and Buoy 46013 (Bay Area, 9.13% in CENCOOS and 14.20% in NEPAC). In the summer, Buoy 46015 (off the coast of Southern Oregon) has about twice as frequent wind direction errors than wind speed errors. At this buoy, wind direction exceeds TAF amendment criteria 13.37% for CENCOOS and 11.74% for NEPAC compared wind speed error frequencies $\leq 5\%$ for both models. This result for the Oregon buoy warrants further investigation since storms are more frequent off the Pacific Northwest coast in winter than in summer when atmospheric rivers occur more often (Warner and Mass 2017).

4.3 Implications

The fact that CENCOOS's finer grid has worse performance than NEPAC's coarser grid in temperature suggests that there is a model physics issue that is partially compensated for at coarser grid resolutions. The weak temperature error sensitivity to cloudiness conditions suggests the error is not in the cloud radiation package. In discussions with personnel at NRL-Monterey, it was suggested that the soil moisture input to the model, which is known to have issues, may be contributing to temperature errors. Future work could examine dewpoint errors in California between NEPAC and CENCOOS to help clarify potential sources of the temperature error within the model.

Preliminary investigation of the sea/land breeze diurnal cycle along the California coast did not indicate that it was a primary contributor to temperature or wind speed errors. There is circumstantial evidence that issues related to small gaps or changes in elevation in mountainous terrain could be contributing to errors but this is yet to be investigated in detail. Specifically, small gaps in the coastline of the San Francisco Bay are not being resolved properly by either CENCOOS or NEPAC likely leading to larger errors at KSFO. The representation of mountain/valley breezes in the model would be a partially function of terrain detail and land surface types. Stations worth further focus to examine the mountain/valley breeze issues include KNID (China Lake, CA) which is in the western portion of the Mojave Desert in southern California and Central Valley stations at KRDD (Redding, CA), KFAT (Fresno, CA) and KSMF (Sacramento, CA). As in the analysis of temperature biases for North America (Section 3.1.1), differences between model grid and weather station

observation elevations likely contribute to temperature biases in some cases but these elevation differences do not appear to be the primary source of the biases.

CHAPTER

5

CONCLUSIONS AND FUTURE WORK

5.1 Summary of Results

We analyzed forecast skill of the numerical weather prediction models COAMPS and GFS to see if they performed better or worse under different meteorological conditions. COAMPS and GFS do not use the same physical core but, like all operational weather forecast models, they have a common ancestry for their model physics. We assessed models' temperature and dewpoint biases for different observed cloud cover conditions and how frequently model wind speed and direction errors exceeded Terminal Area Forecast amendment criteria. We also examined biases for the subset of conditions when forecast and/or observed temperatures were either above the 90th percentile or below the 10th percentile of the long term climatology. In addition, we looked at both models' ability to forecast the timing of low pressure system passages and the timing and duration of precipitation events. This analysis has yielded detailed information on COAMPS and GFS forecast skill across a range of weather conditions and regions and can help identify and constrain potentially more and less important factors to target for refinements.

The key findings from the study are:

- Temperature biases:
  - In winter, in COAMPS there are regional and diurnal patterns in temperature biases > 2K with warm biases primarily in southern North America and cold biases in northern North America.
  - In summer, COAMPS diurnal temperature biases are generally smaller than in winter (summer biases typically < 2K).
  - In both models, winter temperature biases are sensitive to cloud cover variations, with larger biases for conditions with < 25% observed cloud cover than for all cloud cover conditions. In contrast, summer temperature biases are not strongly a function of observed cloudiness conditions.
  - COAMPS and GFS usually underestimate the severity of temperature events outside of the 90th and 10th climatological percentiles. Forecasts for events > 90th percentile are often too cold and forecasts for events < 10th percentile are often too warm. The magnitudes of these forecast biases station by station are usually larger than the seasonal median values.
- The overall bulk analysis distribution of forecast temperatures for 48 and 72 hour lead times shows that forecast temperature distributions are reasonably close to the observed temperature distribution. This indicates that event timing is a likely contributor to larger station by station matched valid time biases.
- COAMPS displays a regional pattern for dewpoint biases where those in the eastern US are often too moist while those in the western US are too dry. In comparison, GFS tends to be too dry in most locations in the US and Canada.
- Stations within and near mountainous terrain often have the largest errors. Larger errors in mountainous terrain are notable in temperature, low pressure passage timing, and frequency of larger wind speed/direction biases.
- Most forecast low pressure passages in COAMPS and GFS occur within +/- 2 hours of observed low pressure passage.
- The duration of forecast precipitation events is usually too long in comparison to the observed duration of precipitation events. In both models, median start time biases are > 2 hours too early and median end time biases are > 2 hours too late.

- Unexpectedly, the smaller grid spacing COAMPS model CENCOOS has larger temperature bias errors than the larger grid spacing model NEPAC for the same region for summer at 7AM and 3PM and for 7AM in winter.

Regional, time of day, and weather condition variations in the spatial patterns and magnitudes of biases would make domain-wide seasonal or monthly bias corrections problematic. Rather than brute force number crunching to reduce average root mean square error, groups using machine learning to post-process forecast model output can likely improve their outcomes with consideration of bias variations across the diurnal cycle and in different weather conditions.

In the analysis of precipitation durations, the current method to identify events has difficulties resolving short duration storms including isolated convection and fast moving squall lines. Given that shorter duration storms are more common during the summer season, we instead focus our discussion of precipitation results only on the winter season. Our analysis showed that both COAMPS and GFS overforecasted the duration precipitation events in the winter season. Lu et al. (2011) found that COAMPS overforecasted cool season precipitation amounts related to its tendency to overestimate rainfall areas when low to moderate precipitation rates (<1.1 in) were forecast. Precipitation event duration is not the same as amount but in winter when most storm systems are large in areal extent and precipitation rates are moderate to weak they are related. Examination of precipitation feature area in COAMPS and GFS was beyond the scope of this study.

California is the only region where the Navy Fleet Operations runs the COAMPS model at two different grid resolutions. This showed that despite smaller grid spacing and more detailed representation of terrain in California, CENCOOS had larger temperature biases than the coarser grid NEPAC which demonstrates that improved skill cannot be necessarily be gained simply by running a given model at finer grid resolution and is in agreement with Mass et al. (2002); Hoadley et al. (2004).

All the buoys off the coast of California had wind speed errors exceeding the TAF amendment criteria more than 5% of the time. Wind speed forecast errors are particularly relevant for naval operations. Further work is needed to examine wind forecast error frequency at buoys in other locations to determine if this issue is more general or specific to the California coast.

Both COAMPS and GFS have larger errors in several variables in the elevated complex terrain. In order for a feature to be properly resolved in a model it must be at least eight times the size of grid spacing within the model (Parker 2015). When the terrain resolution

is inadequate to adequately represent gaps in mountainous terrain and/or reduces the steepness of ridge slopes it could impact temperature, dewpoint, and winds.

Both GFS and COAMPS tended to underestimate the diurnal temperature range with warm biases at the time of the observed daily minimum temperature (overnight) and cold biases at the time of the observed daily high temperature (daytime). Our GFS results for a different winter season were similar to those in Patel et al. (2021). Massey et al. (2016) also found that WRF temperature forecasts with a 3.3 km grid spacing for the Intermountain West displayed a positive (warm) bias in the early morning and a negative (cold) bias in the afternoon.

