Verification Metrics for National Center for Environmental Prediction (NCEP) Models

Bibliography

Jamie Roberts, Librarian, NOAA Central Library

NCRL subject guide 2019-04
https://doi.org/10.25923/8e9n-9n28

May 2019

U.S. Department of Commerce
National Oceanic and Atmospheric Administration
Office of Oceanic and Atmospheric Research
NOAA Central Library – Silver Spring, Maryland

Table of Contents

Background & Scope

The National Weather Service (NWS) has created multiple Numerical Weather Prediction (NWP) models. The purpose of an NWP model can be inferred from various measurable qualities. The goal of this bibliography is to explore how seven NWP models are assessed for quality and performance. It is organized into eight sections; seven sections are dedicated to the evaluation of a specific NWS weather model, the eighth section explores the use of Observing System Experiment (OSE) & Observing System Simulation Experiments

Section I – Aviation Model (AVN)

Section one is intended to give an overview of verification metrics used to assess the NCEP Aviation Model.

Section II – Global Forecast System (GFS)

Section two is intended to give an overview of verification metrics used to assess the NCEP Global Forecast System.

Section III – ETA Coordinate Model (ETA)

Section three is intended to give an overview of verification metrics used to assess the NCEP ETA Coordinate Model

Section IV – Rapid Update Cycle Model (RUC)

Section four is intended to give an overview of verification metrics used to assess the NCEP Rapid Update Cycle Model.

Section V – North American Mesoscale Forecast System (NAM)

Section five is intended to give an overview of verification metrics used to assess the NCEP North American Mesoscale Forecast System Model.

Section VI – Rapid Refresh Model (RAP)

Section six is intended to give an overview of verification metrics used to assess the NCEP Rapid Refresh Model

Section VII – High Resolution Rapid Refresh Model (HRRR)

Section seven is intended to give an overview of verification metrics used to assess the NCEP High Resolution Rapid Refresh Model.

Section VIII – Observing System Experiment (OSE) & Observing System Simulation Experiments (OSSE)

Section seven is intended to give of an overview of Observing System Experiment and Observing System Simulation Experiments in weather models.

Sources Reviewed

The following databases were used to identify sources: Clarivate's Web of Science: Science Citation Index Expanded, BioOne Complete, Econ-Lit Full Text, JSTOR, ProQuest's Meteorological & Geophysical Abstracts, National Science Digital Laboratory. Only English language materials were included and there

was no date range specification. Though many citations apply to more than one model, each appears only once, under the model or metric search it was discovered with.

Section I: Aviation Model (AVN)

Charba, J. P., Reynolds, D. W., McDonald, B. E., & Carter, G. M. (2003). Comparative Verification of Recent Quantitative Precipitation Forecasts in the National Weather Service: A Simple Approach for Scoring Forecast Accuracy. Weather and Forecasting, 18(2), 161-183 https://doi.org/10.1175/1520-0434(2003)018<0161:CVORQP>2.0.CO;2

Comparative verification of operational 6-h quantitative precipitation forecast (QPF) products used for stream-flow models run at National Weather Service (NWS) River Forecast Centers (RFCs) is presented. The QPF products include 1) national guidance produced by operational numerical weather prediction (NWP) models run at the National Centers for Environmental Prediction (NCEP), 2) guidance produced by forecasters at the Hydrometeorological Prediction Center (HPC) of NCEP for the conterminous United States, 3) local forecasts produced by forecasters at NWS Weather Forecast Offices (WFOs), and 4) the final QPF product for multi-WFO areas prepared by forecasters at RFCs. A major component of the study was development of a simple scoring methodology to indicate the relative accuracy of the various QPF products for NWS managers and possibly hydrologic users. The method is based on mean absolute error (MAE) and bias scores for continuous precipitation amounts grouped into mutually exclusive intervals. The grouping (stratification) was conducted on the basis of observed precipitation, which is customary, and also forecast precipitation. For ranking overall accuracy of each QPF product, the MAE for the two stratifications was objectively combined. The combined MAE could be particularly useful when the accuracy rankings for the individual stratifications are not consistent. MAE and bias scores from the comparative verification of 6-h QPF products during the 1998/99 cool season in the eastern United States for day 1 (0-24-h period) indicated that the HPC guidance performed slightly better than corresponding products issued by WFOs and RFCs. Nevertheless, the HPC product was only marginally better than the best-performing NCEP NWP model for QPF in the eastern United States, the Aviation (AVN) Model. In the western United States during the 1999/2000 cool season, the WFOs improved on the HPC guidance for day 1 but not for day 2 or day 3 (24-48- and 48-72-h periods, respectively). Also, both of these human QPF products improved on the AVN Model on day 1, but by day 3 neither did. These findings contributed to changes in the NWS QPF process for hydrologic model input.

Ellrod, G. P., & Knapp, D. I. (1992). An Objective Clear-Air Turbulence Forecasting Technique: Verification and Operational Use. Weather and Forecasting, 7(1), 150-165 https://doi.org/10.1175/1520-0434(1992)007%3C0150:AOCATF%3E2.0.CO;2

An objective technique for forecasting clear-air turbulence (CAT) is described. An index is calculated based on the product of horizontal deformation and vertical wind shear derived from numerical model forecast winds aloft. The forecast technique has been evaluated and is now in operational use at two forecast centers with international aviation responsibilities: the National Meteorological Center (NMC) in Washington, D.C., and the Air Force Global Weather Central (AFGWC) in Omaha, Nebraska. The index is also an operational forecast tool at the Canadian Atmospheric Environment Service (AES), and the National Aviation Weather Advisory Unit (NAWAU) in Kansas City, Missouri, both responsible for domestic aviation forecasts. The AFGWC index also includes horizontal convergence in its calculation. Thresholds were selected empirically by comparing index values with the location and intensity of

observed CAT. Verification indicates that the index is quite reliable. The probability of detection (POD) varied from 70%-84%. False-alarm ratios (FAR) ranged from a low of 22% for the NMC aviation model to more than 40% for the AFGWC global model. An average threat score of 0.17 was calculated for the aviation model 24-h forecast. The operational capabilities of the NMC and AFGWC indices are compared in two CAT episodes that differ in synoptic-scale conditions and times of the year.

Fitzpatrick, P., Knaff, J., Landsea, C., & Finley, S. (1995). Documentation of a Systematic Bias in the Aviation Models Forecast of the Atlantic Tropical Upper-Tropospheric Trough - Implications for Tropical Cyclone Forecasting. Weather and Forecasting, 10(2), 433-446 https://doi.org/10.1175/1520-0434(1995)010<0433:DOASBI>2.0.CO;2

This study uncovers what appears to be a systematic bias in the National Meteorological Center's aviation (AVN) model at 200 mb over the Caribbean Sea. In general, the 48-h forecast in the vicinity of the Tropical Upper Tropospheric Trough (TUTT) underpredicts the magnitude of the westerly 200-mb winds on the order of 5-10 m s(-1). This unrealistic weakening of the TUTT and associated cold lows by the AVN results in erroneous values of the vertical (850-200 mb) wind shear. These systematic errors are in the same order of magnitude as the minimum shear threshold for tropical cyclone genesis and development. Thus, 48-h tropical cyclone formation and intensity forecasts based upon the AVN model are often incorrect in the vicinity of the TUTT. Knowing the correct future upper-wind regime is also cRUCial for track forecasting of more intense tropical cyclones, especially in cases of recurvature. It is shown that simple persistence or climatology of the 200-mb winds south of a TUTT axis is superior to the AVN model's 48-h forecast. Until this bias in the AVN is successfully removed, the tropical cyclone forecaster for the Atlantic basin should be aware of this systematic error and make subjective changes in his/her forecasts. For 200-mb west winds greater than or equal to 10 m s(-1), forecasts based on persistence are best, while for west winds less than 10 m s(-1), half climatology and half persistence is the preferable predictor. If the TUTT is weak such that 200-mb easterly winds occur, climatology tends to be the best predictor as it nudges the forecast back to a normal westerly wind regime.

Grumm, R., & Siebers, A. (1990). Systematic Model Forecast Errors of Surface Cyclones in the NGM and AVN, January 1990. Weather and Forecasting, 5(4), 672-682 https://doi.org/10.1175/1520-0434(1990)005<0672:SMFEOS>2.0.CO;2

Results from a study examining the performance of the nested grid model (NGM) and the aviation run of the global spectral model (AVN) in predicting surface cyclones during January 1990 revealed that the AVN slightly outperformed the NGM in forecasting cyclone central pressures and placement. Although both models performed better for deepening systems than filling systems, the AVN outperformed the NGM in predicting the characteristics of filling cyclones. Overall, the NGM tended to overdeepen surface cyclones. A large part of the pressure error was due to the model's inability to properly fill cyclones and a tendency to forecast systems to deepen when they were observed to be filling. The AVN tended to underdeepen surface cyclones with the deepening rate errors near 2 mb at 12 h and less than 1 mb by 48 h. The overall pressure errors for deepening cyclones appeared to be linked to a spin-up problem in the AVN and may have also been associated with the AVN cold bias in 1000- to 500-mb thickness forecasts.

Marzban, C., Leyton, S., & Colman, B. (2007). Ceiling and Visibility Forecasts via Neural Networks. Weather and Forecasting, 22(3), 466-479 https://doi.org/10.1175/WAF994.1

Statistical postprocessing of numerical model output can improve forecast quality, especially when model output is combined with surface observations. In this article, the development of nonlinear postprocessors for the prediction of ceiling and visibility is discussed. The forecast period is approximately 2001-05, involving data from hourly surface observations, and from the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model. The statistical model for mapping these data to ceiling and visibility is a neural network. A total of 39 such neural networks are developed for each of 39 terminal aerodrome forecast stations in the northwest United States. These postprocessors are compared with a number of alternatives, including logistic regression, and model output statistics (MOS) derived from the Aviation Model/Global Forecast System. It is found that the performance of the neural networks is generally superior to logistic regression and MOS. Depending on the comparison, different measures of performance are examined, including the Heidke skill statistic, cross-entropy, relative operating characteristic curves, discrimination plots, and attributes diagrams. The extent of the improvement brought about by the neural network depends on the measure of performance, and the specific station.

Monaghan, A. J., Bromwich, D. H., Wei, H. L., Cayette, A. M., Powers, J. G., Kuo, Y. H., & Lazzara, M. A. (2003). Performance of Weather Forecast Models in the Rescue of Dr. Ronald Shemenski from the South Pole in April 2001. Weather and Forecasting, 18(2), 142-160 https://doi.org/10.1175/1520-0434(2003)018<0142:POWFMI>2.0.CO;2

In late April 2001, an unprecedented late-season flight to Amundsen-Scott South Pole Station was made in the evacuation of Dr. Ronald Shemenski, a medical doctor seriously ill with pancreatitis. This case study analyzes the performance of four of the numerical weather prediction models that aided meteorologists in forecasting weather throughout the operation: 1) the Antarctic Mesoscale Prediction System (AMPS) Polar MM5 (fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model), 2) the National Centers for Environmental Prediction Aviation Model (AVN), 3) the European Centre for Medium-Range Weather Forecasts (ECMWF) global forecast model, and 4) the NCAR Global MM5. To identify specific strengths and weaknesses, key variables for each model are statistically analyzed for all forecasts initialized between 21 and 25 April for several points over West Antarctica at the surface and at 500- and 700-hPa levels. The ECMWF model performs with the highest overall skill, generally having the lowest bias and rms errors and highest correlations for the examined fields. The AMPS Polar MM5 exhibits the next best skill, followed by AVN and Global MM5. For the surface variables, all of the models show high skill in predicting surface pressure but demonstrate modest skill in predicting temperature, wind speed, and wind direction. In the free atmosphere, the models show high skill in forecasting geopotential height, considerable skill in predicting temperature and wind direction, and good skill in predicting wind speed. In general, the models produce very useful forecasts in the free atmosphere, but substantial efforts are still needed to improve the surface prediction. The spatial resolution of each model exerts an important influence on forecast accuracy, especially in the complex topography of the Antarctic coastal regions. The initial and boundary conditions for the AMPS Polar MM5 exert a significant influence on forecasts.

Rudack, D. E., & Ghirardelli, J. E. (2010). A Comparative Verification of Localized Aviation Model Output Statistics Program (LAMP) and Numerical Weather Prediction (NWP) Model Forecasts of Ceiling

Height and Visibility. Weather and Forecasting, 25(4), 1161-1178 https://doi.org/10.1175/2010WAF2222383.1

In an effort to support aviation forecasting, the National Weather Service's Meteorological Development Laboratory (MDL) has recently redeveloped the Localized Aviation Model Output Statistics (MOS) Program (LAMP) system. LAMP is designed to run hourly in NWS operations and produce short-range aviation forecast guidance at 1-h projections out to 25 h. This paper compares and contrasts LAMP ceiling height and visibility forecasts with forecasts produced by the 20-km Rapid Update Cycle model (RUC20), the Weather Research and Forecasting Nonhydrostatic Mesoscale Model (WRF-NMM), and the Short-Range Ensemble Forecast system (SREF). RUC20 and WRF-NMM forecasts of continuous ceiling height and visibility were interpolated to stations and converted into categorical forecasts. These interpolated forecasts were also categorized into instrument flight rule (IFR) or lower conditions and verified against LAMP forecasts at stations in the contiguous United States. LAMP and SREF probabilistic forecasts of ceiling height and visibility from LAMP and the SREF system were also verified. This study demonstrates that for the 0000 and 1200 UTC cycles over the contiguous United States, LAMP station-based categorical forecasts of ceiling height, visibility, and IFR conditions or lower are more accurate than the RUC20 and WRF-NMM ceiling height and visibility forecasts interpolated to stations. Moreover, for the 0900 and 2100 UTC forecast cycles and verification periods studied here, LAMP ceiling height and visibility probabilities exhibit better reliability and skill than the SREF system.

Tolman, H. L. (1998). Validation of NCEP's Ocean Winds for the Use in Wind Wave Models. Global Atmosphere and Ocean System, 6(3), 243-268 Retrieved from https://polar.ncep.noaa.gov/mmab/papers/tn150/OMB150.pdf

The quality of analyzed ocean surface winds from the Global Data Assimilation System (GDAS) and forecasted winds from the early or `aviation' cycle of the global medium range forecast model (AVN) of the National Centers for Environmental Prediction (NCEP) is assessed as part of a validation study of a new wave forecast system. This validation is performed using conventional buoy data and satellite retrieved wind speeds from the ERS1 altimeter and scatterometer. Both GDAS and AVN wind fields are shown to include moderate systematic biases, for which statistical corrections based on both satellite and buoy data are presented. Furthermore, buoy data are shown not to be representative for a global validation study. The altimeter data are potentially of significant importance for wave model validations, as they include collocated wind and wave measurements. The altimeter winds, however, are shown to be seriously contaminated by the development stage of the wave field. As it does not appear to be possible to remove this contamination, altimeter wind data should not be used in the validation of wave models.

Section II: Global Forecast System (GFS)

Baek, S.-J., Szunyogh, I., Hunt, B. R., & Ott, E. (2009). Correcting for Surface Pressure Background Bias in Ensemble-Based Analyses. *Monthly Weather Review, 137*(7), 2349-2364 https://doi.org/10.1175/2008MWR2787.1

Model error is the component of the forecast error that is due to the difference between the dynamics of the atmosphere and the dynamics of the numerical prediction model. The systematic, slowly varying part of the model error is called model bias. This paper evaluates three different ensemble-based

strategies to account for the surface pressure model bias in the analysis scheme. These strategies are based on modifying the observation operator for the surface pressure observations by the addition of a bias-correction term. One estimates the correction term adaptively, while another uses the hydrostatic balance equation to obtain the correction term. The third strategy combines an adaptively estimated correction term and the hydrostatic-balance-based correction term. Numerical experiments are carried out in an idealized setting, where the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model is integrated at resolution T62L28 to simulate the evolution of the atmosphere and the T30L7 resolution Simplified Parameterization Primitive Equation Dynamics (SPEEDY) model is used for data assimilation. The results suggest that the adaptive bias-correction term is effective in correcting the bias in the data-rich regions, while the hydrostatic-balance-based approach is effective in data-sparse regions. The adaptive bias-correction approach also has the benefit that it leads to a significant improvement of the temperature and wind analysis at the higher model levels. The best results are obtained when the two bias-correction approaches are combined.