Overall, we were surprised to see how similar biases were across multiple weather-conditioned evaluation metrics between COAMPS and GFS, two models that utilize very different sets of parameterizations (Table 2.1). GFS uses a newer set of parameterizations which are intended to improve performance but, while there are some differences depending on the metric examined, the older set of parameterizations utilized in COAMPS yield a roughly comparable performance to that of the GFS.

Similarities between COAMPS and GFS bias distributions and magnitudes suggests that there are limits to the skill of current operational models with larger grid spacings between 15 to 25 km. However, the difficulties both COAMPS and GFS have in representing the minimum and maximum of the diurnal cycle especially in conditions with lower cloud amounts were also seen in Massey et al. (2016) which assessed WRF model runs, utilizing much smaller grid spacing, in the Intermountain West. These problems in representing the diurnal cycle of temperatures in periods with low cloud cover cannot be attributed to limitations of parameterized convection. The representation of the boundary layer, particularly the inability of current forecast models to address shallow temperature inversions that are 10s of m in thickness, suggests more work is needed in this area to either increase vertical resolution to explicitly represent these fine scale features or to develop improved boundary layer parameterizations which better take shallow temperature inversions into account.

5.2 Future Work

In this study, we have assessed numerical weather prediction model biases in North America in the summer and winter. As part of funded grant work, we still need to assess model biases in the Navy COAMPS domains in Asia and Europe. Future work will involve replicating the assessments completed for North America for Asia and for Europe to see if

the biases noted by this study are consistent across similar climatic zones (e.g. California and the Mediterranean coast) and terrains (e.g. Intermountain West and the Alps).

Another area we would like to address is numerical weather prediction model skill during the Spring and Fall transition seasons. Spring and fall weather can often be more variable week to week than summer and winter which may manifest as larger errors in longer lead time forecasts.

With regards to the specific metrics used in this study, further work is needed to see if forecast precipitation event skill increases or decreases if we use different meteorological conditional subsets that precede a precipitation event (i.e. the amount of observed cloud cover, the observed wind speed). As done in Patel et al. (2021), another potential assessment is to test temperature biases in combinations with more than one type of meteorological criteria (i.e. amount of observed cloud cover and wind speed or amount of observed cloud cover and dewpoint, etc). This will help us diagnose if model biases are a combination of different factors and if certain factors become more influential when viewed in combination with other meteorological conditions.

Ongoing work at NRL-Monterey is examining and potentially refining several model initialization fields for COAMPS including the surface roughness and soil moisture (J. Doyle, personal communication). If soil moisture is modeled to be either too dry or too moist, it can directly increase errors in dewpoint. Lin et al. (2017) argues that errors in soil moisture modeling will lead to errors in the amount of moisture in the atmosphere which impacts the models ability to forecast cloud cover and precipitation (i.e. if the model initializes the soil as too dry, there will be less moisture in the air and the model is less likely to forecast clouds or precipitation).As discussed in Lu et al. (2011), previous studies have found that improving initialization of soil moisture reduced model temperature biases (Massey et al. 2016; Colle et al. 2003). Model underestimates of cloud cover, potentially due to a too dry atmosphere, can contribute to warm biases in 3PM surface air temperature. However, testing model sensitivity to soil moisture inputs would require a modeling study and is outside the scope of this research.

REFERENCES

Arguez, A., Durre, I., Applequist, S., Vose, R. S., Squires, M. F., Yin, X., Heim, R. R., and Owen, T. W. (2012). NOAA's 1981–2010 U.S. Climate Normals: An Overview. *Bulletin of the American Meteorological Society*, 93(11):1687 – 1697. Place: Boston MA, USA Publisher: American Meteorological Society.

AWS (2022). NOAA Global Forecast System (GFS) - Registry of Open Data on AWS.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., Olson, J. B., James, E. P., Dowell, D. C., Grell, G. A., Lin, H., Peckham, S. E., Smith, T. L., Moninger, W. R., Kenyon, J. S., and Manikin, G. S. (2016). A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Monthly Weather Review*, 144(4):1669–1694.

Bouallègue, Z. B., Cooper, F., Chantry, M., Düben, P., Bechtold, P., and Sandu, I. (2023). Statistical Modeling of 2-m Temperature and 10-m Wind Speed Forecast Errors. *Monthly Weather Review*, 151(4):897–911. Publisher: American Meteorological Society Section: Monthly Weather Review.

Bullock, R. G., Brown, B. G., and Fowler, T. L. (2016). Method for Object-Based Diagnostic Evaluation.

Casaretto, G., Dillon, M. E., Salio, P., Skabar, Y. G., Nesbitt, S. W., Schumacher, R. S., García, C. M., and Catalini, C. (2022). High-Resolution NWP Forecast Precipitation Comparison over Complex Terrain of the Sierras de Córdoba during RELAMPAGO-CACTI. *Weather and Forecasting*, 37(2):241–266. Publisher: American Meteorological Society Section: Weather and Forecasting.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S. (2008). Forecast verification: current status and future directions. *Meteorological Applications*, 15(1):3–18. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.52.

Center, E. P. I. (2020). 2. Technical Overview — Unified Post Processor Users Guide documentation.

Chen, S., Cummings, J., Doyle, J., Hodur, R., Holt, T., Liou, C., Liu, M., Mirin, A., Ridout, J., Schmidt, J., and others (2003). COAMPS version 3 model description: General theory and equations. *NRL Publ. NRL/PU/7500-03*, 448:145.

Cheng, W. Y. Y. and Steenburgh, W. J. (2005). Evaluation of Surface Sensible Weather Forecasts by the WRF and the Eta Models over the Western United States. *Weather and Forecasting*, 20(5):812–821. Publisher: American Meteorological Society Section: Weather and Forecasting.

Chien, F.-C., Kuo, Y.-H., and Yang, M.-J. (2002). Precipitation Forecast of MM5 in the Taiwan Area during the 1998 Mei-yu Season. *Weather and Forecasting*, 17(4):739–754. Publisher: American Meteorological Society Section: Weather and Forecasting.

Colle, B. A., Olson, J. B., and Tongue, J. S. (2003). Multiseason Verification of the MM5. Part I: Comparison with the Eta Model over the Central and Eastern United States and Impact of MM5 Resolution. *Weather and Forecasting*, 18(3):431–457. Publisher: American Meteorological Society Section: Weather and Forecasting.

Colle, B. A., Westrick, K. J., and Mass, C. F. (1999). Evaluation of MM5 and Eta-10 Precipitation Forecasts over the Pacific Northwest during the Cool Season. *Weather and Forecasting*, 14(2):137–154. Publisher: American Meteorological Society Section: Weather and Forecasting.

Department of the Air Force (2020). Air and Space Weather Operations. Technical Report Air Force Manual 15-129, Department of the Air Force.

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., Blake, B. T., Brown, J. M., Olson, J. B., Hu, M., Smirnova, T. G., Ladwig, T., Kenyon, J. S., Ahmadov, R., Turner, D. D., Duda, J. D., and Alcott, T. I. (2022). The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. *Weather and Forecasting*, 37(8):1371–1395.

Durham, B. F. (2020). Designated Mountainous Areas.

Dutra, E., Johannsen, F., and Magnusson, L. (2021). Late Spring and Summer Subseasonal Forecasts in the Northern Hemisphere Midlatitudes: Biases and Skill in the ECMWF Model. *Monthly Weather Review*, 149(8):2659–2671. Publisher: American Meteorological Society Section: Monthly Weather Review.

Evans, C., Weiss, S. J., Jirak, I. L., Dean, A. R., and Nevius, D. S. (2018). An Evaluation of Paired Regional/Convection-Allowing Forecast Vertical Thermodynamic Profiles in Warm-Season, Thunderstorm-Supporting Environments. *Weather and Forecasting*, 33(6):1547–1566. Publisher: American Meteorological Society Section: Weather and Forecasting.

Fritz, J., Yuter, S., Tomkins, L., Kennedy, R., and Miller, M. (2023). Evaluating Weather Forecasts of Winter Precipitation Start Times and End Times.

Glahn, H. R. and Lowry, D. A. (1972). The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology and Climatology*, 11(8):1203–1211. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., Alexander, C. R., Jensen, T. L., and Wolff, J. K. (2017). Seasonal Analysis of Cloud Objects in the High-Resolution

Rapid Refresh (HRRR) Model Using Object-Based Verification. *Journal of Applied Meteorology and Climatology*, 56(8):2317–2334. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Hoadley, J. L., Westrick, K., Ferguson, S. A., Goodrick, S. L., Bradshaw, L., and Werth, P. (2004). The Effect of Model Resolution in Predicting Meteorological Parameters Used in Fire Danger Rating. *Journal of Applied Meteorology and Climatology*, 43(10):1333–1347.

Hodur, R. M. (1997). The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Monthly Weather Review*, 125(7):1414–1430.

Irina Sandu, P. B. (2020). On the causes of systematic forecast biases in near-surface wind direction over the oceans.

Ji, L., Luo, Q., Ji, Y., and Zhi, X. (2021). Probabilistic Forecasting of the 500 hPa Geopotential Height over the Northern Hemisphere Using TIGGE Multi-model Ensemble Forecasts. *Atmosphere*, 12(2):253. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Johnson, R. H. (2003). Thermal Low. *Encyclopedia of Atmospheric Science*, pages 2269–2273.

Kain, J. S. (2004). The Kain–Fritsch Convective Parameterization: An Update. *Journal of Applied Meteorology and Climatology*, 43(1):170–181. Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Køltzow, M., Casati, B., Bazile, E., Haiden, T., and Valkonen, T. (2019). An NWP Model Intercomparison of Surface Weather Parameters in the European Arctic during the Year of Polar Prediction Special Observing Period Northern Hemisphere 1. *Weather and Forecasting*, 34(4):959–983. Publisher: American Meteorological Society Section: Weather and Forecasting.

Lavdas, L. G. (1997). Accuracy of National Weather Service wind-direction forecasts at Macon and Augusta, Georgia. *National Weather Digest. 22(1): 22-26.*

Li, W., Song, J., Hsu, P.-c., and Wang, Y. (2022). Evaluation of the Forecast Performance for Week-2 Winter Surface Air Temperature from the Model for Prediction Across Scales–Atmosphere (MPAS-A). *Weather and Forecasting*, 37(11):2035–2047. Publisher: American Meteorological Society Section: Weather and Forecasting.

Lin, Y., Dong, W., Zhang, M., Xie, Y., Xue, W., Huang, J., and Luo, Y. (2017). Causes of model dry and warm bias over central U.S. and impact on climate projections. *Nature Communications*, 8(1):881. Number: 1 Publisher: Nature Publishing Group.

Lu, D., White, L., Reddy, R. S., Williams, Q. L., and Croft, P. J. (2011). Multiseason evaluation of the MM5, COAMPS and WRF over southeast United States. *Meteorology and Atmospheric Physics*, 111(3):75–90.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A. (2002). DOES INCREASING HORIZONTAL RESOLUTION PRODUCE MORE SKILLFUL FORECASTS?: The Results of Two Years of Real-Time Numerical Weather Prediction over the Pacific Northwest. *Bulletin of the American Meteorological Society*, 83(3):407–430. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Massey, J. D., Steenburgh, W. J., Knievel, J. C., and Cheng, W. Y. Y. (2016). Regional Soil Moisture Biases and Their Influence on WRF Model Temperature Forecasts over the Intermountain West. *Weather and Forecasting*, 31(1):197–216. Publisher: American Meteorological Society Section: Weather and Forecasting.

National Centers for Environmental Information, N. (2021). Global Hourly - Integrated Surface Database (ISD).

NCEI (2023). National Trends. Utilized Summer and Winter Precipitation Subsets.

NCEI, I. (1971). Meteorological and oceanographic data collected from the National Data Buoy Center Coastal-Marine Automated Network (C-MAN) and moored (weather) buoys. Last Modified: 2023-05-16.

NOAA, N. I. D. I. S. (2023). California-Nevada | Drought.gov.

Parker, M. (2015). NUMERICAL MODELS | Convective Storm Modeling. In *Encyclopedia of Atmospheric Sciences*, pages 246–254. Elsevier.

Parker, W. S. (2016). Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society*, 97(9):1565–1572. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Patel, R. N., Yuter, S. E., Miller, M. A., Rhodes, S. R., Bain, L., and Peele, T. W. (2021). The Diurnal Cycle of Winter Season Temperature Errors in the Operational Global Forecast System (GFS). *Geophysical Research Letters*, 48(20):e2021GL095101.

Rowson, D. R. and Colucci, S. J. (1992). Synoptic climatology of thermal low-pressure systems over south-western north America. *International Journal of Climatology*, 12(6):529–545. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3370120602.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200097. Publisher: Royal Society.

Sobash, R. A. and Kain, J. S. (2017). Seasonal Variations in Severe Weather Forecast Skill in an Experimental Convection-Allowing Model. *Weather and Forecasting*, 32(5):1885–1902. Publisher: American Meteorological Society Section: Weather and Forecasting.

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., and Novak, D. R. (2014). Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011. *Weather and Forecasting*, 29(4):894–911. Publisher: American Meteorological Society Section: Weather and Forecasting.

Sun, S., Li, L., Zhao, B., Ma, Y., and Hu, J. (2023). Multiscale feature analysis of forecast errors of 500 hPa geopotential height for the CMA-GFS model. *Atmospheric Science Letters*, n/a(n/a):e1174. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asl.1174.

Vaittinada Ayar, P., Vrac, M., and Mailhot, A. (2021). Ensemble bias correction of climate simulations: preserving internal variability. *Scientific Reports*, 11(1):3098. Number: 1 Publisher: Nature Publishing Group.

Warner, M. D. and Mass, C. F. (2017). Changes in the Climatology, Structure, and Seasonality of Northeast Pacific Atmospheric Rivers in CMIP5 Climate Simulations. *Journal of Hydrometeorology*, 18(8):2131–2141. Publisher: American Meteorological Society Section: Journal of Hydrometeorology.

Wong, M., Romine, G., and Snyder, C. (2020). Model Improvement via Systematic Investigation of Physics Tendencies. *Monthly Weather Review*, 148(2):671–688. Publisher: American Meteorological Society Section: Monthly Weather Review.