Bhargava, K., Kalnay, E., Carton, J. A., & Yang, F. (2018). Estimation of Systematic Errors in the GFS Using Analysis Increments. *Journal of Geophysical Research-Atmospheres, 123*(3), 1626-1637 https://doi.org/10.1002/2017JD027423

We estimate the effect of model deficiencies in the Global Forecast System that lead to systematic forecast errors, as a first step toward correcting them online (i.e., within the model) as in Danforth & Kalnay (2008a, 2008b). Since the analysis increments represent the corrections that new observations make on the 6h forecast in the analysis cycle, we estimate the model bias corrections from the time average of the analysis increments divided by 6h, assuming that initial model errors grow linearly and first ignoring the impact of observation bias. During 2012-2016, seasonal means of the 6h model bias are generally robust despite changes in model resolution and data assimilation systems, and their broad continental scales explain their insensitivity to model resolution. The daily bias dominates the submonthly analysis increments and consists primarily of diurnal and semidiurnal components, also requiring a low dimensional correction. Analysis increments in 2015 and 2016 are reduced over oceans, which we attribute to improvements in the specification of the sea surface temperatures. These results provide support for future efforts to make online correction of the mean, seasonal, and diurnal and semidiurnal model biases of Global Forecast System to reduce both systematic and random errors, as suggested by Danforth & Kalnay (2008a, 2008b). It also raises the possibility that analysis increments could be used to provide guidance in testing new physical parameterizations.

Boukabara, S.-A., Garrett, K., & Kumar, V. K. (2016). Potential Gaps in the Satellite Observing System Coverage: Assessment of Impact on NOAA's Numerical Weather Prediction Overall Skills. *Monthly Weather Review, 144*(7), 2547-2563 https://doi.org/10.1175/mwr-d-16-0013.1

The current constellation of environmental satellites is at risk of degrading due to several factors. This includes the following: 1) loss of secondary polar-orbiting satellites due to reaching their nominal lifetimes, 2) decrease in the density of extratropical radio-occultation (RO) observations due to a likely delayed launch of the Constellation Observing System for Meteorology, Ionosphere and Climate-2 (COSMIC-2) high inclination orbit constellation, and 3) the risk of losing afternoon polar-orbiting satellite coverage due to potential launch delays in the Joint Polar Satellite System (JPSS) programs. In this study, the impacts from these scenarios on the National Oceanic and Atmospheric Administration (NOAA) Global Forecast System skill are quantified. Performances for several metrics are assessed, but to

encapsulate the results the authors introduce an overall forecast score combining metrics for all parameters, atmospheric levels, and forecast lead times. The first result suggests that removing secondary satellites results in significant degradation of the forecast. This is unexpected since it is generally assumed that secondary sensors contribute to system's robustness but not necessarily to forecast performance. Second, losing the afternoon orbit on top of losing secondary satellites further degrades forecast performances by a significant margin. Finally, losing extratropical RO observations on top of losing secondary satellites also negatively impacts the forecast performances, but to a lesser degree. These results provide a benchmark that will allow for the assessment of the added value of projects being implemented at NOAA in support of mitigation strategies designed to alleviate the negative impacts associated with these data gaps, and additionally help NOAA to define requirements of the future global observing system architecture.

Chien, F.-C., Hong, J.-S., & Kuo, Y.-H. (2019). The Marine Boundary Layer Height over the Western North Pacific Based on GPS Radio Occultation, Island Soundings, and Numerical Models. *Sensors, 19*(1), 155 https://doi.org/10.3390/s19010155

This paper estimates marine boundary layer height (MBLH) over the western North Pacific (WNP) based on Global Positioning System Radio Occultation (GPS-RO) profiles from the Formosa Satellite Mission 3 (FORMOSAT-3)/Constellation Observing System for Meteorology, Ionosphere, and Climate (COSMIC) satellites, island soundings, and numerical models. The seasonally-averaged MBLHs computed from nine years (2007-2015) of GPS-RO data are inter-compared with those obtained from sounding observations at 15 island stations and from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis (ERA-Interim) and National Centers for Environmental Prediction Global Forecast System (NCEP GFS) data over the WNP from 2012 to 2015. It is found that the MBLH using nine years of GPS-RO data is smoother and more consistent with that obtained from sounding observations than is the MBLH using four years of GPS-RO data in a previous study. In winter, higher MBLHs are found around the subtropical latitudes and over oceans east of Japan, which are approximately located within the paths of the North Equatorial Current and the Kuroshio Current. The MBLH is also significantly higher in winter than in summer over the WNP. The above MBLH pattern is generally similar to those obtained from the analysis data of the ERA-Interim and NCEP GFS, but the heights are about 200 m higher. The verification with soundings suggests that the ERA-Interim has a better MBLH estimation than the NCEP GFS. Thus, the MBLH distributions obtained from both the nine-year GPS-RO and the ERA-Interim data can represent well the climatological MBLH over the WNP, but the heights should be adjusted about 30 m lower for the former and similar to 200 m higher for the latter. A positive correlation between the MBLH and the instability of the lower atmosphere exists over large near-shore areas of the WNP, where cold air can move over warm oceans from the land in winter, resulting in an increase in lower-atmospheric instability and providing favorable conditions for convection to yield a higher MBLH. During summer, the lower-atmospheric instability becomes smaller and the MBLH is thus lower over near-shore oceans.

Dawson, N., Broxton, P., Zeng, X., Leuthold, M., Barlage, M., & Holbrook, P. (2016). An Evaluation of Snow Initializations in NCEP Global and Regional Forecasting Models. *Journal of Hydrometeorology, 17*(6), 1885-1901 https://doi.org/10.1175/JHM-D-15-0227.1

Snow plays a major role in land-atmosphere interactions, but strong spatial heterogeneity in snow depth (SD) and snow water equivalent (SWE) makes it challenging to evaluate gridded snow quantities using in situ measurements. First, a new method is developed to upscale point measurements into gridded

datasets that is superior to other tested methods. It is then utilized to generate daily SD and SWE datasets for water years 2012-14 using measurements from two networks (COOP and SNOTEL) in the United States. These datasets are used to evaluate daily SD and SWE initializations in NCEP global forecasting models (GFS and CFSv2, both on 0.5° × 0.5° grids) and regional models (NAM on 12 km × 12 km grids and RAP on 13 km × 13 km grids) across eight 2° × 2° boxes. Initialized SD from three models (GFS, CFSv2, and NAM) that utilize Air Force Weather Agency (AFWA) SD data for initialization is 77% below the area-averaged values, on average. RAP initializations, which cycle snow instead of using the AFWA SD, underestimate SD to a lesser degree. Compared with SD errors, SWE errors from GFS, CFSv2, and NAM are larger because of the application of unrealistically low and globally constant snow densities. Furthermore, the widely used daily gridded SD data produced by the Canadian Meteorological Centre (CMC) are also found to underestimate SD (similar to GFS, CFSv2, and NAM), but are worse than RAP. These results suggest an urgent need to improve SD and SWE initializations in these operational models.

Fletcher, J. K., Bretherton, C. S., Xiao, H., Sun, R., & Han, J. (2014). Improving Subtropical Boundary Layer Cloudiness in the 2011 NCEP GFS. *Geoscientific Model Development, 7*(5), 2107-2120 https://doi.org/10.5194/gmd-7-2107-2014

The current operational version of National Centers for Environmental Prediction (NCEP) Global Forecasting System (GFS) shows significant low cloud bias. These biases also appear in the Coupled Forecast System (CFS), which is developed from the GFS. These low cloud biases degrade seasonal and longer climate forecasts, particularly of short-wave cloud radiative forcing, and affect predicted sea surface temperature. Reducing this bias in the GFS will aid the development of future CFS versions and contributes to NCEP's goal of unified weather and climate modelling. Changes are made to the shallow convection and planetary boundary layer parameterisations to make them more consistent with current knowledge of these processes and to reduce the low cloud bias. These changes are tested in a single-column version of GFS and in global simulations with GFS coupled to a dynamical ocean model. In the single-column model, we focus on changing parameters that set the following: the strength of shallow cumulus lateral entrainment, the conversion of updraught liquid water to precipitation and grid-scale condensate, shallow cumulus cloud top, and the effect of shallow convection in stratocumulus environments. Results show that these changes improve the single-column simulations when compared to large eddy simulations, in particular through decreasing the precipitation efficiency of boundary layer clouds. These changes, combined with a few other model improvements, also reduce boundary layer cloud and albedo biases in global coupled simulations.

Werth, D., & Garrett, A. (2011). Patterns of Land Surface Errors and Biases in the Global Forecast System. *Monthly Weather Review, 139*(5), 1569-1582 https://doi.org/10.1175/2010MWR3423.1

One year's worth of Global Forecast System (GFS) predictions of surface meteorological variables (wind speed, air temperature, dewpoint temperature, sea level pressure) are validated for land-based stations over the entire planet for forecasts extending from 0 h into the future (an analysis) to 7 days. Approximately 12 000 surface stations worldwide were included in this analysis. Root-mean-square errors (RMSEs) increased as the forecast period increased from 0 to 36 h, but the initial RMSEs were almost as large as the 36-h forecast RMSEs for all variables. Typical RMSEs were 3 degrees C for air temperature, 2-3 mb for sea level pressure, 3.5 degrees C for dewpoint temperature, and 2.5 m s(-1) for wind speed. An analysis of the biases at each station shows that the biggest errors are associated with

mountain ranges and other areas of steep topography, with land-sea contrasts also playing a role. When the error is decomposed into the bias, variance, and correlation terms, the large initial RMSEs for the 0-h forecasts are seen to be due to a large forecast bias (which persisted into the longer forecasts) with errors in forecast correlation also making a large contribution. A validation of two subdomains showed results similar to the global validation, but the dependence of the biases on the forecast time was clearer. Finally, the RMSE values climb as forecasts go out when validated out to a period of 7 days as the correlation error term grows.

Wolff, J. K., Harrold, M., Fowler, T., Gotway, J. H., Nance, L., & Brown, B. G. (2014). Beyond the Basics: Evaluating Model-Based Precipitation Forecasts Using Traditional, Spatial, and Object-Based Methods. *Weather and Forecasting, 29*(6), 1451-1472 https://doi.org/10.1175/WAF-D-13-00135.1

While traditional verification methods are commonly used to assess numerical model quantitative precipitation forecasts (QPFs) using a grid-to-grid approach, they generally offer little diagnostic information or reasoning behind the computed statistic. On the other hand, advanced spatial verification techniques, such as neighborhood and object-based methods, can provide more meaningful insight into differences between forecast and observed features in terms of skill with spatial scale, coverage area, displacement, orientation, and intensity. To demonstrate the utility of applying advanced verification techniques to mid- and coarse-resolution models, the Developmental Testbed Center (DTC) applied several traditional metrics and spatial verification techniques to QPFs provided by the Global Forecast System (GFS) and operational North American Mesoscale Model (NAM). Along with frequency bias and Gilbert skill score (GSS) adjusted for bias, both the fractions skill score (FSS) and Method for Object-Based Diagnostic Evaluation (MODE) were utilized for this study with careful consideration given to how these methods were applied and how the results were interpreted. By illustrating the types of forecast attributes appropriate to assess with the spatial verification techniques, this paper provides examples of how to obtain advanced diagnostic information to help identify what aspects of the forecast are or are not performing well.

Yang, F., Pan, H.-L., Krueger, S. K., Moorthi, S., & Lord, S. J. (2006). Evaluation of the NCEP Global Forecast System at the ARM SGP Site. *Monthly Weather Review, 134*(12), 3668-3690 https://doi.org/10.1175/MWR3264.1

This study evaluates the performance of the National Centers for Environmental Prediction Global Forecast System (GFS) against observations made by the U.S. Department of Energy Atmospheric Radiation Measurement (ARM) Program at the southern Great Plains site for the years 2001-04. The spatial and temporal scales of the observations are examined to search for an optimum approach for comparing grid-mean model forecasts with single-point observations. A single-column model (SCM) based upon the GFS was also used to aid in understanding certain forecast errors. The investigation is focused on the surface energy fluxes and clouds. Results show that the overall performance of the GFS model has been improving, although certain forecast errors remain. The model overestimated the daily maximum latent heat flux by 76 W m(-2) and the daily maximum surface downward solar flux by 44 W m(-2), and underestimated the daily maximum sensible heat flux by 44 W m(-2). The model's surface energy balance was reached by a cancellation of errors. For clouds, the GFS was able to capture the observed evolutions of cloud systems during major synoptic events. However. on average, the model largely underestimated cloud fraction in the lower and midtroposphere, especially for daytime

nonprecipitating low clouds because shallow convection in the GFS does not produce clouds. Analyses of surface radiative fluxes revealed that the diurnal cycle of the model's surface downward longwave flux (SDLW) was not in phase with that of the ARM-observed SDLW. SCM experiments showed that this error was caused by an inaccurate scaling factor, which was a function of ground skin temperature and was used to adjust the SDLW at each model time step to that computed by the model's longwave radiative transfer routine once every 3 h. A method has been proposed to correct this error in the operational forecast model. It was also noticed that the SDLW biases changed from mostly negative in 2003 to slightly positive in 2004. This change was traced back to errors in the near-surface air temperature. In addition, the SDLW simulated with the newly implemented Rapid Radiative Transfer Model longwave routine in the GFS is usually 5(-10) W m(-2) larger than that simulated with the previous routine. The forecasts of surface downward shortwave flux (SDSW) were relatively accurate under clear-sky conditions. The errors in SDSW were primarily caused by inaccurate forecasts of cloud properties. Results from this study can be used as guidance for the further development of the GFS.

Yang, X., DelSole, T., & Pan, H.-L. (2008). Empirical Correction of the NCEP Global Forecast System. *Monthly Weather Review, 136*(12), 5224-5233 https://doi.org/10.1175/2008MWR2527.1

This paper examines the extent to which an empirical correction method can improve forecasts of the National Centers for Environmental Prediction (NCEP) operational Global Forecast System. The empirical correction is based on adding a forcing term to the prognostic equations equal to the negative of the climatological tendency errors. The tendency errors are estimated by a least squares method using 6-, 12-, 18-, and 24-h forecast errors. Tests on independent verification data show that the empirical correction significantly reduces temperature biases nearly everywhere at all lead times up to at least 5 days but does not significantly reduce biases in forecast winds and humidity. Decomposing mean-square error into bias and random components reveals that the reduction in total mean-square error arises solely from reduction in bias. Interestingly, the empirical correction increases the random error slightly, but this increase is argued to be an artifact of the change in variance in the forecasts. The empirical correction also is found to reduce the bias more than traditional “after the fact” corrections. The latter result might be a consequence of the very different sample sizes available for estimation, but this difference in sample size is unavoidable in operational situations in which limited calibration data are available for a given forecast model.

Yin, J., & Zhan, X. (2018). Impact of Bias-Correction Methods on Effectiveness of Assimilating SMAP Soil Moisture Data into NCEP Global Forecast System Using the Ensemble Kalman Filter. *IEEE Geoscience and Remote Sensing Letters, 15*(5), 659-663 https://doi.org/10.1109/LGRS.2018.2806092

Improving numerical weather prediction was one of the main justifications for National Aeronautics and Space Administration's Soil Moisture Active/Passive (SMAP) Mission. The ensemble Kalman filter (EnKF) has been extensively applied to assimilate the SM observations into numerical weather predication models. Implementation of EnKF requires the observations and model simulations to be Gaussian distributed and not biased from each other. In this letter, we tested the impacts of three bias-correction methods on effectiveness of assimilating SMAP retrievals into the National Oceanic and Atmospheric Administration-National Centers for Environmental Prediction Global Forecast System (GFS). They are: 1) global cumulative distribution function (CDF) matching with only one CDF for all grids and time series; 2) monthly CDF matching with one CDF for each grid; 3) the linear transformation technique that matches

monthly mean and standard deviation of the SMAP retrievals and model simulations for each grid; and 4) assimilating SMAP SM data into GFS without any bias-correction procedure. With respect to the global land data assimilation (DA) system precipitation product, the results demonstrate that the effectiveness of assimilating SMAP retrievals into GFS is significantly impacted by the bias-correction methods. Relative to other DA cases, the monthly CDF matching produces the best precipitation forecast performance. Improvements of the three-hourly GFS precipitation prediction with SMAP assimilation using the monthly CDF matching can reach to 8% and 10% in sparsely and densely vegetated areas, respectively, and marginally positive in medium vegetation areas. Based on these results, assimilating SMAP retrievals into the GFS with the EnKF algorithm using the monthly CDF matching method is suggested for enhancing accuracy of the precipitation forecasts.