APPENDIX

APPENDIX

A

SUPPLEMENTAL MATERIALS

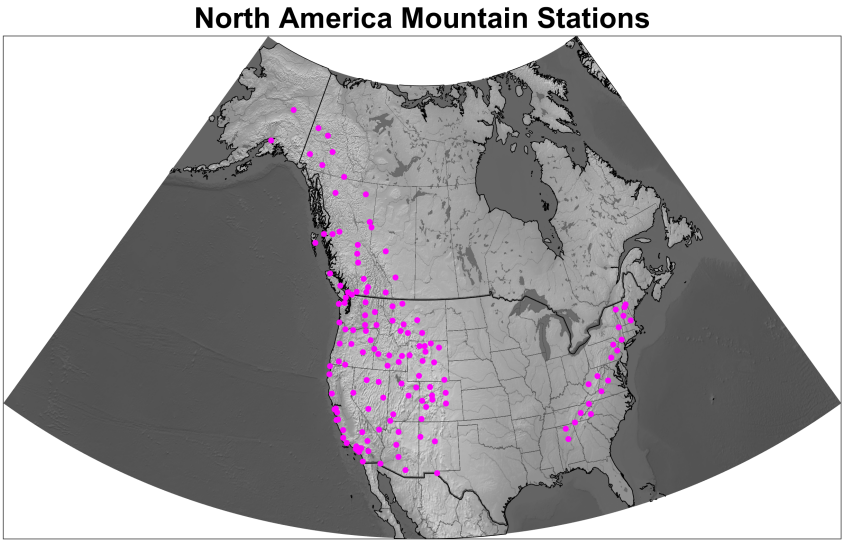

Figure A.1: Map of ASOS stations considered part of mountainous terrain per Federal Aviation Administration guidelines (Durham 2020).

Figure A.2: Model elevation differences for COAMPS (left) and GFS (right).

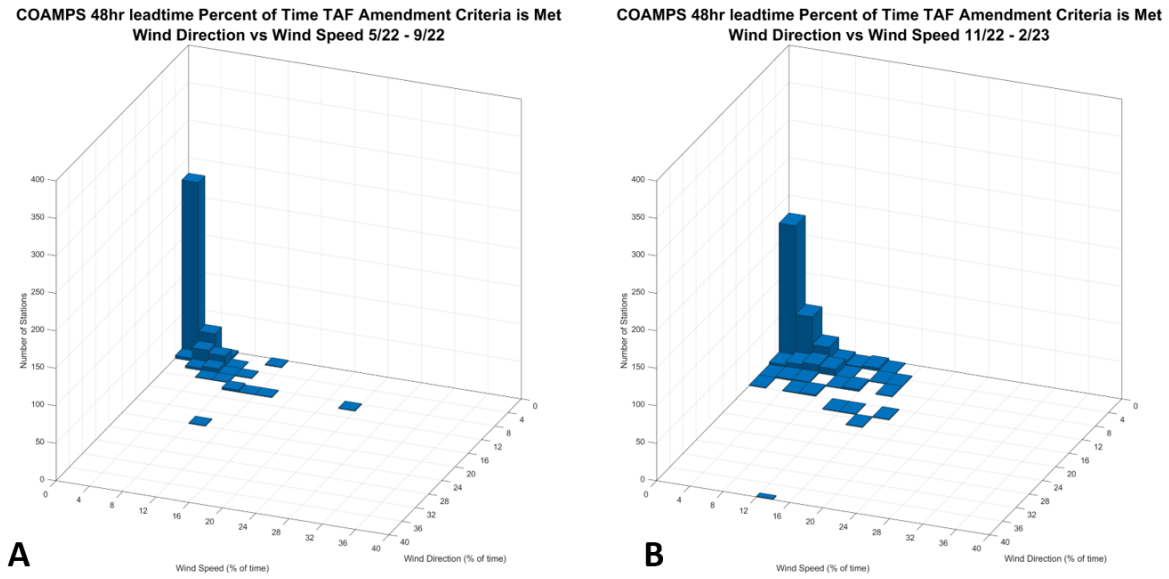

Figure A.3: Percent of time that COAMPS forecast wind speeds meet TAF amendment criteria vs the percent of time that COAMPS forecast wind directions meet TAF amendment criteria. Z axis is the number of stations that meet criteria for that wind speed and wind direction percent category (both the x and y axes are in intervals of 2). The y axis is the percent of time wind speed meets TAF amendment criteria. The x axis is the percent of time wind direction meets TAF amendment criteria. A) COAMPS 5/22 - 9/22 B) COAMPS 11/22 - 2/23