Yoo, H., & Li, Z. (2012). Evaluation of Cloud Properties in the NOAA/NCEP Global Forecast System Using Multiple Satellite Products. *Climate Dynamics, 39*(12), 2769-2787 https://doi.org/10.1007/s00382-012-1430-0

Knowledge of cloud properties and their vertical structure is important for meteorological studies due to their impact on both the Earth's radiation budget and adiabatic heating within the atmosphere. The objective of this study is to evaluate bulk cloud properties and vertical distribution simulated by the US National Oceanic and Atmospheric Administration National Centers for Environmental Prediction Global Forecast System (GFS) using three global satellite products. Cloud variables evaluated include the occurrence and fraction of clouds in up to three layers, cloud optical depth, liquid water path, and ice water path. Cloud vertical structure data are retrieved from both active (CloudSat/CALIPSO) and passive sensors and are subsequently compared with GFS model results. In general, the GFS model captures the spatial patterns of hydrometeors reasonably well and follows the general features seen in satellite measurements, but large discrepancies exist in low-level cloud properties. More boundary layer clouds over the interior continents were generated by the GFS model whereas satellite retrievals showed more low-level clouds over oceans. Although the frequencies of global multi-layer clouds from observations are similar to those from the model, latitudinal variations show discrepancies in terms of structure and pattern. The modeled cloud optical depth over storm track region and subtropical region is less than that from the passive sensor and is overestimated for deep convective clouds. The distributions of ice water path (IWP) agree better with satellite observations than do liquid water path (LWP) distributions. Discrepancies in LWP/IWP distributions between observations and the model are attributed to differences in cloud water mixing ratio and mean relative humidity fields, which are major control variables determining the formation of clouds.

Zheng, W., Ek, M., Mitchell, K., Wei, H., & Meng, J. (2017). Improving the Stable Surface Layer in the NCEP Global Forecast System. *Monthly Weather Review, 145*(10), 3969-3987 https://doi.org/10.1175/MWR-D-16-0438.1

This study examines the performance of the NCEP Global Forecast System (GFS) surface layer parameterization scheme for strongly stable conditions over land in which turbulence is weak or even disappears because of high near-surface atmospheric stability. Cases of both deep snowpack and snow-free conditions are investigated. The results show that decoupling and excessive near-surface cooling may appear in the late afternoon and nighttime, manifesting as a severe cold bias of the 2-m surface air temperature that persists for several hours or more. Concurrently, because of negligible downward heat transport from the atmosphere to the land, a warm temperature bias develops at the first model level.

The authors test changes to the stable surface layer scheme that include introduction of a stability parameter constraint that prevents the land-atmosphere system from fully decoupling and modification to the roughness-length formulation. GFS sensitivity runs with these two changes demonstrate the ability of the proposed surface layer changes to reduce the excessive near-surface cooling in forecasts of 2-m surface air temperature. The proposed changes prevent both the collapse of turbulence in the stable surface layer over land and the possibility of numerical instability resulting from thermal decoupling between the atmosphere and the surface. The authors also execute and evaluate daily GFS 7-day test forecasts with the proposed changes spanning a one-month period in winter. The assessment reveals that the systematic deficiencies and substantial errors in GFS near-surface 2-m air temperature forecasts are considerably reduced, along with a notable reduction of temperature errors throughout the lower atmosphere and improvement of forecast skill scores for light and medium precipitation amounts.

Zheng, W., Wei, H., Wang, Z., Zeng, X., Meng, J., Ek, M., . . . Derber, J. (2012). Improvement of Daytime Land Surface Skin Temperature over Arid Regions in the NCEP GFS Model and Its Impact on Satellite Data Assimilation. *Journal of Geophysical Research-Atmospheres, 117*, D06117 https://doi.org/10.1029/2011JD015901

Comparison of the land surface skin temperature (LST) from the National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS) against satellite and in situ data in summer 2007 indicates that the GFS has a large and cold bias in LST over the arid western continental United States (CONUS) during daytime. This LST bias contributes to large errors in simulated satellite brightness temperatures over land by the Community Radiative Transfer Model (CRTM) and hence the rejection of satellite data in the NCEP Gridpoint Statistical Interpolation (GSI) system, especially for surface-sensitive satellite channels. The new vegetation-dependent formulations of momentum and thermal roughness lengths are tested in the GFS. They substantially reduce the large cold bias of daytime LST over the arid western CONUS in the warm season. This, in turn, significantly reduces the large biases of calculated satellite brightness temperatures found for infrared and microwave sensors in window or near-window channels, so that many more satellite data can be assimilated in the GSI system. In the arid western CONUS, the calculation of surface emissivity for microwave sensors in the CRTM can be further improved, and the new microwave land emissivity model together with increased LST via changes in surface roughness length formulations reduces biases and root-mean-square errors in the calculated brightness temperature.

Section III: ETA Coordinate Model (ETA)

Chou, S. C., Tanajura, C. a. S., Xue, Y. K., & Nobre, C. A. (2002). Validation of the Coupled ETA/SSiB Model over South America. *Journal of Geophysical Research-Atmospheres, 107*(D20), 8088 https://doi.org/10.1029/2000JD000270

Two 1-month integrations were performed with the regional ETA model coupled with the Simplified Simple Biosphere model (SSiB) over South America. The goal of the present work is to validate the model and to investigate its biases and skill on the simulations of South American climate. This is an initial step on the use of this model for climate research. The ETA model was set up with 80-km horizontal resolution and 38 vertical layers over the South American continent and part of the adjacent oceans. Analyses from the National Centers for Environmental Prediction (NCEP) were used as initial and

lateral boundary conditions. The selected months were August and November 1997, which are in opposite phases of the precipitation annual cycle observed in the central part of South America. The model was integrated continuously for each 1-month period. Monthly means and daily variations of simulated precipitation and surface temperature compare well with observations. The patterns of simulated outgoing longwave radiation are also similar to the observed ones. However, a positive bias is verified in the simulations. The model shows a positive bias in latent and sensible heat surface fluxes due to an excessive shortwave incoming radiation at the surface. Comparisons with a version of the ETA model coupled with the bucket model shows that the ETA/SSiB version improves the surface temperature and increases precipitation in the interior of the continent during wet months.

Colle, B. A., Mass, C. F., & Ovens, D. (2001). Evaluation of the Timing and Strength of MM5 and ETA Surface Trough Passages over the Eastern Pacific. *Weather and Forecasting, 16*(5), 553-572 https://doi.org/10.1175/1520-0434(2001)016%3C0553:EOTTAS%3E2.0.CO;2

The timing and strength of suirface trough forecast by the Pennsylvania State University-National Center for Atmospheric Research fifth-generation Mesoscale Model (MM5) and the National Centers for Environmental Prediction's (NCEP's) ETA Model are evaluated over the eastern Pacific during the 1997-2000 cool seasons (Sep-Mar).

Colle, B. A., Olson, J. B., & 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 https://doi.org/10.1175/1520-0434(2003)18<431:MVOTMP>2.0.CO;2

This paper describes the multiseason verification of the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model (MM5) and the National Centers for Environmental Prediction (NCEP) ETA Model over the eastern two-thirds of the United States and surrounding coastal waters during the cool (1 November - 31 March) and warm (1 May - 30 September) seasons from the autumn of 1999 through the summer of 2001. Verification statistics are calculated by interpolating model forecasts to the observation sites. The horizontal and vertical distributions of model errors are presented as are the diurnal and intraseasonal trends. During the cool season, both the MM5 and ETA have a low-level cool and moist bias over land, a significant surface warm bias over water, and surface winds that are too strong over land to the east of the Rockies and too weak over water. The low-level cool and moist bias is maximized during the day, and the cool bias is largest during late winter. During the warm season, the MM5 and ETA have little temperature bias over water and a negative wind speed bias over the Rockies. Both the MM5 and ETA have a surface dry and warm bias over land during the warm season; however, the ETA warm-season biases were reduced between 2000 and 2001 because of recent improvements to the land surface model and soil moisture initialization. Using the NCEP Aviation Model rather than the ETA to initialize the MM5 during the 2000/01 cool season resulted in slightly better sea level pressure forecasts over the Northeast on average, but not for wind and temperature. In order to quantify the impact of increased resolution, the MM5 was verified down to 4-km grid spacing around coastal southern New England. For 32 objectively identified sea-breeze events, the 12-km MM5 has significantly greater wind and temperature skill along the coast than the 36-km version, but there is little improvement from 12 to 4 km. The sea breezes in the MM5 are too early on average and are associated with a late afternoon cool bias. Many of the MM5 and ETA errors have slowly evolving intraseasonal trends. In particular, the cool bias amplifies during the winter and the

summer dry bias in the MM5 increases during prolonged wet periods. The sea level pressure errors are episodic, with clusters of negative and positive mean errors lasting approximately 3 - 6 weeks, thereby suggesting a dependence on the large-scale flow.

Colle, B. A., Olson, J. B., & Tongue, J. S. (2003). Multiseason Verification of the MM5. Part II: Evaluation of High-Resolution Precipitation Forecasts over the Northeastern United States. *Weather and Forecasting, 18*(3), 458-480 https://doi.org/10.1175/1520-0434(2003)18<458:MVOTMP>2.0.CO;2

This paper evaluates the fifth-generation Pennsylvania State University - National Center for Atmospheric Research Mesoscale Model (MM5) precipitation forecasts over the northeastern United States to show the effects of increasing resolution, the spatial variations in model skill, and the impact of convective parameterizations on the MM5 precipitation forecasts. The MM5 is verified during the cool seasons (November - March) of 1999 2001 and the warm season (May - September) of 2000 using approximately 500 cooperative observer and National Weather Service precipitation sites. During the cool season, the 12-km MM5 produces excessive precipitation immediately downwind of the Great Lakes and along the windward slopes of the Appalachians and too little precipitation in the lee of the barrier. The 36-km MM5 has slightly more skill than at 12-km grid spacing for the light to moderate thresholds, while the 12-km precipitation forecasts are slightly better on average for the heavy precipitation events. During the 2000/01 cool season, two separate MM5 runs were completed twice daily using the National Centers for Environmental Prediction ETA (ETA - MM5) and Aviation (AVN - MM5) Models for initial and boundary conditions. The ETA - MM5 had slightly lower (better) rms errors than the AVN - MM5 for the weak to moderate thresholds (2.54 - 50.8 mm in 24 h), while for the heavier thresholds the AVN - MM5 had significantly lower rms errors than the ETA - MM5. As a result, the 36-km AVN - MM5 was as skillful as the 12-km ETA - MM5 for these higher thresholds. During the warm season, both the 36- and 12-km grid spacings overpredict precipitation just inland of the coast and significantly underpredict farther inland over the Appalachians. This coastal overprediction originated from an overactive Kain-Fritsch (KF) convective parameterization, while the inland underprediction is associated with a low-level dry bias during the warm season. A representative case study shows that both the Betts-Miller and Grell parameterizations produce less precipitation near the coast than the overactive KF scheme. A new and alternate version of KF (KF2) in the MM5 may also help to reduce this coastal overprediction. The 4-km MM5 explicit precipitation during the summer is sensitive to which convective parameterization is applied in the outer domains. Using KF or KF2 in the 36- and 12-km domains suppresses the explicit precipitation in the 4-km nest, especially for weak to moderate events over western sections of the 4-km domain. For a representative event, the Betts - Miller and Grell convective schemes allowed for a more realistic 4-km precipitation distribution, while a simulation using no convective parameterization in the 36- and 12-km domains produced excessive rain rates in the 4-km forecasts.

Gallus, W. A., Baldwin, M. E., & Elmore, K. L. (2007). Evaluation of Probabilistic Precipitation Forecasts Determined from ETA and AVN Forecasted Amounts. *Weather and Forecasting, 22*(1), 207-215 https://doi.org/10.1175/WAF976.1

This note examines the connection between the probability of precipitation and forecasted amounts from the NCEP ETA (now known as the North American Mesoscale model) and Aviation (AVN; now known as the Global Forecast System) models run over a 2-yr period on a contiguous U.S. domain.

Specifically, the quantitative precipitation forecast (QPF)-probability relationship found recently by Gallus and Segal in 10-km grid spacing model runs for 20 warm season mesoscale convective systems is tested over this much larger temporal and spatial dataset. A 1-yr period was used to investigate the QPF-probability relationship, and the predictive capability of this relationship was then tested on an independent 1-yr sample of data. The same relationship of a substantial increase in the likelihood of observed rainfall exceeding a specified threshold in areas where model runs forecasted higher rainfall amounts is found to hold over all seasons. Rainfall is less likely to occur in those areas where the models indicate none than it is elsewhere in the domain; it is more likely to occur in those regions where rainfall is predicted, especially where the predicted rainfall amounts are largest. The probability of rainfall forecasts based on this relationship are found to possess skill as measured by relative operating characteristic curves, reliability diagrams, and Brier skill scores. Skillful forecasts from the technique exist throughout the 48-h periods for which ETA and AVN output were available. The results suggest that this forecasting tool might assist forecasters throughout the year in a wide variety of weather events and not only in areas of difficult-to-forecast convective systems.

Hirschberg, P. A., Shafran, P. C., Elsberry, R. L., & Ritchie, E. A. (2001). An Observing System Experiment with the West Coast Picket Fence. *Monthly Weather Review, 129*(10), 2585-2599 https://doi.org/10.1175/1520-0493(2001)129<2585:Aosewt>2.0.Co;2

Analyses and forecasts from a modern data assimilation and modeling system are used to evaluate the impact of a special rawinsonde dataset of 3-h soundings at seven sites interspersed with the seven regular sites along the West Coast (to form a so-called picket fence to intercept all transiting circulations) plus special 6-h rawin-sondes over the National Weather Service Western Region. Whereas four intensive observing periods (IOPs) are available, only two representative IOPs (IOP-3 and IOP-4) are described here. The special observations collected during each 12-h cycle are analyzed with the National Centers for Environmental Prediction (NCEP) ETA Data Assimilation System in a cold start from the NCEP-National Center for Atmospheric Research reanalyses as the initial condition. Forecasts up to 48 h with and without the special picket fence observations are generated by the 32-km horizontal resolution ETA Model with 45 vertical levels. The picket fence observations had little impact in some cases with smooth environmental flow. In other cases, relatively large initial increments were introduced offshore of the picket fence observations. However, these increments usually damped as they translated downstream. During IOP-3, the increments amplified east of the Rocky Mountains after only 24 h. Even though initially small, the increments in IOP-4 grew rapidly to 500-mb height increments similar to 20-25 m with accompanying meridional wind increments of 5-8 m s(-1) that contributed to maxima in shear vorticity. Many of the downstream amplifying circulations had associated precipitation increments similar to6 mm (6 h)(-1) between the control and experimental forecasts. The equitable threat scores against the cooperative station set for the first 24-h forecasts during IOP-3 had higher values at the 0.50- and 0.75-in-thresholds for the picket fence dataset. However, the overall four-IOP equitable threat scores were similar. Although the classical synoptic case was not achieved during the picket fence, these model forecasts suggest that such observations around the coast of the United States would impact the downstream forecasts when added in dynamically unstable regions. An ultimate picket fence of continuous remotely observing systems should be studied further.

Listemaa, S. A. (2002). *Workstation ETA Verification Efforts at the Lower Mississippi River Forecast Center*. Retrieved from https://ams.confex.com/ams/annual2002/webprogram/Paper29660.html

With the release of the National Centers for Environmental Prediction (NCEP) Workstation ETA, local National Weather Service (NWS) offices and others have the ability to run numerical models at the local level. The LMRFC is currently running the Workstation ETA twice a day (0000 and 1200 UTC cycles), with the output available for the forecaster. Verification of numerical model output is an import step in determining how well a model performs. Verification statistics, including absolute error, mean absolute error, and root mean square error (RMSE), are calculated using output from the Workstation ETA and quality- controlled Stage III (mulisensor) data. This presentation will show verification statistics from the Workstation ETA, and compare them to statistics from the NCEP operational models.

McMurdie, L., & Mass, C. (2004). Major Numerical Forecast Failures over the Northeast Pacific. *Weather and Forecasting, 19*(2), 338-356 https://doi.org/10.1175/1520-0434(2004)019%3C0338:MNFFOT%3E2.0.CO;2

Strong North Pacific storms that impact the North American west coast are sometimes poorly predicted in the short term (up to 48 h) by operational models, with cyclone position errors of hundreds of kilometers and central pressure errors of tens of millibars. These major numerical forecast failures still occur despite continuing improvements in modeling and data assimilation. The frequency and intensity of sea level pressure errors at buoy and coastal locations are documented by comparing the National Centers for Environmental Prediction ETA Model forecasts to observations and through case studies of two poorly forecast cyclones from the 2001/2002 winter season.