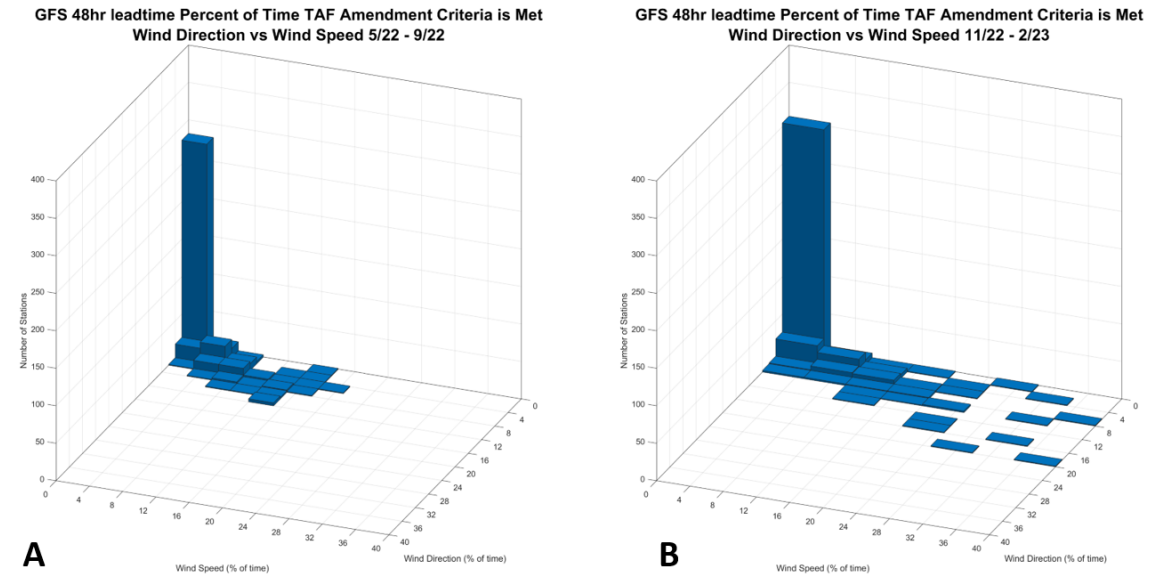

Figure A.4: Same as A.3 but for GFS. A) GFS 5/22 - 9/22 B) GFS 11/22 - 2/23

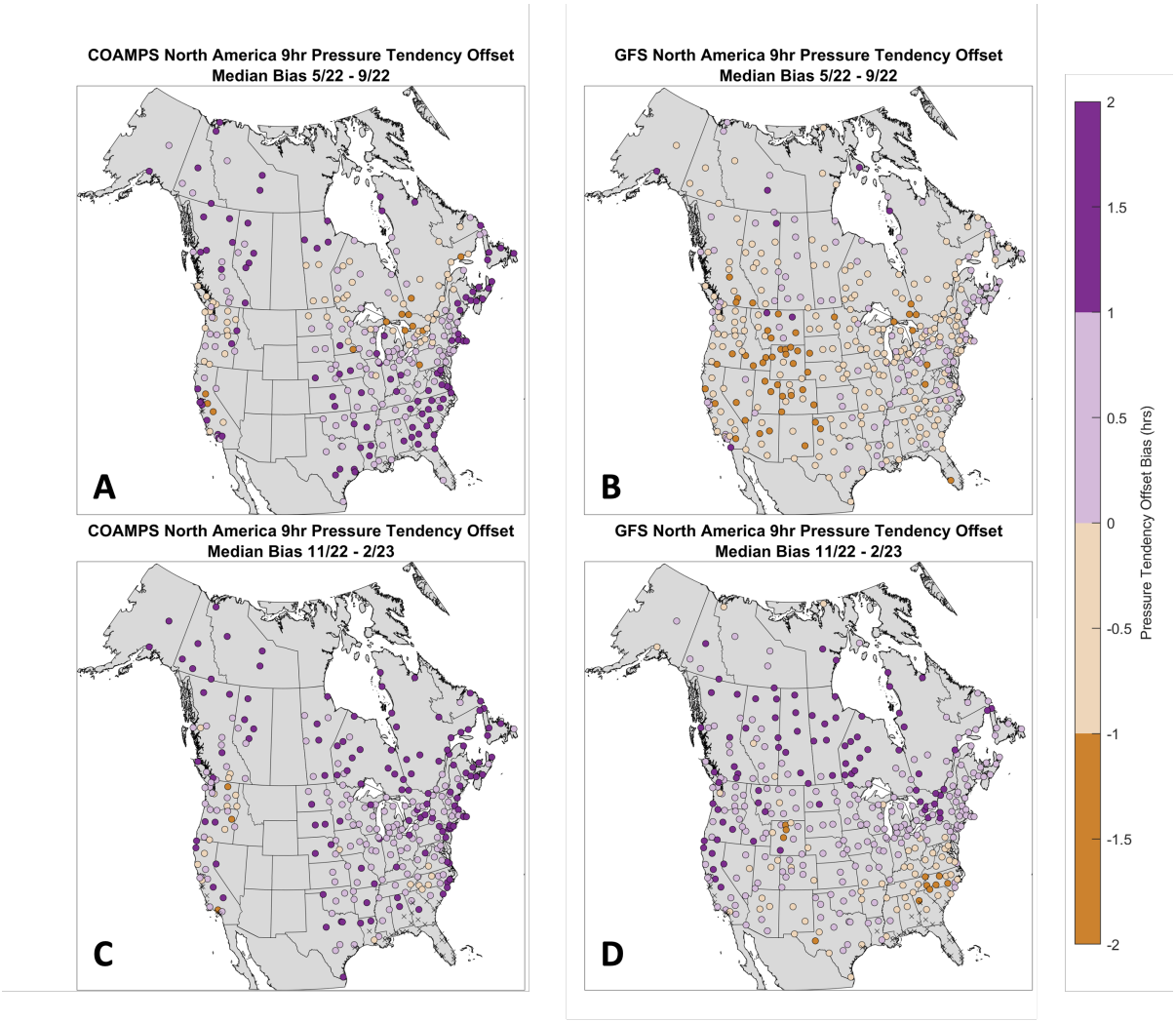

Figure A.5: Median low pressure passage model timing biases using the 9 hour pressure tendency at each station for COAMPS (left) and GFS (right) from 5/22 - 9/22 and 11/22 - 2/23. Positive indicates the model forecasts low pressure systems to arrive too late and negative values indicates the model forecasts low pressure systems to arrive too early. As in figures 3.27 and A.6, biases are not studied at stations marked with a black 'X'. A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23.

Figure A.6: Number of forecast low pressure passages per week across North America by station for COAMPS (left column) and GFS (right column). Pink X's indicate stations where fewer than 12 forecast low pressure passages were paired with observed low pressure passages. Biases produced at these stations are considered non-representative and were excluded from bias analysis. A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23

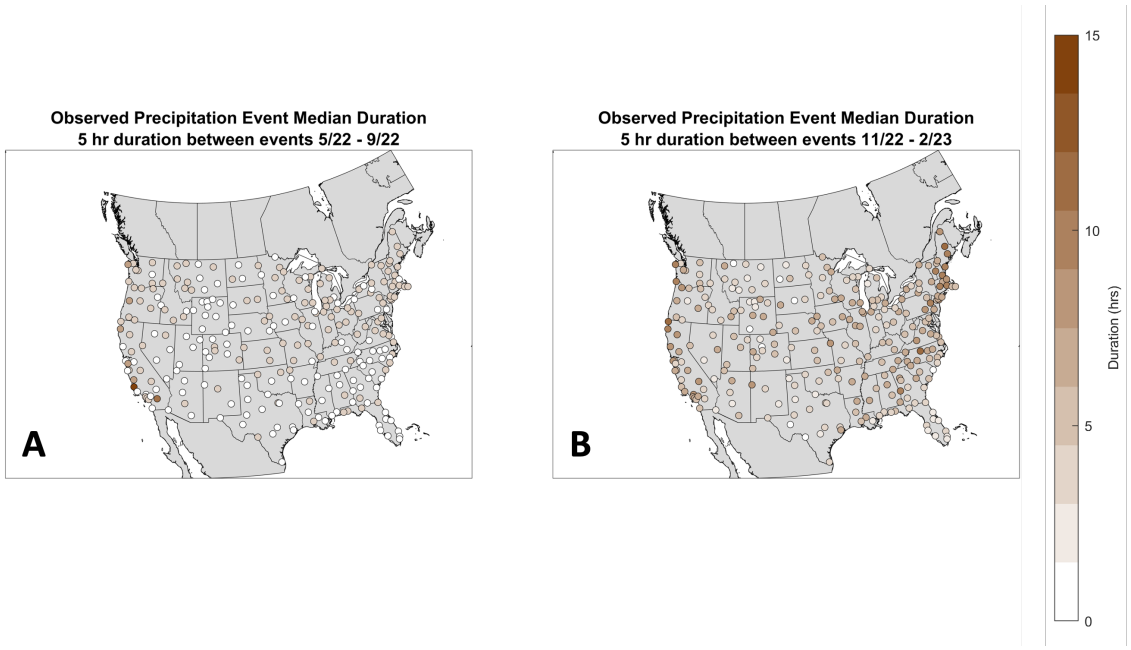

Figure A.7: Observed event durations of all paired and unpaired events at each station. A) 5/22 - 9/22 B) 11/22 - 2/23.