Section IV: Rapid Update Cycle Model (RUC)

Ancell, B. C., Mass, C. F., & Hakim, G. J. (2011). Evaluation of Surface Analyses and Forecasts with a Multiscale Ensemble Kalman Filter in Regions of Complex Terrain. *Monthly Weather Review, 139*(6), 2008-2024 https://doi.org/10.1175/2010MWR3612.1

Previous research suggests that an ensemble Kalman filter (EnKF) data assimilation and modeling system can produce accurate atmospheric analyses and forecasts at 30-50-km grid spacing. This study examines the ability of a mesoscale EnKF system using multiscale (36/12 km) Weather Research and Forecasting (WRF) model simulations to produce high-resolution, accurate, regional surface analyses, and 6-h forecasts. This study takes place over the complex terrain of the Pacific Northwest, where the small-scale features of the near-surface flow field make the region particularly attractive for testing an EnKF and its flow-dependent background error covariances. A variety of EnKF experiments are performed over a 5-week period to test the impact of decreasing the grid spacing from 36 to 12 km and to evaluate new approaches for dealing with representativeness error, lack of surface background variance, and low-level bias. All verification in this study is performed with independent, unassimilated observations. Significant surface analysis and 6-h forecast improvements are found when EnKF grid spacing is reduced from 36 to 12 km. Forecast improvements appear to be a consequence of increased resolution during model integration, whereas analysis improvements also benefit from high-resolution ensemble covariances during data assimilation. On the 12-km domain, additional analysis improvements are found by reducing observation error variance in order to address representativeness error. Removing model surface biases prior to assimilation significantly enhances the analysis. Inflating surface wind and temperature background error variance has large impacts on analyses, but only produces small improvements in analysis RMS errors. Both surface and upper-air 6-h forecasts are nearly unchanged in

the 12-km experiments. Last, 12-km WRF EnKF surface analyses and 6-h forecasts are shown to generally outperform those of the Global Forecast System (GFS), North American Model (NAM), and the Rapid Update Cycle (RUC) by about 10%-30%, although these improvements do not extend above the surface. Based on these results, future improvements in multiscale EnKF are suggested.

Benjamin, S. G., Devenyi, D., Weygandt, S. S., Brundage, K. J., Brown, J. M., Grell, G. A., . . . Manikin, G. S. (2004). An Hourly Assimilation-Forecast Cycle: The RUC. *Monthly Weather Review, 132*(2), 495-518 https://doi.org/10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2

The Rapid Update Cycle (RUC), an operational regional analysis - forecast system among the suite of models at the National Centers for Environmental Prediction (NCEP), is distinctive in two primary aspects: its hourly assimilation cycle and its use of a hybrid isentropic - sigma vertical coordinate. The use of a quasi-isentropic coordinate for the analysis increment allows the influence of observations to be adaptively shaped by the potential temperature structure around the observation, while the hourly update cycle allows for a very current analysis and short-range forecast. Herein, the RUC analysis framework in the hybrid coordinate is described, and some considerations for high-frequency cycling are discussed. A 20-km 50-level hourly version of the RUC was implemented into operations at NCEP in April 2002. This followed an initial implementation with 60-km horizontal grid spacing and a 3-h cycle in 1994 and a major upgrade including 40-km horizontal grid spacing in 1998. Verification of forecasts from the latest 20-km version is presented using rawinsonde and surface observations. These verification statistics show that the hourly RUC assimilation cycle improves short-range forecasts (compared to longer-range forecasts valid at the same time) even down to the 1-h projection.

Benjamin, S. G., Grell, G. A., Brown, J. M., Smirnova, T. G., & Bleck, R. (2004). Mesoscale Weather Prediction with the RUC Hybrid Isentropic-Terrain-Following Coordinate Model. *Monthly Weather Review, 132*(2), 473-494 https://doi.org/10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2

A mesoscale atmospheric forecast model configured in a hybrid isentropic - sigma vertical coordinate and used in the NOAA Rapid Update Cycle (RUC) for operational numerical guidance is presented. The RUC model is the only quasi-isentropic forecast model running operationally in the world and is distinguished from other hybrid isentropic models by its application at fairly high horizontal resolution (10 - 20 km) and a generalized vertical coordinate formulation that allows model levels to remain continuous and yet be purely isentropic well into the middle and even lower troposphere. The RUC model is fully described in its 2003 operational version, including numerics and physical parameterizations. The use of these parameterizations, including mixed-phase cloud microphysics and an ensemble-closure-based cumulus parameterization, is fully consistent with the RUC vertical coordinate without any loss of generality. A series of experiments confirm that the RUC hybrid theta-sigma coordinate reduces cross-coordinate transport over a quasi-horizontal sigma coordinate. This reduction in cross-coordinate vertical transport results in less numerical vertical diffusion and thereby improves numerical accuracy for moist reversible processes. Finally, a forecast is presented of a strong cyclogenesis case over the eastern United States in which the RUC model produced an accurate 36-h prediction, especially in a 10-km nested version. Horizontal and vertical plots from these forecasts give evidence of detailed yet coherent structures of potential vorticity, moisture, and vertical motion.

Benjamin, S. G., Jamison, B. D., Moninger, W. R., Sahm, S. R., Schwartz, B. E., & Schlatter, T. W. (2010). Relative Short-Range Forecast Impact from Aircraft, Profiler, Radiosonde, VAD, GPS-PW, METAR, and Mesonet Observations via the RUC Hourly Assimilation Cycle. *Monthly Weather Review, 138*(4), 1319–1342 https://doi.org/10.1175/2009MWR3097.1

An assessment is presented on the relative forecast impact on the performance of a numerical weather prediction model from eight different observation data types: aircraft, profiler, radiosonde, velocity azimuth display (VAD), GPS-derived precipitable water, aviation routine weather report (METAR; surface), surface mesonet, and satellite-based atmospheric motion vectors. A series of observation sensitivity experiments was conducted using the Rapid Update Cycle (RUC) model/assimilation system in which various data sources were denied to assess the relative importance of the different data types for short-range (3-12 h) wind, temperature, and relative humidity forecasts at different vertical levels and near the surface. These experiments were conducted for two 10-day periods, one in November-December 2006 and one in August 2007. These experiments show positive short-range forecast impacts from most of the contributors to the heterogeneous observing system over the RUC domain. In particular, aircraft observations had the largest overall impact for forecasts initialized 3-6 h before 0000 or 1200 UTC, considered over the full depth (1000-100 hPa), followed by radiosonde observations, even though the latter are available only every 12 h. Profiler data (including at a hypothetical 8-km depth), GPS-precipitable water estimates, and surface observations also led to significant improvements in short-range forecast skill.

Benjamin, S. G., Smirnova, T. G., Brundage, K., Weygandt, S. S., Grell, G. A., Brown, J. M., . . . Smith, T. L. (2004). *Application of the Rapid Update Cycle at 10-13 Km-Initial Testing.* Paper presented at the 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction. Retrieved from http://ams.confex.com/ams/84Annual/wrfredirect.cgi?paperid=71092

To gain experience with the RUC system at anticipated higher horizontal resolution in operations at NCEP,testing of the RUC model at 10km resolution in regional domains of 2500 x 2000 km has been performed since winter 2000-2001. Real-time testing of a full national-scale domain at 13-km resolution will begin in fall 2003. Most recently, the 10km RUC model has been run over a domain covering much of eastern U.S. and southeastern Canada and having its southwestern corner over Oklahoma. This has been run in support of NOAA pilot programs to improve surface temperature forecasts with special operations in summers of 2002 and 2003. The northeast 10km RUC continued to run through winter 2002-2003. We will present results of objective verification of the 10km RUC forecasts against rawinsonde observations over this domain (about 35 rawinsondes) and compare against the 20km RUC running over the CONUS domain for those same stations. Many local forecasts related to surface effects have been shown to be considerably improved over those from 20km RUC (operational in April 2002),including orographic precipitation, lake-effect snow, and terrain-induced circulations. We will concentrate on behavior of surface wind, temperature and precipitation fields, forecast fields known to be sensitive to improved resolution of orography. We will also present results and example forecasts from the 13-km CONUS version of the RUC, which will be the next proposed resolution for the operational RUC at NCEP.

Caumont, O., Cimini, D., Loehnert, U., Alados-Arboledas, L., Bleisch, R., Buffa, F., . . . Pace, G. (2016). Assimilation of Humidity and Temperature Observations Retrieved from Ground-Based

Microwave Radiometers into a Convective-Scale NWP Model. *Quarterly Journal of the Royal Meteorological Society, 142*(700), 2692-2704 https://doi.org/10.1002/qj.2860

Temperature and humidity retrievals from an international network of ground-based microwave radiometers (MWRs) have been collected to assess the potential of their assimilation into a convective-scale numerical weather prediction (NWP) system. Thirteen stations over a domain encompassing the western Mediterranean basin were considered for a time period of 41 days in autumn, when heavy precipitation events most often plague this area. Prior to their assimilation, MWR data were compared to very-short-term forecasts. Observation-minus-background statistics revealed some biases, but standard deviations were comparable to that obtained with radiosondes. The MWR data were then assimilated in a three-dimensional variational data assimilation system through the use of a rapid update cycle. A first set of four different experiments were designed to assess the impact of the assimilation of temperature and humidity profiles, both separately and jointly. This assessment was done through the use of a comprehensive dataset of upper-air and surface observations collected in the framework of the HyMeX programme. The results showed that the impact was generally very limited on all verified parameters, except for precipitation. The impact was found to be generally beneficial in terms of most verification metrics for about 18 h, especially for larger accumulations. Two additional data-denial experiments showed that even more positive impact could be obtained when MWR data were assimilated without other redundant observations. The conclusion of the study points to possible ways of enhancing the impact of the assimilation of MWR data in convective-scale NWP systems.

Cole, R. E., Green, S. M., & Jardin, M. R. (2000). Improving RUC-1 Wind Estimates by Incorporating Near-Real-Time Aircraft Reports. *Weather and Forecasting, 15*(4), 447-460 https://doi.org/10.1175/1520-0434(2000)015<0447:IRWEBI>2.0.CO;2

A verification study of wind accuracy is presented for wind nowcasts generated by augmenting Rapid Update Cycle (RUC) wind forecasts with near-real-time aircraft reports using the Integrated Terminal Weather System (ITWS) gridded winds algorithm. Aircraft wind reports collected between the end of the RUC data collection interval and the time each RUC forecasts is valid are available for use in augmenting the RUC wind forecast to form a wind nowcast. The 60-km resolution, hourly RUG-I wind forecasts are used. ITWS-based nowcast wind errors and RUC forecast wind errors are examined statistically over a 1-yr dataset. The addition of the recent aircraft reports significantly reduces therms vector error and the 90th percentile vector error. Also reduced is the number of hours of sustained large errors and the correlation among errors. The errors increase with increasing wind speed, in part due to an underestimation of wind speed that increases with increasing wind speed. The errors in the augmented wind fields decrease with increasing numbers of Aircraft Communications Addressing and Reporting System reports. Different types of weather are also seen to influence wind field accuracy.

Coniglio, M. C. (2012). Verification of RUC 0-1-h Forecasts and SPC Mesoscale Analyses Using VORTEX2 Soundings. *Weather and Forecasting, 27*(3), 667-683 https://doi.org/10.1175/WAF-D-11-00096.1

This study uses radiosonde observations obtained during the second phase of the Verification of the Origins of Rotation in Tornadoes Experiment (VORTEX2) to verify base-state variables and severe-weather-related parameters calculated from Rapid Update Cycle (RUC) analyses and 1-h forecasts, as well as those calculated from the operational surface objective analysis system used at the Storm

Prediction Center (the SFCOA). The rapid growth in temperature, humidity, and wind errors from 0 to 1 h seen at all levels in a past RUC verification study by Benjamin et al. is not seen in the present study. This could be because the verification observations are also assimilated into the RUC in the Benjamin et al. study, whereas the verification observations in the present study are not. In the upper troposphere, the present study shows large errors in relative humidity, mostly related to a large moist bias. The planetary boundary layer tends to be too shallow in the RUC analyses and 1-h forecasts. Wind speeds tend to be too fast in the lowest 1 km and too slow in the 2-4-km layer. RUC and SECOA 1-h forecast errors for many important severe weather parameters are large relative to their potential impact on convective evolution. However, the SFCOA significantly improves upon the biases seen in most of the 1-h RUC forecasts for the base-state surface variables and most of the other severe-weather-related parameters, indicating that the SECOA has a more significant impact in reducing the biases in the 1-h RUC forecasts than on the root-mean-squared errors.

Horel, J. D., Colman, B., & Jackson, M. (2004). *Verification of NDFD Gridded Forecasts in the Western United States.* Paper presented at the 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction.

Validation of forecast grids over complex terrain remains a difficult problem. Initial attempts at verifying experimental forecasts over the western United States issued by National Weather Service forecasters using the Interactive Forecast Preparation System (IFPS) are presented. Forecast grids of temperature and wind issued during the Operational Readiness Demonstration period, 16 June - 15 July 2003, are evaluated and compared to objective analysis grids derived by the ARPS Data Assimilation System (ADAS). ADAS grids of surface temperature, wind, and relative humidity at horizontal resolutions of 10 km and 2.5 km reflect an adjustment of the Rapid Update Cycle gridded fields by local observations provided by MesoWest.

Myrick, D. T., & Horel, J. D. (2006). Verification of Surface Temperature Forecasts from the National Digital Forecast Database over the Western United States. *Weather and Forecasting, 21*(5), 869-892 https://doi.org/10.1175/WAF946.1

Experimental gridded forecasts of surface temperature issued by National Weather Service offices in the western United States during the 2003/04 winter season (18 November 2003-29 February 2004) are evaluated relative to surface observations and gridded analyses. The 5-km horizontal resolution gridded forecasts issued at 0000 UTC for forecast lead times at 12-h intervals from 12 to 168 h were obtained from the National Digital Forecast Database (NDFD). Forecast accuracy and skill are determined relative to observations at over 3000 locations archived by MesoWest. Forecast quality is also determined relative to Rapid Update Cycle (RUC) analyses at 20-km resolution that are interpolated to the 5-km NDFD grid as well as objective analyses obtained from the Advanced Regional Prediction System Data Assimilation System that rely upon the MesoWest observations and RUC analyses. For the West as a whole, the experimental temperature forecasts issued at 0000 UTC during the 2003/04 winter season exhibit skill at lead times of 12, 24, 36, and 48 h on the basis of several verification approaches. Subgrid-scale temperature variations and observational and analysis errors undoubtedly contribute some uncertainty regarding these results. Even though the "true" values appropriate to evaluate the forecast values on the NDFD grid are unknown, it is estimated that the root-mean-square errors of the NDFD temperature forecasts are on the order of 3 degrees C at lead times shorter than 48 h and greater than 4 degrees C at lead times longer than 120 h. However, such estimates are derived from only a small

fraction of the NDFD grid boxes. Incremental improvements in forecast accuracy as a result of forecaster adjustments to the 0000 UTC temperature grids from 144- to 24-h lead times are estimated to be on the order of 13%.

Schwartz, B., & Benjamin, S. (2004). *Observation Sensitivity Experiments Using the Rapid Update Cycle.* Paper presented at the Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface. Retrieved from https://ams.confex.com/ams/84Annual/techprogram/paper_71188.htm

There is currently much discussion within the meteorological community regarding the design and implementation of current and future observing systems, with particular emphasis on mesoscale observing systems. Assessing the relative value of observational platforms is useful for both scientific and budgetary interests. One way to evaluate the relative value in each observing system is to assess its contribution to the reduction of error in numerical weather prediction model forecasts. The Rapid Update Cycle (RUC), running operationally at the National Centers for Environmental Prediction (NCEP), is a good model to use for this evaluation because it assimilates a variety of asynoptic observations on an hourly basis. Using the RUC on its operational continental United States (CONUS) domain, we have run a series of experiments where various data sources where systematically removed from the model in an attempt to measure the relative contribution of each in reducing forecast error. The data sources included in our tests were rawinsonde, automated aircraft (ACARS), profilers, surface (METARS), and VAD winds. In addition, we have run the RUC with no data, other than that supplied indirectly through the lateral boundary conditions, as a "worse case" calibration. In this paper we discuss the results of retrospective tests using the RUC model for the period 4-13 January 2001. This is the same period used by NCEP for testing the most recent operational versions of both the ETA and RUC models. Data were denied from the RUC over various portions of the RUC domain, including an area that contains the operational NOAA profiler network. For all the experiments, including a control run that contains all possible data, average verification statistics for RUC wind, temperature, height, and relative humidity forecasts against rawinsonde observations were compiled for the test period. In addition to the average errors, we examine the largest errors at individual rawinsonde locations to identify how the absence of each data type elates to the occurrence of large error events associated with active weather periods. Statistical tests were performed for the difference between each denial experiment and the control run in order to access the significance of the results.

Schwartz, B. E., & Benjamin, S. G. (2002). *Verification of RUC Surface Forecasts at Major US Airport Hubs.* Paper presented at the 10th Conference on Aviation Meteorology, Portland, OR. Retrieved from https://www.researchgate.net/publication/242478649_99_Verification_of_RUC_Surface_Forecasts_at_Major_US_Airport_Hubs

The Rapid Update Cycle (RUC) model running operationally at the National Centers for Environmental Prediction (NCEP) provides high-frequency mesoscale analyses and short-range numerical weather prediction guidance for aviation, severe weather, and general weather forecasting. In spring of 2002, a new 20-km version of the RUC (henceforth referred to as the RUC20) will replace the operational 40-km version (RUC40) running at NCEP. In addition to higher horizontal resolution, the new version of the RUC features a cloud analysis scheme and various other enhancements that include improvements to cloud microphysics, land-surface, and convective parameterizations. In addition, the RUC20 contains more detailed specifications of topography, land-use and soil-type fields. These changes have led to

improvements in surface temperature, humidity, wind, and precipitation forecasts in the sample of forecasts examined by Schwartz and Benjamin (2001). Here, surface METAR observations are used to verify surface RUC40 and RUC20 3-h forecasts of 2 m temperature and 10 m wind speed at 27 major U.S. airport hubs

Smith, T. L., Benjamin, S. G., Gutman, S. I., & Sahm, S. (2007). Short-Range Forecast Impact from Assimilation of GPS-IPW Observations into the Rapid Update Cycle. *Monthly Weather Review, 135*(8), 2914-2930 https://doi.org/10.1175/MWR3436.1

Integrated precipitable water (IPW) estimates derived from time delays in the arrival of global positioning system (GPS) satellite signals are a relatively recent, high- frequency source of atmospheric moisture information available for real- time data assimilation. Different experimental versions of the Rapid Update Cycle (RUC) have assimilated these observations to assess GPS-IPW impact on moisture forecasts. In these tests, GPS-IPW data have proven to be a useful real- time source of moisture information, leading to more accurate short- range moisture forecasts when added to other observations. A multiyear experiment with parallel (one with GPS- IPW processed 24 h after the fact, one without) 3- h cycles using the original 60- km RUC was run from 1999 to 2004 with verification of each cycle against rawinsonde observations. This experiment showed a steady increase in the positive impact in short- range relative humidity (RH) forecasts due to the GPS- IPW data as the number of observing sites increased from 18 to almost 300 (as of 2004) across the United States and Canada. Positive impact from GPS-IPW on 850 - 700- hPa RH forecasts was also evident in 6- and 12- h forecasts. The impact of GPS-IPW data was also examined on forecasts from the more recent 20- km RUC, including a 1- h assimilation cycle and improved assimilation and physical parameterizations, now using real- time GPS-IPW retrievals available 30 min after valid time. In a 3- month comparison during the March - May 2004 period, 20-km RUC cycles with and without assimilation of GPS-IPW were compared with IPW for 3-, 6-, 9-, and 12-h forecasts. Using this measure, assimilation of GPS-IPW data led to the strongest improvements in the 3- and 6-h forecasts and smaller but still evident improvements in 9and 12- h forecasts. In a severe convective weather case, inclusion of GPS-IPW data improved forecasts of convective available potential energy, an important predictor of severe storm potential, and relative humidity. Positive impact from GPS-IPW assimilation was found to vary over season, geographical location, and time of day, apparently related to variations in vertical mixing. For example, GPS-IPW has a stronger effect on improving RH forecasts at 850 hPa at nighttime (than daytime) and in cooler seasons (than warmer seasons) when surface moisture observations are less representative of conditions aloft. As a result of these studies, assimilation of GPS-IPW was added to the operational RUC run at NOAA/ NCEP in June 2005 and to the operational North American Mesoscale model (also at NCEP) in June 2006 to improve their accuracy for short-range moisture forecasts.

Section V: North American Mesoscale Forecast System (NAM)

Charles, M. E., & Colle, B. A. (2009). Verification of Extratropical Cyclones within the NCEP Operational Models. Part I: Analysis Errors and Short-Term NAM and GFS Forecasts. *Weather and Forecasting, 24*(5), 1173-1190 https://doi.org/10.1175/2009WAF2222169.1

This paper verifies extratropical cyclones around North America and the adjacent oceans within the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) and North American Mesoscale (NAM) models during the 2002-07 cool seasons (October-March). The analyzed cyclones in

the Global Forecast System (GFS) model, North American Mesoscale (NAM) model, and the North American Regional Reanalysis (NARR) were also compared against sea level pressure (SLP) observations around extratropical cyclones. The GFS analysis of SLP was clearly superior to the NAM and NARR analyses. The analyzed cyclone pressures in the NAM improved in 2006-07 when its data assimilation was switched to the Gridpoint Statistical Interpolation (GSI). The NCEP GFS has more skillful cyclone intensity and position forecasts than the NAM over the continental United States and adjacent oceans, especially over the eastern Pacific, where the NAM has a large positive (underdeepening) bias in cyclone central pressure. For the short-term (0-60 h) forecasts, the GFS and NAM cyclone errors over the eastern Pacific are larger than the other regions to the east. There are relatively large biases in cyclone position for both models, which vary spatially around North America. The eastern Pacific has four to eight cyclone events per year on average, with errors > 10 mb at hour 48 in the GFS; this number has not decreased in recent years. There has been little improvement in the 0-2-day cyclone forecasts during the past 5 yr over the eastern United States, while there has been a relatively large improvement in the cyclone pressure predictions over the eastern Pacific in the NAM.

Charles, M. E., & Colle, B. A. (2009). Verification of Extratropical Cyclones within the NCEP Operational Models. Part II: The Short-Range Ensemble Forecast System. *Weather and Forecasting, 24*(5), 1191-1214 https://doi.org/10.1175/2009WAF2222170.1

This paper verifies the strengths and positions of extratropical cyclones around North America and the adjacent oceans within the Short Range Ensemble Forecast (SREF) system at the National Centers for Environmental Prediction (NCEP) during the 2004-07 cool seasons (October-March). The SREF mean for cyclone position and central pressure has a smaller error than the various subgroups within SREF and the operational North American Mesoscale (NAM) model in many regions on average, but not the operational Global Forecast System (GFS) for many forecast times. Inclusion of six additional Weather Research and Forecasting (WRF) model members into SREF during the 2006-07 cool season did not improve the SREF mean predictions. The SREF has slightly more probabilistic skill over the eastern United States and western Atlantic than the western portions of the domain for cyclone central pressure. The SREF also has slightly greater probabilistic skill than the combined GFS and NAM for central pressure, which is significant at the 90% level for many regions and thresholds. The SREF probabilities are fairly reliable, although the SREF is overconfident at higher probabilities in all regions. The inclusion of WRF did not improve the SREF probabilistic skill. Over the eastern Pacific, eastern Canada, and western Atlantic, the SREF is overdispersed on average, especially early in the forecast, while across the central and eastern United States the SREF is underdispersed later in the forecast. There are relatively large biases in cyclone central pressure within each SREF subgroup. As a result, the best-member diagrams reveal that the SREF members are not equally accurate for the cyclone central pressure and displacement. Two cases are presented to illustrate examples of SREF developing large errors early in the forecast for cyclones over the eastern United States.

Clark, A. J., Coniglio, M. C., Coffer, B. E., Thompson, G., Xue, M., & Kong, F. (2015). Sensitivity of 24-H Forecast Dryline Position and Structure to Boundary Layer Parameterizations in Convection-Allowing WRF Model Simulations. *Weather and Forecasting, 30*(3), 613-638 https://doi.org/10.1175/WAF-D-14-00078.1

Recent NOAA Hazardous Weather Testbed Spring Forecasting Experiments have emphasized the sensitivity of forecast sensible weather fields to how boundary layer processes are represented in the

Weather Research and Forecasting (WRF) Model. Thus, since 2010, the Center for Analysis and Prediction of Storms has configured at least three members of their WRF-based Storm-Scale Ensemble Forecast (SSEF) system specifically for examination of sensitivities to parameterizations of turbulent mixing, including the Mellor-Yamada-Janjic (MYJ); quasi-normal scale elimination (QNSE); Asymmetrical Convective Model, version 2 (ACM2); Yonsei University (YSU); and Mellor-Yamada-Nakanishi-Niino (MYNN) schemes (hereafter PBL members). In postexperiment analyses, significant differences in forecast boundary layer structure and evolution have been observed, and for preconvective environments MYNN was found to have a superior depiction of temperature and moisture profiles. This study evaluates the 24-h forecast dryline positions in the SSEF system PBL members during the period April-June 2010-12 and documents sensitivities of the vertical distribution of thermodynamic and kinematic variables in near-dryline environments. Main results include the following. Despite having superior temperature and moisture profiles, as indicated by a previous study, MYNN was one of the worst-performing PBL members, exhibiting large eastward errors in forecast dryline position. During April-June 2010-11, a dry bias in the North American Mesoscale Forecast System (NAM) initial conditions largely contributed to eastward dryline errors in all PBL members. An upgrade to the NAM and assimilation system in October 2011 apparently fixed the dry bias, reducing eastward errors. Large sensitivities of CAPE and low-level shear to the PBL schemes were found, which were largest between 1.0 degrees and 3.0 degrees to the east of drylines. Finally, modifications to YSU to decrease vertical mixing and mitigate its warm and dry bias greatly reduced eastward dryline errors.

Clark, A. J., Gallus, W. A., & Weisman, M. L. (2010). Neighborhood-Based Verification of Precipitation Forecasts from Convection-Allowing NCAR WRF Model Simulations and the Operational NAM. *Weather and Forecasting, 25*(5), 1495-1509 https://doi.org/10.1175/2010WAF2222404.1

Since 2003 the National Center for Atmospheric Research (NCAR) has been running various experimental convection-allowing configurations of the Weather Research and Forecasting Model (WRF) for domains covering a large portion of the central United States during the warm season (April-July). In this study, the skill of 3-hourly accumulated precipitation forecasts from a large sample of these convection-allowing simulations conducted during 2004-05 and 2007-08 is compared to that from operational North American Mesoscale (NAM) model forecasts using a neighborhood-based equitable threat score (ETS). Separate analyses were conducted for simulations run before and after the implementation in 2007 of positive-definite (PD) moisture transport for the NCAR-WRF simulations. The neighborhood-based ETS (denoted < ETS >(r)) relaxes the criteria for "hits" (i.e., correct forecasts) by considering grid points within a specified radius r. It is shown that < ETS >(r) is more useful than the traditional ETS because < ETS >(r) can be used to diagnose differences in precipitation forecast skill between different models as a function of spatial scale, whereas the traditional ETS only considers the spatial scale of the verification grid. It was found that differences in < ETS >(r) between NCAR-WRF and NAM generally increased with increasing r, with NCAR-WRF having higher scores. Examining time series of < ETS >(r) for r - 100 and r - 0 km (which simply reduces to the "traditional" ETS), statistically significant differences between NCAR-WRF and NAM were found at many forecast lead times for < ETS >(100) but only a few times for < ETS >(0). Larger and more statistically significant differences occurred with the 2007-08 cases relative to the 2004-05 cases. Because of differences in model configurations and dominant large-scale weather regimes, a more controlled experiment would have been needed to diagnose the reason for the larger differences that occurred with the 2007-08 cases. Finally, a compositing technique was used to diagnose the differences in the spatial distribution of the forecasts. This technique implied westward displacement errors for NAM model forecasts in both sets of cases and in NCAR-WRF model forecasts for the 2007-08 cases. Generally, the results are encouraging because

they imply that advantages in convection-allowing relative to convection-parameterizing simulations noted in recent studies are reflected in an objective neighborhood-based metric.

Elmore, K. L., Grams, H. M., Apps, D., & Reeves, H. D. (2015). Verifying Forecast Precipitation Type with mPING*. *Weather and Forecasting, 30*(3), 656-667 https://doi.org/10.1175/WAF-D-14-00068.1

In winter weather, precipitation type is a pivotal characteristic because it determines the nature of most preparations that need to be made. Decisions about how to protect critical infrastructure, such as power lines and transportation systems, and optimize how best to get aid to people are all fundamentally precipitation-type dependent. However, current understanding of the microphysical processes that govern precipitation type and how they interplay with physics-based numerical forecast models is incomplete, degrading precipitation-type forecasts, but by how much? This work demonstrates the utility of crowd-sourced surface observations of precipitation type from the Meteorological Phenomena Identification Near the Ground (mPING) project in estimating the skill of numerical model precipitation-type forecasts and, as an extension, assessing the current model performance regarding precipitation type in areas that are otherwise without surface observations. In general, forecast precipitation type is biased high for snow and rain and biased low for freezing rain and ice pellets. For both the North American Mesoscale Forecast System and Global Forecast System models, Gilbert skill scores are between 0.4 and 0.5 and from 0.35 to 0.45 for the Rapid Refresh model, depending on lead time. Peirce skill scores for individual precipitation types are 0.7-0.8 for both rain and snow, 0.2-0.4 for freezing rain and freezing rain, and 0.25 or less for ice pellets. The RAPid Refresh model displays somewhat lower scores except for ice pellets, which are severely underforecast, compared to the other models.

Evans, C., Weiss, S. J., Jirak, I. L., Dean, A. R., & 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 https://doi.org/10.1175/WAF-D-18-0124.1

This study evaluates forecast vertical thermodynamic profiles and derived thermodynamic parameters from two regional/convection-allowing model pairs, the North American Mesoscale Forecast System and the North American Mesoscale Nest model pair and the Rapid Refresh and High Resolution Rapid Refresh model pair, in warm-season, thunderstorm-supporting environments. Differences in bias and mean absolute error between the regional and convection-allowing models in each of the two pairs, while often statistically significant, are practically small for the variables, parameters, and vertical levels considered, such that the smaller-scale variability resolved by convection-allowing models does not degrade their forecast skill. Model biases shared by the regional and convection-allowing models in each pair are documented, particularly the substantial cool and moist biases in the planetary boundary layer arising from the Mellor-Yamada-Janji planetary boundary layer parameterization used by the North American Mesoscale model and the Nest version as well as the middle-tropospheric moist bias shared by the Rapid Refresh and High Resolution Rapid Refresh models. Bias and mean absolute errors typically have larger magnitudes in the evening, when buoyancy is a significant contributor to turbulent vertical mixing, than in the morning. Vertical thermodynamic profile biases extend over a deep vertical layer in the western United States given strong sensible heating of the underlying surface. The results suggest that convection-allowing models can fulfill the use cases typically and historically met by regional models in operations at forecast entities such as the Storm Prediction Center, a fruitful finding given the proposed elimination of regional models with the Next-Generation Global Prediction System initiative.

Gowan, T. M., Steenburgh, W. J., & Schwartz, C. S. (2018). Validation of Mountain Precipitation Forecasts from the Convection-Permitting NCAR Ensemble and Operational Forecast Systems over the Western United States. *Weather and Forecasting, 33*(3), 739-765 https://doi.org/10.1175/WAF-D-17-0144.1

Convection-permitting ensembles can capture the large spatial variability and quantify the inherent uncertainty of precipitation in areas of complex terrain; however, such systems remain largely untested over the western United States. In this study, we assess the capabilities of deterministic and probabilistic cool-season quantitative precipitation forecasts (QPFs) produced by the 10-member, convection-permitting (3-km horizontal grid spacing) NCAR Ensemble using observations collected by SNOTEL stations at mountain locations across the western United States and precipitation analyses from PRISM. We also examine the performance of operational forecast systems run by NCEP including the High Resolution Rapid Refresh (HRRR) model, the NAM forecast system with a 3-km continental United States (CONUS) nest, GFS, and the Short-Range Ensemble Forecast system (SREF). Overall, we find that higher-resolution models, such as the HRRR, NAM-3km CONUS nest, and an individual member of the NCAR Ensemble, are more deterministically skillful than coarser models, especially over the narrow interior ranges of the western United States, likely because they better resolve topography and thus better simulate orographic precipitation. The 10-member NCAR Ensemble is also more probabilistically skillful than 13-member subensembles composed of each SREF dynamical core, but less probabilistically skillful than the full 26-member SREF, as a result of insufficient spread. These results should help guide future short-range model development and inform forecasters about the capabilities and limitations of several widely used deterministic and probabilistic modeling systems over the western United States.