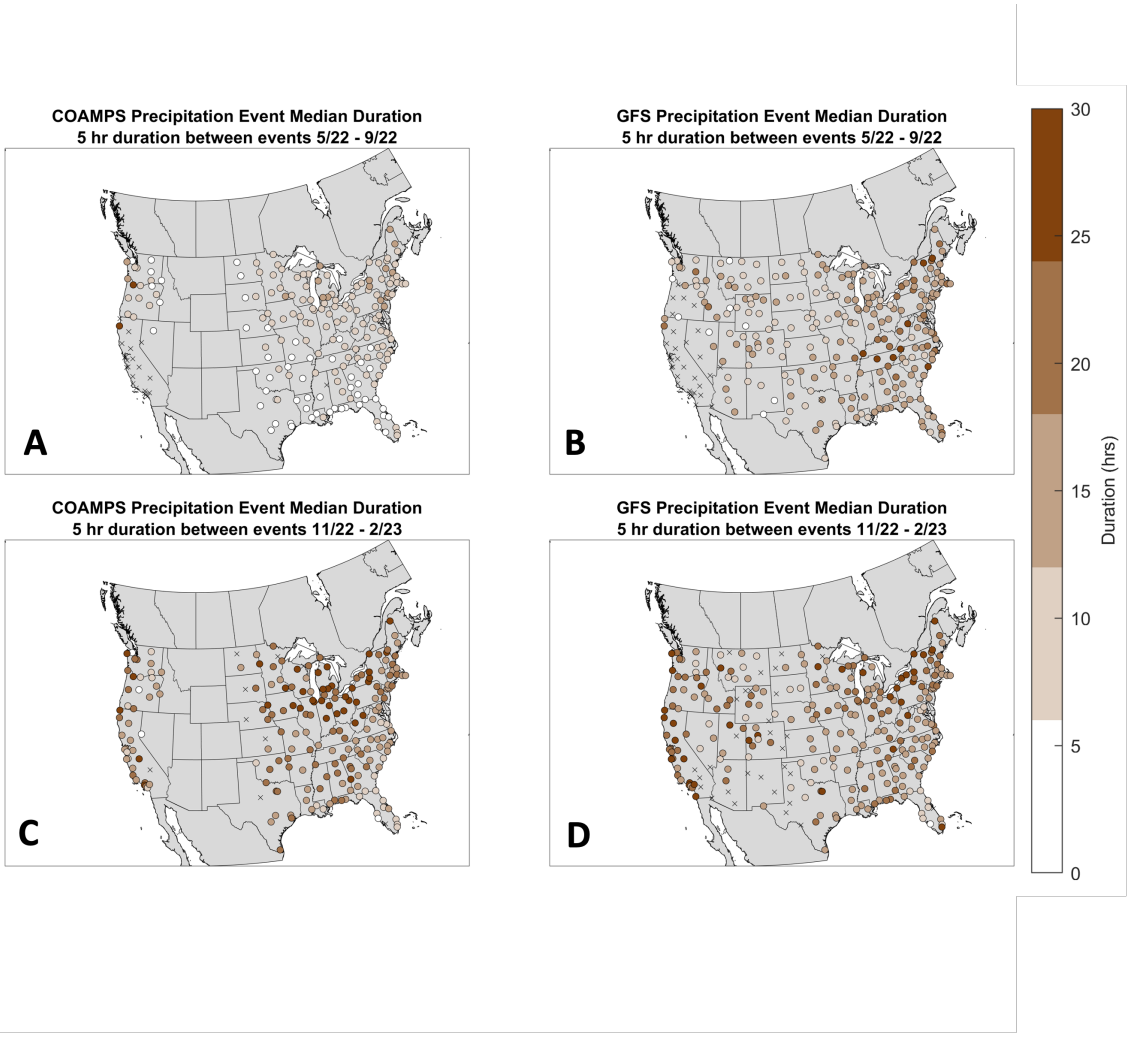

Figure A.8: Paired model event durations for 48-hour forecasts on a station by station analysis. Stations with less than ten paired precipitation events are marked with a black 'X' as not enough paired precipitation events in place of a precipitation event bias. A) COAMPS 5/22 - 9/22 B) GFS 5/22 - 9/22 C) COAMPS 11/22 - 2/23 D) GFS 11/22 - 2/23

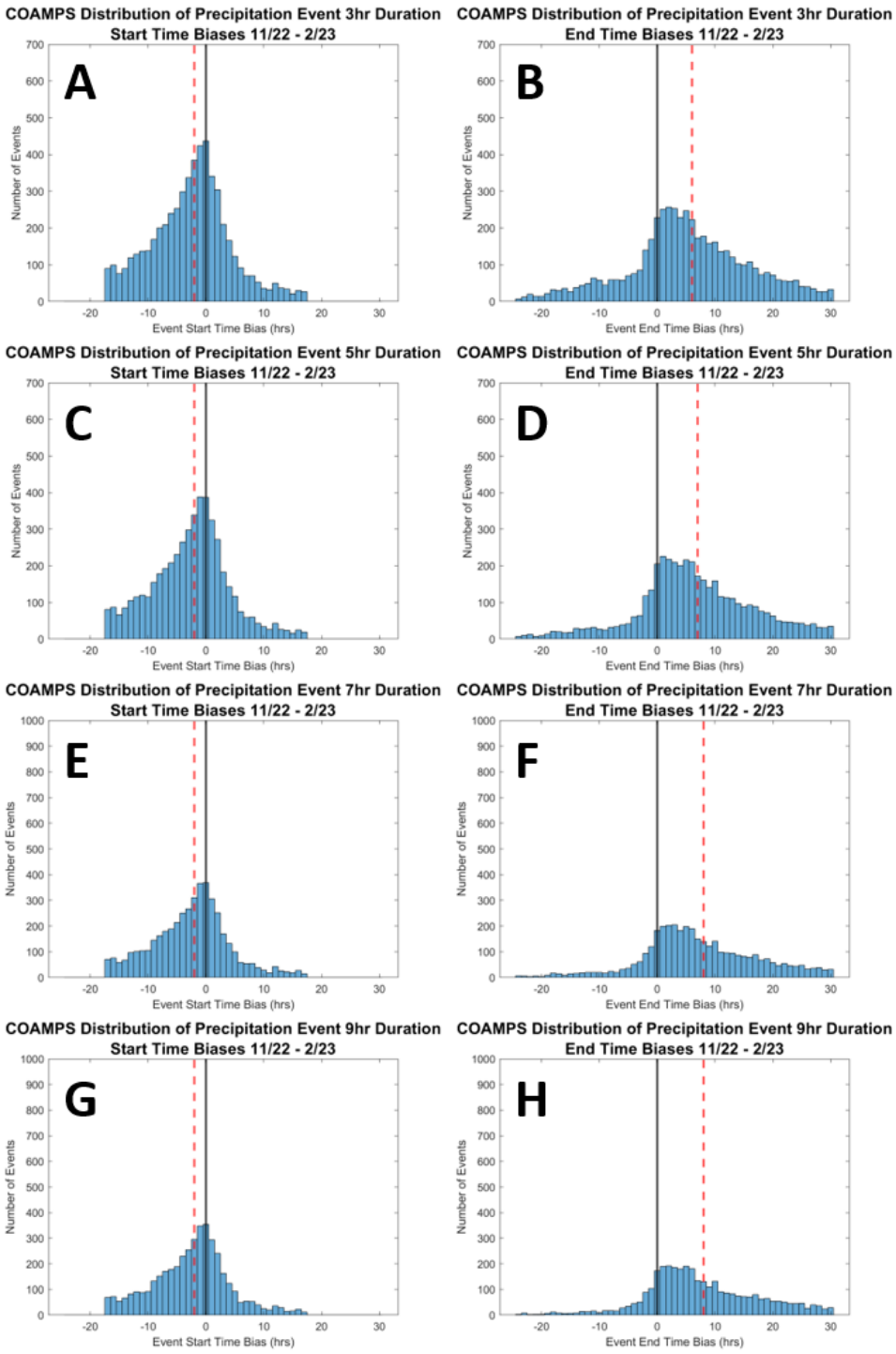

Figure A.9: COAMPS November 2022 - February 2023 precipitation start and end time sensitivity tests with a pairing window of +/- 18 hours with gaps of less than 3 hours, 5 hours, 7 hours, and 9 hours between events