Herman, G. R., & Schumacher, R. S. (2016). Extreme Precipitation in Models: An Evaluation. *Weather and Forecasting, 31*(6), 1853-1879 https://doi.org/10.1175/WAF-D-16-0093.1

A continental United States (CONUS)-wide framework for analyzing quantitative precipitation forecasts (QPFs) from NWP models from the perspective of precipitation return period (RP) exceedances is introduced using threshold estimates derived from a combination of NOAA Atlas 14 and older sources. Forecasts between 2009 and 2015 from several different NWP models of varying configurations and spatial resolutions are analyzed to assess bias characteristics and forecast skill for predicting RP exceedances. Specifically, NOAA's Global Ensemble Forecast System Reforecast (GEFS/R) and the National Severe Storms Laboratory WRF (NSSL-WRF) model are evaluated for 24-h precipitation accumulations. The climatology of extreme precipitation events for 6-h accumulations is also explored in three convection-allowing models: 1) NSSL-WRF, 2) the North American Mesoscale 4-km nest (NAM-NEST), and 3) the experimental High Resolution Rapid Refresh (HRRR). The GEFS/R and NSSL-WRF are both found to exhibit similar 24-h accumulation RP exceedance climatologies over the U.S. West Coast to those found in observations and are found to be approximately equally skillful at predicting these exceedance events in this region. In contrast, over the eastern two-thirds of the CONUS, GEFS/R struggles to predict the predominantly convectively driven extreme QPFs, predicting far fewer events than are observed and exhibiting inferior forecast skill to the NSSL-WRF. The NSSL-WRF and HRRR are found to produce 6-h extreme precipitation climatologies that are approximately in accord with those found in the observations, while NAM-NEST produces many more RP exceedances than are observed across all of the CONUS.

Novak, D. R., Bailey, C., Brill, K. F., Burke, P., Hogsett, W. A., Rausch, R., & Schichtel, M. (2014). Precipitation and Temperature Forecast Performance at the Weather Prediction Center. *Weather and Forecasting, 29*(3), 489-504 https://doi.org/10.1175/WAF-D-13-00066.1

The role of the human forecaster in improving upon the accuracy of numerical weather prediction is explored using multiyear verification of human-generated short-range precipitation forecasts and medium-range maximum temperature forecasts from the Weather Prediction Center (WPC). Results show that human-generated forecasts improve over raw deterministic model guidance. Over the past two decades, WPC human forecasters achieved a 20%-40% improvement over the North American Mesoscale (NAM) model and the Global Forecast System (GFS) for the 1 in. (25.4 mm) (24 h)(-1) threshold for day 1 precipitation forecasts, with a smaller, but statistically significant, 5%-15% improvement over the deterministic ECMWF model. Medium-range maximum temperature forecasts also exhibit statistically significant improvement over GFS model output statistics (MOS), and the improvement has been increasing over the past 5 yr. The quality added by humans for forecasts of high-impact events varies by element and forecast projection, with generally large improvements when the forecaster makes changes >= 8 degrees F (4.4 degrees C) to MOS temperatures. Human improvement over guidance for extreme rainfall events [3 in. (76.2 mm) (24 h)(-1)] is largest in the short-range forecast. However, human-generated forecasts failed to outperform the most skillful downscaled, bias-corrected ensemble guidance for precipitation and maximum temperature available near the same time as the human-modified forecasts. Thus, as additional downscaled and bias-corrected sensible weather element guidance becomes operationally available, and with the support of near-real-time verification, forecaster training, and tools to guide forecaster interventions, a key test is whether forecasters can learn to make statistically significant improvements over the most skillful of this guidance. Such a test can inform to what degree, and just how quickly, the role of the forecaster changes.

Van Thien, L., Gallus, W. A., Olsen, M. A., & Livesey, N. (2010). Comparison of Aura MLS Water Vapor Measurements with GFS and NAM Analyses in the Upper Troposphere-Lower Stratosphere. *Journal of Atmospheric and Oceanic Technology, 27*(2), 274-289 https://doi.org/10.1175/2009JTECHA1317.1

Water vapor mixing ratios in the upper troposphere and lower stratosphere measured by the Aura Microwave Limb Sounder (MLS) version 2.2 instrument have been compared with Global Forecast System(GFS) analyses at five levels within the 300-100-hPa layer and North American Mesoscale (NAM) model analyses at six levels within the 300-50-hPa layer over the two years of 2005 and 2006 at four analysis times (e.g., 0000, 0600, 1200, and 1800 UTC). Probability density functions of the vapor mixing ratios suggest that both analyses are often moister than Aura MLS values, but NAM model analyses agree somewhat better with Aura MLS measurements than GFS model analyses over the same North American domain at the five common levels. Examining five subsets of the global GFS domain, the GFS model analysis is moister than Aura MLS estimates everywhere but at 150 and 100 hPa in all regions outside of the tropics. NAM model analysis water vapor mixing ratios exceeded the Aura MLS values at all levels from 250 to 150 hPa in all four seasons of both years and some seasons at 100 and 50 hPa. Moist biases in winter and spring of both years were similar at all levels, but these moist biases in summer and fall were smaller in 2005 than in 2006 at all levels. These differences may be due to the change in the NAM from using the ETA Model to using the Weather Research and Forecasting model (WRF) in June 2006.

Yan, H., & Gallus, W. A. (2016). An Evaluation of QPF from the WRF, NAM, and GFS Models Using Multiple Verification Methods over a Small Domain. *Weather and Forecasting, 31*(4), 1363-1379 https://doi.org/10.1175/WAF-D-16-0020.1

The ARW model was run over a small domain centered on Iowa for 9 months with 4-km grid spacing to better understand the limits of predictability of short-term (12 h) quantitative precipitation forecasts (QPFs) that might be used in hydrology models. Radar data assimilation was performed to reduce spinup problems. Three grid-to-grid verification methods, as well as two spatial techniques, neighborhood and object based, were used to compare the QPFs from the high-resolution runs with coarser operational GFS and NAM QPFs to verify QPFs for various precipitation accumulation intervals and on two grid configurations with different resolutions. In general, NAM had the worst performance not only for model skill but also for spatial feature attributes as a result of the existence of large dry bias and location errors. The finer resolution of NAM did not offer any advantage in predicting small-scale storms compared to the coarser GFS. WRF had a large advantage for high precipitation thresholds. A greater improvement in skill was noted when the accumulation time interval was increased, compared to an increase in the spatial neighborhood size. At the same neighborhood scale, the high-resolution WRF Model was less influenced by the grid on which the verification was done than the other two models. All models had the highest skill from midnight to early morning, because the least wet bias, location, and coverage errors were present then. The lowest skill was shown from late morning through afternoon. The main cause of poor skill during this period was large displacement errors.

Section VI: Rapid Refresh Model (RAP)

Benjamin, S. G., Brown, J. M., & Smirnova, T. G. (2016). Explicit Precipitation-Type Diagnosis from a Model Using a Mixed-Phase Bulk Cloud-Precipitation Microphysics Parameterization. *Weather and Forecasting, 31*(2), 609-619 https://doi.org/10.1175/WAF-D-15-0136.1

The Rapid Refresh (RAP) and High-Resolution Rapid Refresh (HRRR), both operational at NOAA's National Centers for Environmental Prediction (NCEP) use the Thompson et al. mixed-phase bulk cloud microphysics scheme. This scheme permits predicted surface precipitation to simultaneously consist of rain, snow, and graupel at the same location under certain conditions. Here, the explicit precipitation-type diagnostic method is described as used in conjunction with the Thompson et al. scheme in the RAP and HRRR models. The postprocessing logic combines the explicitly predicted multispecies hydrometeor data and other information from the model forecasts to produce fields of surface precipitation type that distinguish between rain and freezing rain, and to also portray areas of mixed precipitation. This explicit precipitation-type diagnostic method is used with the NOAA operational RAP and HRRR models. Verification from two winter seasons from 2013 to 2015 is provided against METAR surface observations. An example of this product from a January 2015 south-central United States winter storm is also shown.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., . . . Manikin, G. S. (2016). A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Monthly Weather Review, 144*(4), 1669-1694 https://doi.org/10.1175/MWR-D-15-0242.1

The Rapid Refresh (RAP), an hourly updated assimilation and model forecast system, replaced the Rapid Update Cycle (RUC) as an operational regional analysis and forecast system among the suite of models at the NOAA/National Centers for Environmental Prediction (NCEP) in 2012. The need for an effective hourly updated assimilation and modeling system for the United States for situational awareness and related decision-making has continued to increase for various applications including aviation (and transportation in general), severe weather, and energy. The RAP is distinct from the previous RUC in three primary aspects: a larger geographical domain (covering North America), use of the community-based Advanced Research version of the Weather Research and Forecasting (WRF) Model (ARW) replacing the RUC forecast model, and use of the Gridpoint Statistical Interpolation analysis system (GSI) instead of the RUC three-dimensional variational data assimilation (3DVar). As part of the RAP development, modifications have been made to the community ARW model (especially in model physics) and GSI assimilation systems, some based on previous model and assimilation design innovations developed initially with the RUC. Upper-air comparison is included for forecast verification against both rawinsondes and aircraft reports, the latter allowing hourly verification. In general, the RAP produces superior forecasts to those from the RUC, and its skill has continued to increase from 2012 up to RAP version 3 as of 2015. In addition, the RAP can improve on persistence forecasts for the 1-3-h forecast range for surface, upper-air, and ceiling forecasts.

Burg, T., Elmore, K. L., & Grams, H. M. (2017). Assessing the Skill of Updated Precipitation-Type Diagnostics for the Rapid Refresh with mPING. *Weather and Forecasting, 32*(2), 725-732 https://doi.org/10.1175/WAF-D-16-0132.1

Previous work has shown that the Rapid Refresh (RAP) model severely underrepresents ice pellets in its grid, with a skill near zero and a very low bias. An ice pellet diagnostic upgrade was devised at the Earth System Research Laboratory (ESRL) to resolve this issue. Parallel runs of the experimental ESRL-RAP with the fix and the operational NCEP-RAP without the fix provide an opportunity to assess whether this upgrade has improved the overall performance and the performance of the individual precipitation types of the ESRL-RAP. Verification was conducted using the mobile Phenomena Identification Near the Ground (mPING) project. The overall Gerrity skill score (GSS) for the ESRL-RAP was improved relative to the NCEP-RAP at a 3-h lead time but degraded with increasing lead time; the difference is significant at $p < 0.05$. Whether this difference is practically significant for users is unknown. Some improvement was found in the bias and skill scores of ice pellets and snow in the ESRL-RAP, although the model continues to underrepresent ice pellets, while rain and freezing rain were generally the same or slightly worse with the fix. The ESRL-RAP was also found to depict a more realistic spatial distribution of precipitation types in transition zones involving ice pellets and freezing rain.

James, E. P., & Benjamin, S. G. (2017). Observation System Experiments with the Hourly Updating Rapid Refresh Model Using GSI Hybrid Ensemble-Variational Data Assimilation. *Monthly Weather Review, 145*(8), 2897-2918 https://doi.org/10.1175/MWR-D-16-0398.1

A set of observation system experiments (OSEs) over three seasons using the hourly updated Rapid Refresh (RAP) numerical weather prediction (NWP) assimilation-forecast system identifies the importance of the various components of the North American observing system for 3-12-h RAP forecasts. Aircraft observations emerge as the strongest-impact observation type for wind, relative humidity (RH), and temperature forecasts, permitting a 15%-30% reduction in 6-h forecast error in the troposphere and lower stratosphere. Major positive impacts are also seen from rawinsondes, GOES

satellite cloud observations, and surface observations, with lesser but still significant impacts from GPS precipitable water (PW) observations, satellite atmospheric motion vectors (AMVs), and radar reflectivity observations. A separate experiment revealed that the aircraft-related RH forecast improvement was augmented by 50% due specifically to the addition of aircraft moisture observations. Additionally, observations from en route aircraft and those from ascending or descending aircraft contribute approximately equally to the overall forecast skill, with the strongest impacts in the respective layers of the observations. Initial results from these OSEs supported implementation of an improved assimilation configuration of boundary layer pseudoinnovations from surface observations, as well as allowing the assimilation of satellite AMVs over land. The breadth of these experiments over the three seasons suggests that observation impact results are applicable to general forecasting skill, not just classes of phenomena during limited time periods.

Lin, H., Weygandt, S. S., Benjamin, S. G., & Hu, M. (2017). Satellite Radiance Data Assimilation within the Hourly Updated Rapid Refresh. *Weather and Forecasting, 32*(4), 1273-1287 https://doi.org/10.1175/WAF-D-16-0215.1

Assimilation of satellite radiance data in limited-area, rapidly updating weather model/assimilation systems poses unique challenges compared to those for global model systems. Principal among these is the severe data restriction posed by the short data cutoff time. Also, the limited extent of the model domain reduces the spatial extent of satellite data coverage and the lower model top of regional models reduces the spectral usage of radiance data especially for infrared data. These three factors impact the quality of the feedback to the bias correction procedures, making the procedures potentially less effective. Within the National Oceanic and Atmospheric Administration (NOAA) Rapid Refresh (RAP) hourly updating prediction system, satellite radiance data are assimilated using the standard procedures within the Gridpoint Statistical Interpolation (GSI) analysis package. Experiments for optimizing the operational implementation of radiance data into RAP and for improving benefits of radiance data have been performed. The radiance data impact for short-range forecasts has been documented to be consistent and statistically significantly positive in systematic RAP retrospective runs using real-time datasets. The radiance data impact has also been compared with conventional observation datasets within RAP. The configuration for RAP satellite radiance assimilation evaluated here is that implemented at the National Centers for Environmental Prediction (NCEP) in August 2016.

Lin, H., Weygandt, S. S., Lim, A. H. N., Hu, M., Brown, J. M., & Benjamin, S. G. (2017). Radiance Preprocessing for Assimilation in the Hourly Updating Rapid Refresh Mesoscale Model: A Study Using AIRS Data. *Weather and Forecasting, 32*(5), 1781-1800 https://doi.org/10.1175/WAF-D-17-0028.1

This study describes the initial application of radiance bias correction and channel selection in the hourly updated RAPid Refresh model. For this initial application, data from the Atmospheric Infrared Sounder (AIRS) are used; this dataset gives atmospheric temperature and water vapor information at higher vertical resolution and accuracy than previously launched low-spectral resolution satellite systems. In this preliminary study, data from AIRS are shown to add skill to short-range weather forecasts over a relatively data-rich area. Two 1-month retrospective runs were conducted to evaluate the impact of assimilating clear-sky AIRS radiance data on 1-12-h forecasts using a research version of the National Oceanic and Atmospheric Administration (NOAA) Rapid Refresh (RAP) regional mesoscale model already assimilating conventional and other radiance [AMSU-A, Microwave Humidity Sounder (MHS), HIRS-4]

data. Prior to performing the assimilation, a channel selection and bias-correction spinup procedure was conducted that was appropriate for the RAP configuration. RAP forecasts initialized from analyses with and without AIRS data were verified against radiosonde, surface atmosphere, precipitation, and satellite radiance observations. Results show that the impact from AIRS radiance data on short-range forecast skill in the RAP system is small but positive and statistically significant at the 95% confidence level. The RAP-specific channel selection and bias correction procedures described in this study were the basis for similar applications to other radiance datasets now assimilated in version 3 of RAP implemented at NOAA's National Centers for Environmental Prediction (NCEP) in August 2016.

Pan, Y., Zhu, K., Xue, M., Wang, X., Hu, M., Benjamin, S. G., . . . Whitaker, J. S. (2014). A GSI-Based Coupled EnSRF-En3DVar Hybrid Data Assimilation System for the Operational Rapid Refresh Model: Tests at a Reduced Resolution. *Monthly Weather Review, 142*(10), 3756-3780 https://doi.org/10.1175/MWR-D-13-00242.1

Acoupled ensemble square root filter-three-dimensional ensemble-variational hybrid (EnSRF-En3DVar) data assimilation (DA) system is developed for the operational Rapid Refresh (RAP) forecasting system. The En3DVar hybrid system employs the extended control variable method, and is built on the NCEP operational gridpoint statistical interpolation (GSI) three-dimensional variational data assimilation (3DVar) framework. It is coupled with an EnSRF system for RAP, which provides ensemble perturbations. Recursive filters (RF) are used to localize ensemble covariance in both horizontal and vertical within the En3DVar. The coupled En3DVar hybrid system is evaluated with 3-h cycles over a 9-day period with active convection. All conventional observations used by operational RAP are included. The En3DVar hybrid system is run at 1/3 of the operational RAP horizontal resolution or about 40-km grid spacing, and its performance is compared to parallel GSI 3DVar and EnSRF runs using the same datasets and resolution. Short-term forecasts initialized from the 3-hourly analyses are verified against sounding and surface observations. When using equally weighted static and ensemble background error covariances and 40 ensemble members, the En3DVar hybrid system outperforms the corresponding GSI 3DVar and EnSRF. When the recursive filter coefficients are tuned to achieve a similar height-dependent localization as in the EnSRF, the En3DVar results using pure ensemble covariance are close to EnSRF. Two-way coupling between EnSRF and En3DVar did not produce noticeable improvement over one-way coupling. Downscaled precipitation forecast skill on the 13-km RAP grid from the En3DVar hybrid is better than those from GSI 3DVar analyses.