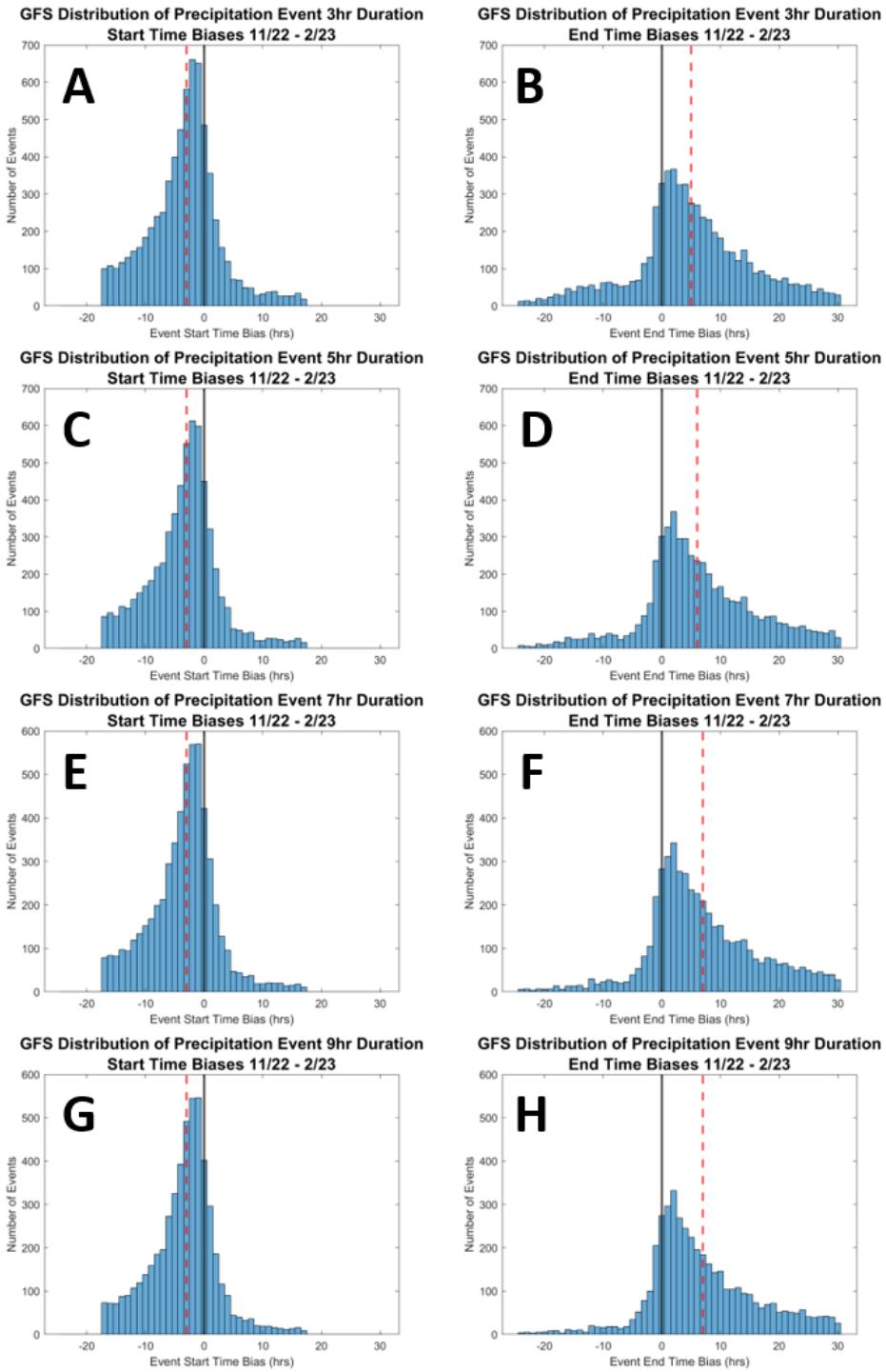

Figure A.10: GFS November 2022 - February 2023 precipitation start and end time sensitivity tests with a pairing window of +/- 18 hours with gaps of less than 3 hours, 5 hours, 7 hours, and 9 hours between events

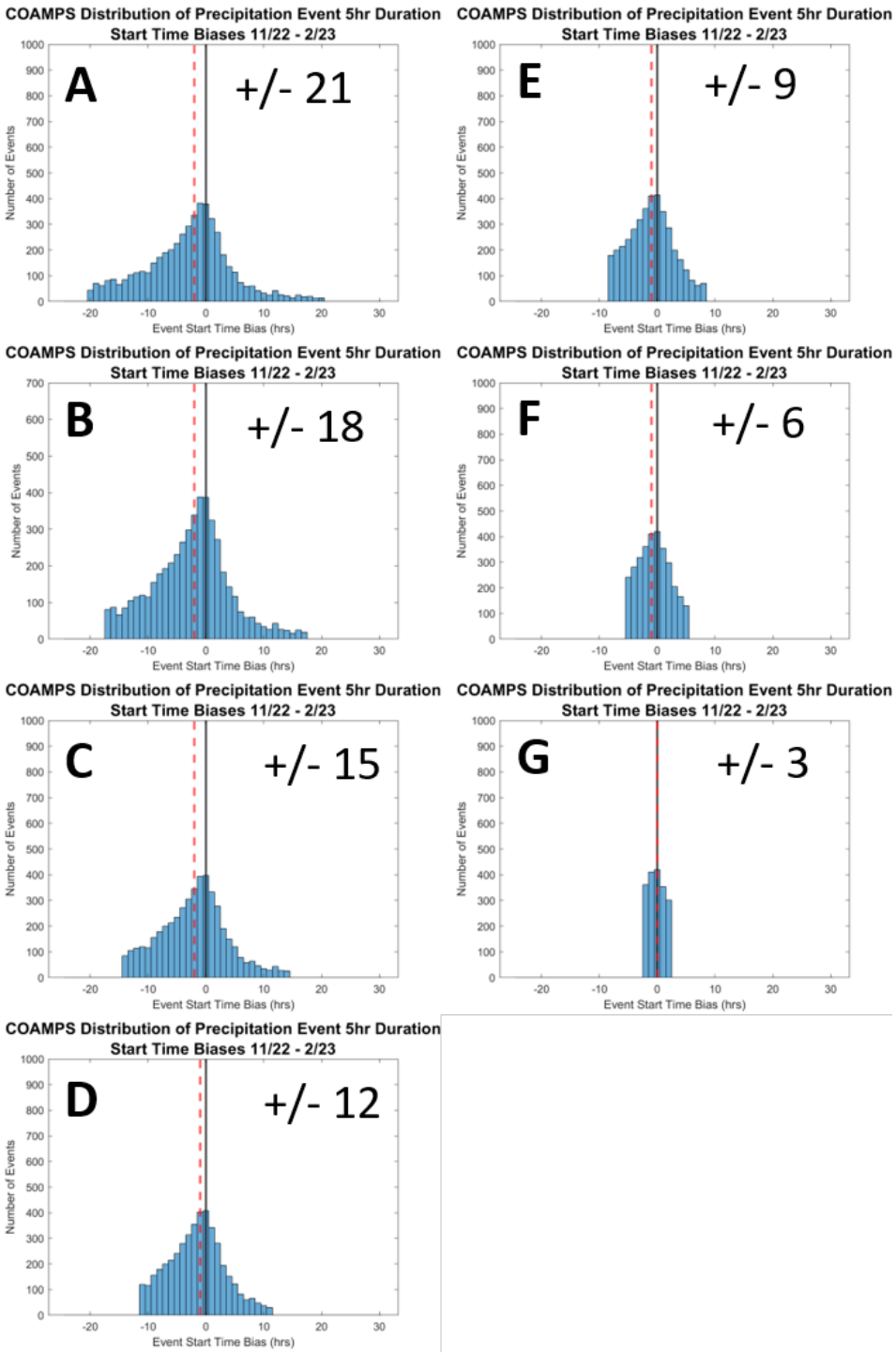

Figure A.11: COAMPS start biases for November 2022 - February 2023 for all pairing windows (+/- 21 hours, +/- 18 hours, +/- 15 hours, +/- 12 hours, +/- 9 hours, +/- 6 hours, +/- 3 hours) with <5 hour gap between precipitation events.