Section VII: High Resolution Rapid Refresh Model (HRRR)

Bytheway, J. L., & Kummerow, C. D. (2015). Toward an Object-Based Assessment of High-Resolution Forecasts of Long-Lived Convective Precipitation in the Central US. *Journal of Advances in Modeling Earth Systems, 7*(3), 1248-1264 https://doi.org/10.1002/2015MS000497

Forecast models have seen vast improvements in recent years, via both increased resolutions and the ability to assimilate observational data, particularly that which has been affected by clouds and precipitation. The High-Resolution Rapid Refresh (HRRR) model is an hourly updated, 3 km model designed for forecasting convective precipitation recently deployed for operational use over the U.S. that initializes latent heating profiles as a function of assimilated radar reflectivity. An object-oriented verification process was developed to validate experimental HRRR convective precipitation forecasts during the 2013 warm season using the NCEP Stage IV multisensor precipitation product. A database of

467 convective precipitation features that were observed during the forecast assimilation period and their corresponding HRRR forecast precipitation features was created. This database was used to evaluate model performance over the entire forecast period, and to relate that performance to model processes, especially those related to precipitation production. Generally, HRRR precipitation is located within 30 km of the observed throughout the forecast period. Validation statistics are best at forecast hour 3, with median biases in mean, maximum, and total rainfall and raining area near 0%. Earlier in the forecast, median biases in the mean and maximum rain rate exceed 30%, with bias values often exceeding 150%. The median bias in areal extent at the beginning of the forecast is near -40%. This low areal bias and POD values <0.6 appear to be related to the model's ability to produce deep convection relative to atmospheric moisture content and concentration of rainfall in convective cores.

Bytheway, J. L., & Kummerow, C. D. (2018). Consistency between Convection Allowing Model Output and Passive Microwave Satellite Observations. *Journal of Geophysical Research-Atmospheres, 123*(2), 1065-1078 https://doi.org/10.1002/2017JD027527

Observations from the Global Precipitation Measurement (GPM) core satellite were used along with precipitation forecasts from the High Resolution Rapid Refresh (HRRR) model to assess and interpret differences between observed and modeled storms. Using a feature-based approach, precipitating objects were identified in both the National Centers for Environmental Prediction Stage IV multisensor precipitation product and HRRR forecast at lead times of 1, 2, and 3h at valid times corresponding to GPM overpasses. Precipitating objects were selected for further study if (a) the observed feature occurred entirely within the swath of the GPM Microwave Imager (GMI) and (b) the HRRR model predicted it at all three forecast lead times. Output from the HRRR model was used to simulate microwave brightness temperatures (Tbs), which were compared to those observed by the GMI. Simulated Tbs were found to have biases at both the warm and cold ends of the distribution, corresponding to the stratiform/anvil and convective areas of the storms, respectively. Several experiments altered both the simulation microphysics and hydrometeor classification in order to evaluate potential shortcomings in the model's representation of precipitating clouds. In general, inconsistencies between observed and simulated brightness temperatures were most improved when transferring snow water content to supercooled liquid hydrometeor classes.

Cai, H., & Dumais, R. E. (2015). Object-Based Evaluation of a Numerical Weather Prediction Model's Performance through Forecast Storm Characteristic Analysis. *Weather and Forecasting, 30*(6), 1451-1468 https://doi.org/10.1175/WAF-D-15-0008.1

Traditional pixel-versus-pixel forecast evaluation scores such as the critical success index (CSI) provide a simple way to compare the performances of different forecasts; however, they offer little information on how to improve a particular forecast. This paper strives to demonstrate what additional information an object-based forecast evaluation tool such as the Method for Object-Based Diagnostic Evaluation (MODE) can provide in terms of assessing numerical weather prediction models' convective storm forecasts. Forecast storm attributes evaluated by MODE in this paper include storm size, intensity, orientation, aspect ratio, complexity, and number of storms. Three weeks of the High Resolution Rapid Refresh (HRRR) model's precipitation forecasts during the summer of 2010 over the eastern two-thirds of the contiguous United States were evaluated as an example to demonstrate the methodology. It is found that the HRRR model was able to forecast convective storm characteristics rather well either as a function of time of day or as a function of storm size, although significant bias does exist, especially in

terms of storm number and storm size. Another interesting finding is that the model's ability of forecasting new storm initiation varies substantially by regions, probably as a result of its different skills in forecasting convection driven by different forcing mechanisms (i.e., diurnal heating vs synoptic-scale frontal systems).

Glahn, B., Schnapp, A. D., Ghirardelli, J. E., & Im, J.-S. (2017). A LAMP-HRRR MELD for Improved Aviation Guidance. *Weather and Forecasting, 32*(2), 391-405 https://doi.org/10.1175/WAF-D-16-0127.1

Localized Aviation MOS Program (LAMP) forecasts of ceiling height, visibility, wind, and other weather elements of interest to the aviation community have been produced and put into the National Digital Guidance Database (NDGD) since 2006. The High Resolution Rapid Refresh (HRRR) model is now producing explicit forecasts of ceiling height and visibility computed by algorithms based on variables directly forecasted by the HRRR. The Meteorological Development Laboratory has improved the LAMP ceiling and visibility forecasts by combining these two sources of information into a LAMP-HRRR MELD. The new forecasts show improvement over the original LAMP and particularly over the HRRR and persistence in terms of bias, threat score, and the Gerrity score. This paper explains how the MELD is produced and shows selected verification and example forecasts. A new guidance product based on this work will be made available to partners and customers.

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., & Alexander, C. R. (2017). Methods for Comparing Simulated and Observed Satellite Infrared Brightness Temperatures and What Do They Tell Us? *Weather and Forecasting, 32*(1), 5-25 https://doi.org/10.1175/WAF-D-16-0098.1

In this study, the utility of dimensioned, neighborhood-based, and object-based forecast verification metrics for cloud verification is assessed using output from the experimental High Resolution RAPid Refresh (HRRRx) model over a 1-day period containing different modes of convection. This is accomplished by comparing observed and simulated Geostationary Operational Environmental Satellite (GOES) 10.7-mu m brightness temperatures (BTs). Traditional dimensioned metrics such as mean absolute error (MAE) and mean bias error (MBE) were used to assess the overall model accuracy. The MBE showed that the HRRRx BTs for forecast hours 0 and 1 are too warm compared with the observations, indicating a lack of cloud cover, but rapidly become too cold in subsequent hours because of the generation of excessive upper-level cloudiness. Neighborhood and object-based statistics were used to investigate the source of the HRRRx cloud cover errors. The neighborhood statistic fractions skill score (FSS) showed that displacement errors between cloud objects identified in the HRRRx and GOES BTs increased with time. Combined with the MBE, the FSS distinguished when changes in MAE were due to differences in the HRRRx BT bias or displacement in cloud features. The Method for Object-Based Diagnostic Evaluation (MODE) analyzed the similarity between HRRRx and GOES cloud features in shape and location. The similarity was summarized using the newly defined MODE composite score (MCS), an area-weighted calculation using the cloud feature match value from MODE. Combined with the FSS, the MCS indicated if HRRRx forecast error is the result of cloud shape, since the MCS is moderately large when forecast and observation objects are similar in size.

Griffin, S. M., Otkin, J. A., Rozoff, C. M., Sieglaff, J. M., Cronce, L. M., Alexander, C. R., . . . 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 https://doi.org/10.1175/JAMC-D-17-0004.1

In this study, object-based verification using the method for object-based diagnostic evaluation(MODE) is used to assess the accuracy of cloud-cover forecasts from the experimental High-Resolution Rapid Refresh (HRRRx) model during the warm and cool seasons. This is accomplished by comparing cloud objects identified by MODE in observed and simulated Geostationary Operational Environmental Satellite 10.7-mm brightness temperatures for August 2015 and January 2016. The analysis revealed that more cloud objects and a more pronounced diurnal cycle occurred during August, with larger object sizes observed in January because of the prevalence of synoptic-scale cloud features. With the exception of the 0-h analyses, the forecasts contained fewer cloud objects than were observed. HRRRx forecast accuracy is assessed using two methods: traditional verification, which compares the locations of grid points identified as observation and forecast objects, and the MODE composite score, an area-weighted calculation using the object-pair interest values computed by MODE. The 1-h forecasts for both August and January were the most accurate for their respective months. Inspection of the individual MODE attribute interest scores showed that, even though displacement errors between the forecast and observation objects increased between the 0-h analyses and 1-h forecasts, the forecasts were more accurate than the analyses because the sizes of the largest cloud objects more closely matched the observations. The 1-h forecasts from August were found to be more accurate than those during January because the spatial displacement between the cloud objects was smaller and the forecast objects better represented the size of the observation objects.

Hwang, Y., Clark, A. J., Lakshmanan, V., & Koch, S. E. (2015). Improved Nowcasts by Blending Extrapolation and Model Forecasts. *Weather and Forecasting, 30*(5), 1201-1217 https://doi.org/10.1175/WAF-D-15-0057.1

Planning and managing commercial airplane routes to avoid thunderstorms requires very skillful and frequently updated 0-8-h forecasts of convection. The National Oceanic and Atmospheric Administration's High-Resolution Rapid Refresh (HRRR) model is well suited for this purpose, being initialized hourly and providing explicit forecasts of convection out to 15 h. However, because of difficulties with depicting convection at the time of model initialization and shortly thereafter (i.e., during model spinup), relatively simple extrapolation techniques, on average, perform better than the HRRR at 0-2-h lead times. Thus, recently developed nowcasting techniques blend extrapolation-based forecasts with numerical weather prediction (NWP)-based forecasts, heavily weighting the extrapolation forecasts at 0-2-h lead times and transitioning emphasis to the NWP-based forecasts at the later lead times. In this study, a new approach to applying different weights to blend extrapolation and model forecasts based on intensities and forecast times is applied and tested. An image-processing method of morphing between extrapolation and model forecasts to create nowcasts is described and the skill is compared to extrapolation forecasts and forecasts from the HRRR. The new approach is called salient cross dissolve (Sal CD), which is compared to a commonly used method called linear cross dissolve (Lin CD). Examinations of forecasts and observations of the maximum altitude of echo-top heights 18 dBZ and measurement of forecast skill using neighborhood-based methods shows that Sal CD significantly improves upon Lin CD, as well as the HRRR at 2-5-h lead times.

Ikeda, K., Steiner, M., Pinto, J., & Alexander, C. (2013). Evaluation of Cold-Season Precipitation Forecasts Generated by the Hourly Updating High-Resolution Rapid Refresh Model. *Weather and Forecasting, 28*(4), 921-939 https://doi.org/10.1175/WAF-D-12-00085.1

The hourly updating High-Resolution Rapid Refresh (HRRR) model is evaluated with regard to its ability to predict the areal extent of cold-season precipitation and accurately depict the timing and location of regions of snow, rain, and mixed-phase precipitation on the ground. Validation of the HRRR forecasts is performed using observations collected by the Automated Surface Observing System (ASOS) stations across the eastern two-thirds of the United States during the 2010-11 cold season. The results show that the HRRR is able to reliably forecast precipitation extent during the cold season. In particular, the location and areal extent of both snow and rain are very well predicted. Depiction of rain-to-snow transitions and freezing rain is reasonably good; however, the associated evaluation scores are significantly lower than for either snow or rain. The analyses suggest the skill in accurately depicting precipitation extent and phase (i.e., rain, snow, and mixed phase) depends on the size and organization of a weather system. Typically, larger synoptically forced weather systems are better predicted than smaller weather systems, including the associated rain-to-snow transition or freezing-rain areas. Offsets in space or time (i.e., causing misses and false alarms) have a larger effect on the model performance for smaller weather systems.

Ikeda, K., Steiner, M., & Thompson, G. (2017). Examination of Mixed-Phase Precipitation Forecasts from the High-Resolution Rapid Refresh Model Using Surface Observations and Sounding Data. *Weather and Forecasting, 32*(3), 949-967 https://doi.org/10.1175/WAF-D-16-0171.1

Accurate prediction of mixed-phase precipitation remains challenging for numerical weather prediction models even at high resolution and with a sophisticated explicit microphysics scheme and diagnostic algorithm to designate the surface precipitation type. Since mixed-phase winter weather precipitation can damage infrastructure and produce significant disruptions to air and road travel, incorrect surface precipitation phase forecasts can have major consequences for local and statewide decision-makers as well as the general public. Building upon earlier work, this study examines the High-Resolution Rapid Refresh (HRRR) model's ability to forecast the surface precipitation phase, with a particular focus on model-predicted vertical temperature profiles associated with mixed-phase precipitation, using upper-air sounding observations as well as the Automated Surface Observing Systems (ASOS) and Meteorological Phenomena Identification Near the Ground (mPING) observations. The analyses concentrate on regions of mixed-phase precipitation from two winter season events. The results show that when both the observational and model data indicated mixed-phase precipitation at the surface, the model represents the observed temperature profile well. Overall, cases where the model predicted rain but the observations indicated mixed-phase precipitation generally show a model surface temperature bias of <2 degrees C and a vertical temperature profile similar to the sounding observations. However, the surface temperature bias was similar to 4 degrees C in weather systems involving cold-air damming in the eastern United States, resulting in an incorrect surface precipitation phase or the duration (areal coverage) of freezing rain being much shorter (smaller) than the observation. Cases with predicted snow in regions of observed mixed-phase precipitation present subtle difference in the elevated layer with temperatures near 0 degrees C and the near-surface layer.

McCorkle, T. A., Horel, J. D., Jacques, A. A., & Alcott, T. (2018). Evaluating the Experimental High-Resolution Rapid Refresh–Alaska Modeling System Using USArray Pressure Observations.

Weather and Forecasting, 33(4), 933-953 https://doi.org/http://dx.doi.org/10.1175/WAF-D-17-0155.1

The High-Resolution Rapid Refresh-Alaska (HRRR-AK) modeling system provides 3-km horizontal resolution and 0-36-h forecast guidance for weather conditions over Alaska. This study evaluated the experimental version of the HRRR-AK system available from December 2016 to June 2017, prior to its operational deployment by the National Centers for Environmental Prediction in July 2018. Surface pressure observations from 158 National Weather Service (NWS) stations assimilated during the model's production cycle and pressure observations from 101 USArray Transportable Array (TA) stations that were not assimilated were used to evaluate 265 complete 0-36-h forecasts of the altimeter setting (surface pressure reduced to sea level). The TA network is the largest recent expansion of Alaskan weather observations and provides an independent evaluation of the model's performance during this period. Throughout the study period, systematic differences in altimeter setting between the HRRR-AK 0-h forecasts were larger relative to the unassimilated TA observations than relative to the assimilated NWS observations. Upon removal of these initial biases from each of the subsequent 1-36-h altimeter setting forecasts, the model's 36-h forecast root-mean-square errors at the NWS and TA locations were comparable. The model's treatment of RAPid warming and downslope winds that developed in the lee of the Alaska Range during 12-15 February is examined. The HRRR-AK 0-h forecasts were used to diagnose the synoptic and mesoscale conditions during this period. The model forecasts underestimated the abrupt increases in the temperature and intensity of the downslope winds with smaller errors as the downslope wind events evolved.

Pinto, J. O., Grim, J. A., & Steiner, M. (2015). Assessment of the High-Resolution Rapid Refresh Model's Ability to Predict Mesoscale Convective Systems Using Object-Based Evaluation. *Weather and Forecasting, 30*(4), 892-913 https://doi.org/10.1175/WAF-D-14-00118.1

An object-based verification technique that keys off the radar-retrieved vertically integrated liquid (VIL) is used to evaluate how well the High-Resolution Rapid Refresh (HRRR) predicted mesoscale convective systems (MCSs) in 2012 and 2013. It is found that the modeled radar VIL values are roughly 50% lower than observed. This mean bias is accounted for by reducing the radar VIL threshold used to identify MCSs in the HRRR. This allows for a more fair evaluation of the model's skill at predicting MCSs. Using an optimized VIL threshold for each summer, it is found that the HRRR reproduces the first (i.e., counts) and second moments (i.e., size distribution) of the observed MCS size distribution averaged over the eastern United States, as well as their aspect ratio, orientation, and diurnal variations. Despite threshold optimization, the HRRR tended to predict too many (few) MCSs at lead times less (greater) than 4 h because of lead time-dependent biases in the modeled radar VIL. The HRRR predicted too many MCSs over the Great Plains and too few MCSs over the southeastern United States during the day. These biases are related to the model's tendency to initiate too many MCSs over the Great Plains and too few MCSs over the southeastern United States. Additional low biases found over the Mississippi River valley region at night revealed a tendency for the HRRR to dissipate MCSs too quickly. The skill of the HRRR at predicting specific MCS events increased between 2012 and 2013, coinciding with changes in both the model physics and in the methods used to assimilate the three-dimensional radar reflectivity.