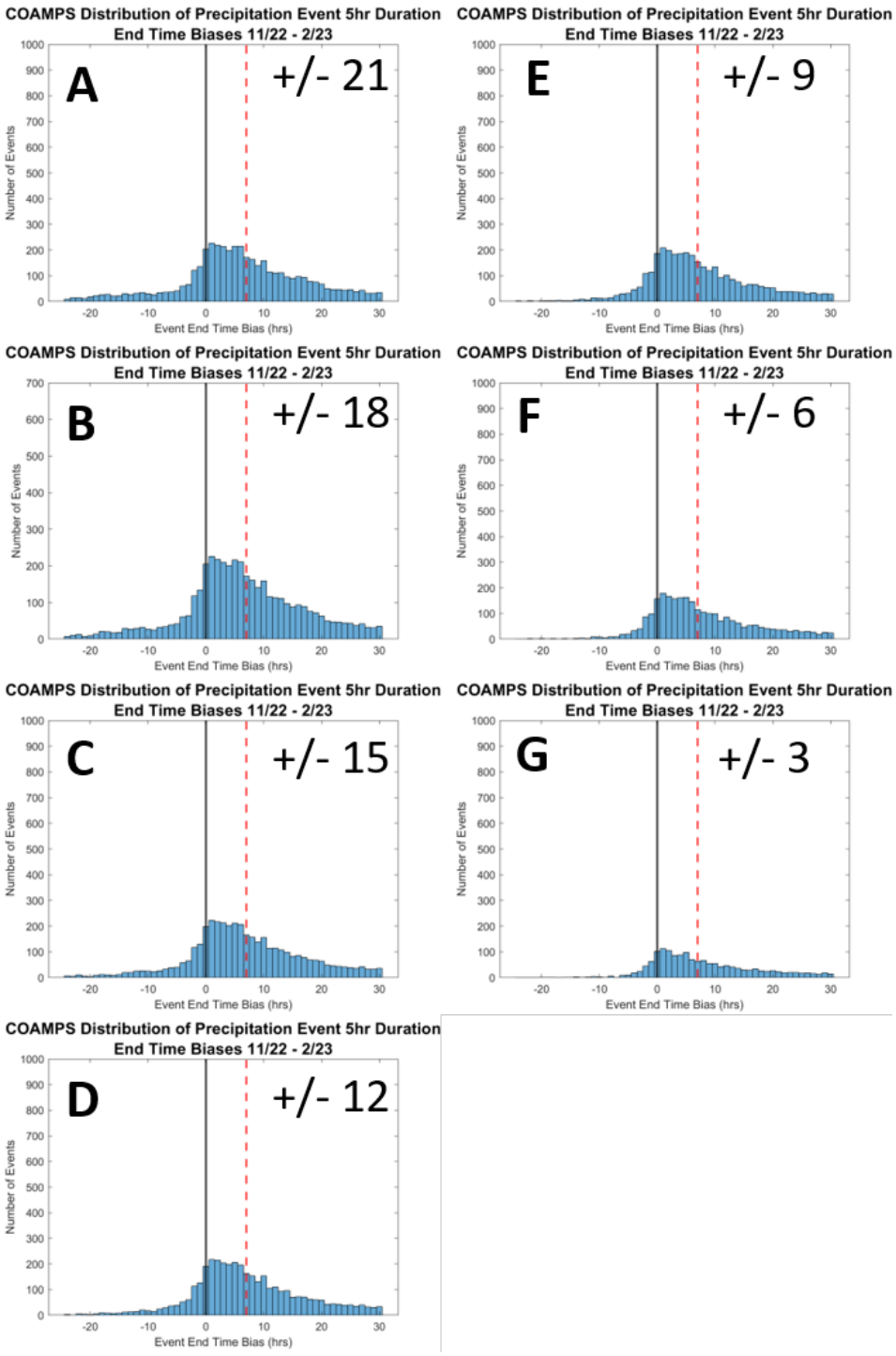

Figure A.12: COAMPS end biases for November 2022 - February 2023 for all pairing windows (+/- 21 hours, +/- 18 hours, +/- 15 hours, +/- 12 hours, +/- 9 hours, +/- 6 hours, +/- 3 hours) with <5 hour gap between precipitation events.

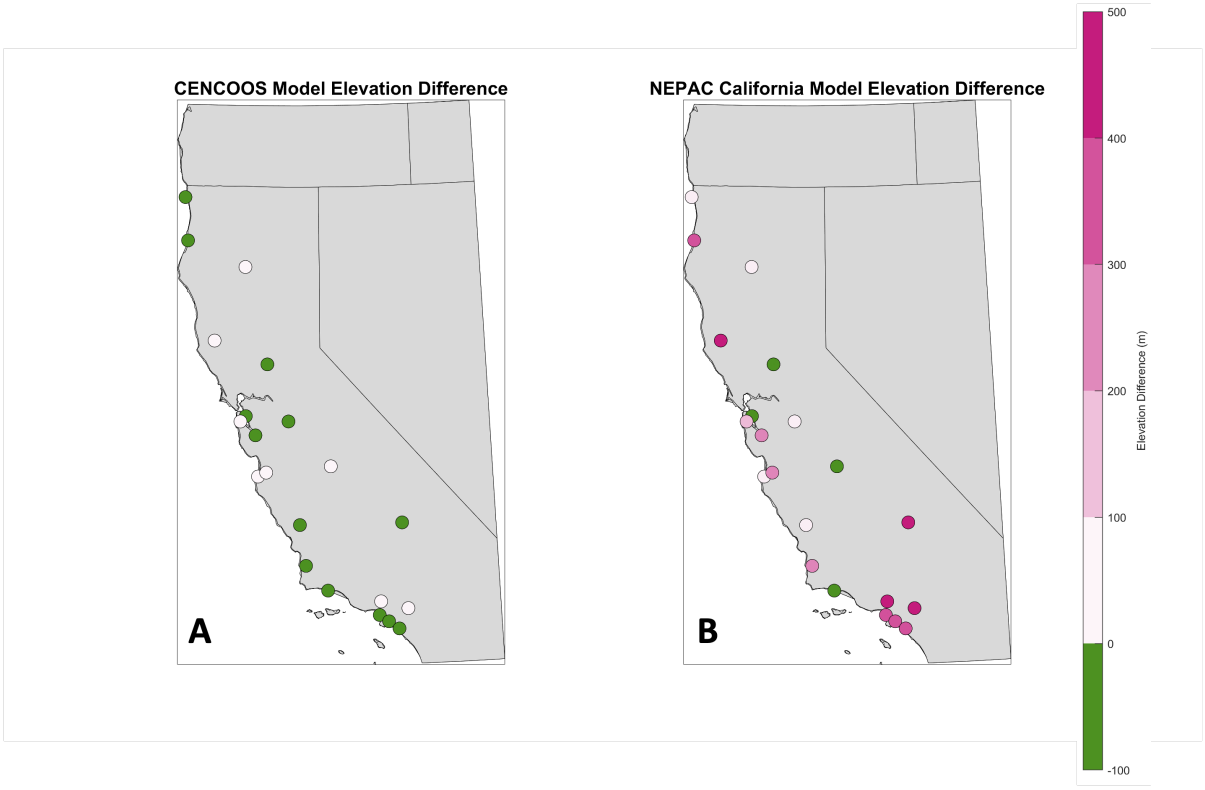

Figure A.13: CENCOOS and NEPAC model elevation errors for land stations in California. Positive bias indicates model heights too high, negative bias indicates model heights too low. A) CENCOOS B) NEPAC.