Seo, B.-C., Quintero, F., & Krajewski, W. F. (2018). High-Resolution QPF Uncertainty and Its Implications for Flood Prediction: A Case Study for the Eastern Iowa Flood of 2016. *Journal of Hydrometeorology, 19*(8), 1289-1304 https://doi.org/10.1175/JHM-D-18-0046.1

This study addresses the uncertainty of High-Resolution Rapid Refresh (HRRR) quantitative precipitation forecasts (QPFs), which were recently appended to the operational hydrologic forecasting framework. In this study, we examine the uncertainty features of HRRR QPFs for an Iowa flooding event that occurred in September 2016. Our evaluation of HRRR QPFs is based on the conventional approach of QPF verification and the analysis of mean areal precipitation (MAP) with respect to forecast lead time. The QPF verification results show that the precipitation forecast skill of HRRR significantly drops during short lead times and then gradually decreases for further lead times. The MAP analysis also demonstrates that the QPF error sharply increases during short lead times and starts decreasing slightly beyond 4-h lead time. We found that the variability of QPF error measured in terms of MAP decreases as basin scale and lead time become larger and longer, respectively. The effects of QPF uncertainty on hydrologic prediction are quantified through the hillslope-link model (HLM) simulations using hydrologic performance metrics (e.g., Kling-Gupta efficiency). The simulation results agree to some degree with those from the MAP analysis, finding that the performance achieved from the QPF forcing decreases during 1-3-h lead times and starts increasing with 4-6-h lead times. The best performance acquired at the 1-h lead time does not seem acceptable because of the large overestimation of the flood peak, along with an erroneous early peak that is not observed in streamflow observations. This study provides further evidence that HRRR contains a well-known weakness at short lead times, and the QPF uncertainty (e.g., bias) described as a function of forecast lead times should be corrected before its use in hydrologic prediction.

Section VIII: Observing System Experiment (OSE) & Observing System Simulation Experiments (OSSE)

Boukabara, S.-A., Ide, K., Zhou, Y., Shahroudi, N., Hoffman, R. N., Garrett, K., . . . Atlas, R. (2018). Community Global Observing System Simulation Experiment (OSSE) Package (CGOP): Assessment and Validation of the OSSE System Using an OSSE-OSE Intercomparison of Summary Assessment Metrics. *Journal of Atmospheric and Oceanic Technology, 35*(10), 2061-2078 https://doi.org/10.1175/jtech-d-18-0061.1

Observing system simulation experiments (OSSEs) are used to simulate and assess the impacts of new observing systems planned for the future or the impacts of adopting new techniques for exploiting data or for forecasting. This study focuses on the impacts of satellite data on global numerical weather prediction (NWP) systems. Since OSSEs are based on simulations of nature and observations, reliable results require that the OSSE system be validated. This validation involves cycles of assessment and calibration of the individual system components, as well as the complete system, with the end goal of reproducing the behavior of real-data observing system experiments (OSEs). This study investigates the accuracy of the calibration of an OSSE systemhere, the Community Global OSSE Package (CGOP) systembefore any explicit tuning has been performed by performing an intercomparison of the OSSE summary assessment metrics (SAMs) with those obtained from parallel real-data OSEs. The main conclusion reached in this study is that, based on the SAMs, the CGOP is able to reproduce aspects of the analysis and forecast performance of parallel OSEs despite the simplifications employed in the OSSEs. This conclusion holds even when the SAMs are stratified by various subsets (the tropics only, temperature only, etc.).

Boukabara, S. A., Ide, K., Shahroudi, N., Zhou, Y., Zhu, T., Li, R. F., . . . Hoffman, R. N. (2018). Community Global Observing System Simulation Experiment (OSSE) Package (CGOP): Perfect Observations Simulation Validation. *Journal of Atmospheric and Oceanic Technology, 35*(1), 207-226 https://doi.org/10.1175/jtech-d-17-0077.1

The simulation of observations-a critical Community Global Observing System Simulation Experiment (OSSE) Package (CGOP) component-is validated first by a comparison of error-free simulated observations for the first 24 h at the start of the nature run (NR) to the real observations for those sensors that operated during that period. Sample results of this validation are presented here for existing low-Earth-orbiting (LEO) infrared (IR) and microwave (MW) brightness temperature (BT) observations, for radio occultation (RO) bending angle observations, and for various types of conventional observations. For sensors not operating at the start of the NR, a qualitative validation is obtained by comparing geographic and statistical characteristics of observations over the initial day for such a sensor and an existing similar sensor. The comparisons agree, with no significant unexplained bias, and to within the uncertainties caused by real observation errors, time and space collocation differences, radiative transfer uncertainties, and differences between the NR and reality. To validate channels of a proposed future MW sensor with no equivalent existing spaceborne sensor channel, multiple linear regression is used to relate these channels to existing similar channels. The validation then compares observations simulated from the NR to observations predicted by the regression relationship applied to actual real observations of the existing channels. Overall, the CGOP simulations of error-free observations from conventional and satellite platforms that make up the global observing system are found to be reasonably accurate and suitable as a starting point for creating realistic simulated observations for OSSEs. These findings complete a critical step in the CGOP validation, thereby reducing the caveats required when interpreting the OSSE results.

de Azevedo, H. B., de Goncalves, L. G. G., Bastarz, C. F., & Silveira, B. B. (2017). Observing System Experiments in a 3DVAR Data Assimilation System at CPTEC/INPE. *Weather and Forecasting, 32*(3), 873-880 https://doi.org/10.1175/waf-d-15-0168.1

The Center for Weather Forecast and Climate Studies [Centro de Previsao e Tempo e Estudos Climaticos (CPTEC)] at the Brazilian National Institute for Space Research [Instituto Nacional de Pesquisas Espaciais (INPE)] has recently operationally implemented a three-dimensional variational data assimilation (3DVAR) scheme based on the Gridpoint Statistical Interpolation analysis system (GSI). Implementation of the GSI system within the atmospheric global circulation model from CPTEC/INPE (AGCM-CPTEC/INPE) is hereafter referred to as the Global 3DVAR (G3DVAR) system. The results of an observing system experiment (OSE) measuring the impacts of radiosonde, satellite radiance, and GPS radio occultation (RO) data on the new G3DVAR system are presented here. The observational impact of each of these platforms was evaluated by measuring the degradation of the geopotential height anomaly correlation and the amplification of the RMSE of the wind. Losing the radiosonde, GPS RO, and satellite radiance data in the OSE resulted in negative impacts on the geopotential height anomaly correlations globally. Nevertheless, the strongest impacts were found over the Southern Hemisphere and South America when satellite radiance data were withheld from the data assimilation system.

English, J. M., Kren, A. C., & Peevey, T. R. (2018). Improving Winter Storm Forecasts with Observing System Simulation Experiments (OSSEs). Part 2: Evaluating a Satellite Gap with Idealized and

Targeted Dropsondes. *Earth and Space Science, 5*(5), 176-196 https://doi.org/10.1002/2017ea000350

Numerous satellites utilized in numerical weather prediction are operating beyond their nominal lifetime, and their replacements are not yet operational. We investigate the impacts of a loss of U.S.-based microwave and infrared satellite data and the addition of dropsonde data on forecast skill by conducting Observing System Simulation Experiments with the European Centre for Medium-range Weather Forecasts T511 Nature Run and the National Center for Environmental Prediction Global Forecast System Model. Removing all U.S.-based microwave and infrared satellite data increases Global Forecast System analysis error, global forecast error, and forecast error during the first 36 hr of three winter storms that impact the United States. Data from Suomi National Polar-orbiting Partnership contributes roughly one third of the total satellite impacts. Assimilating "idealized" dropsondes (sampling over a large region of the Pacific/Arctic Oceans) significantly improves global forecasts and forecasts for all three storms. Assimilating targeted dropsonde flight paths using the Ensemble Transform Sensitivity method for 15 verification dates/locations for the three storms improves roughly 80% of forecasts relative to the control and 50% of forecasts relative to their corresponding experiments without dropsondes. However, removing satellite data degrades only 30% of targeted domain forecasts relative to the control. These results suggest that targeted dropsondes cannot compensate for a gap in satellite data regarding global average forecasts but may be able to compensate for specific targeted storms. However, as with any study of specific weather events, results are variable and more cases are needed to conclude whether targeted observations-as well as satellite data-can be expected to improve forecasts of specific weather events.

Hoffman, R. N., Boukabara, S.-A., Kumar, V. K., Garrett, K., Casey, S. P. F., & Atlas, R. (2017). An Empirical Cumulative Density Function Approach to Defining Summary NWP Forecast Assessment Metrics. *Monthly Weather Review, 145*(4), 1427-1435 https://doi.org/10.1175/mwr-d-16-0271.1

The empirical cumulative density function (ECDF) approach can be used to combine multiple, diverse assessment metrics into summary assessment metrics (SAMs) to analyze the results of impact experiments and preoperational implementation testing with numerical weather prediction (NWP) models. The main advantages of the ECDF approach are that it is amenable to statistical significance testing and produces results that are easy to interpret because the SAMs for various subsets tend to vary smoothly and in a consistent manner. In addition, the ECDF approach can be applied in various contexts thanks to the flexibility allowed in the definition of the reference sample.The interpretations of the examples presented here of the impact of potential future data gaps are consistent with previously reported conclusions. An interesting finding is that the impact of observations decreases with increasing forecast time. This is interpreted as being caused by the masking effect of NWP model errors increasing to become the dominant source of forecast error.

Hoffman, R. N., Kumar, V. K., Boukabara, S.-A., Ide, K., Yang, F., & Atlas, R. (2018). Progress in Forecast Skill at Three Leading Global Operational NWP Centers During 2015–17 as Seen in Summary Assessment Metrics (SAMS). *Weather and Forecasting, 33*(6), 1661-1679 https://doi.org/10.1175/waf-d-18-0117.1

The summary assessment metric (SAM) method is applied to an array of primary assessment metrics (PAMs) for the deterministic forecasts of three leading numerical weather prediction (NWP) centers for

the years 2015–17. The PAMs include anomaly correlation, RMSE, and absolute mean error (i.e., the absolute value of bias) for different forecast times, vertical levels, geographic domains, and variables. SAMs indicate that in terms of forecast skill ECMWF is better than NCEP, which is better than but approximately the same as UKMO. The use of SAMs allows a number of interesting features of the evolution of forecast skill to be observed. All three centers improve over the 3-yr period. NCEP short-term forecast skill substantially increases during the period. Quantitatively, the effect of the 11 May 2016 NCEP upgrade to the four-dimensional ensemble variational data assimilation (4DEnVar) system is a 7.37% increase in the probability of improved skill relative to a randomly chosen forecast metric from 2015 to 2017. This is the largest SAM impact during the study period. However, the observed impacts are within the context of slowly improving forecast skill for operational global NWP as compared to earlier years. Clearly, the systems lagging ECMWF can improve, and there is evidence from SAMs in addition to the 4DEnVar example that improvements in forecast and data assimilation systems are still leading to forecast skill improvements.

Hwang, S. O., & Hong, S. Y. (2012). The Impact of Observation Systems on Medium-Range Weather Forecasting in a Global Forecast System. *Asia-Pacific Journal of Atmospheric Sciences, 48*(2), 159-170 https://doi.org/10.1007/s13143-012-0016-4

To investigate the impact of various types of data on medium-range forecasts, observing system experiments are performed using an assimilation algorithm based on the National Centers for Environmental Prediction (NCEP)/Department of Energy (DOE) reanalysis system. Data-denial experiments for radiosonde, satellite, aircraft, and sea surface observations, and selected data experiments for radiosonde and surface data, are conducted for the boreal summer of 1997 and the boreal winter of 1997/1998. The data assimilation system used in this study is remarkably dependent on radiosonde data, which provides information about the three-dimensional structure of the atmosphere. As expected, the impact of radiosonde observations on medium-range forecasts is strongly positive over the Northern Hemisphere and tropics, whereas the satellite system is most beneficial over the Southern Hemisphere. These results are also found in experiments simulating historical changes in observation systems. Over the tropics, assimilation without radiosonde observations generates unbalanced analyses resulting in unrealistic forecasts that must be corrected by the forecast model. Forecasts based on analysis from the observation data before the era of radiosonde observation are found to be less meaningful. In addition, the impacts on forecasts are closely related to the geographical distribution of observation data. The memory of observation data embedded in the analysis tends to persist throughout forecasts. However, cases exist where the effect of forecast error growth is more dominant than that of analysis error, e.g., over East Asia in summer, and where the deficiency in observations is supplemented or the imbalance in analysis is adjusted by the forecast model during the period of forecasts. Forecast error growth may be related to the synoptic correction performed by the data assimilation system. Over data-rich areas, analysis fields are corrected to a greater extent by the data assimilation system than are those over data-poor areas, which can cause the forecast model to produce more forecast errors in medium-range forecasts. It is found that even one month per season is sufficient for forecast skill verification in data impact experiments. Additionally, the use of upper-air observations is found to benefit areas that are downstream of observation data-rich areas.

Ishibashi, T. (2014). Observing System Simulation Experiments with Multiple Methods. In *Remote Sensing and Modeling of the Atmosphere, Oceans, and Interactions V.* T. N. Krishnamurti & G. Liu (Eds.), (Vol. 9265). Bellingham: SPIE https://doi.org/10.1117/12.2069087

An observing System Simulation Experiment (OSSE) is a method to evaluate impacts of hypothetical observing systems on analysis and forecast accuracy in numerical weather prediction (NWP) systems. Since OSSE requires simulations of hypothetical observations, uncertainty of OSSE results is generally larger than that of observing system experiments (OSEs). To reduce such uncertainty, OSSEs for existing observing systems are often carried out as calibration of the OSSE system. The purpose of this study is to achieve reliable OSSE results based on results of OSSEs with multiple methods. There are three types of OSSE methods. The first one is the sensitivity observing system experiment (SOSE) based OSSE (SOSE-OSSE). The second one is the ensemble of data assimilation cycles (ENDA) based OSSE (ENDA-OSSE). The third one is the nature-run (NR) based OSSE (NR-OSSE). These three OSSE methods have very different properties. The NR-OSSE evaluates hypothetical observations in a virtual (hypothetical) world, NR. The ENDA-OSSE is very simple method but has a sampling error problem due to a small size ensemble. The SOSE-OSSE requires a very highly accurate analysis field as a pseudo truth of the real atmosphere. We construct these three types of OSSE methods in the Japan Meteorological Agency (JMA) global 4D-Var experimental system. In the conference, we will present initial results of these OSSE systems and their comparisons.

Jones, T. A., Otkin, J. A., Stensrud, D. J., & Knopfmeier, K. (2014). Forecast Evaluation of an Observing System Simulation Experiment Assimilating Both Radar and Satellite Data. *Monthly Weather Review, 142*(1), 107-124 https://doi.org/10.1175/mwr-d-13-00151.1

In the first part of this study, Jones et al. compared the relative skill of assimilating simulated radar reflectivity and radial velocity observations and satellite 6.95-m brightness temperatures T-B and found that both improved analyses of water vapor and cloud hydrometeor variables for a cool-season, high-impact weather event across the central United States. In this study, the authors examine the impact of the observations on 1-3-h forecasts and provide additional analysis of the relationship between simulated satellite and radar data observations to various water vapor and cloud hydrometeor variables. Correlation statistics showed that the radar and satellite observations are sensitive to different variables. Assimilating 6.95-m T-B primarily improved the atmospheric water vapor and frozen cloud hydrometeor variables such as ice and snow. Radar reflectivity proved more effective in both the lower and midtroposphere with the best results observed for rainwater, graupel, and snow. The impacts of assimilating both datasets decrease rapidly as a function of forecast time. By 1 h, the effects of satellite data become small on forecast cloud hydrometeor values, though it remains useful for atmospheric water vapor. The impacts of radar data last somewhat longer, sometimes up to 3 h, but also display a large decrease in effectiveness by 1 h. Generally, assimilating both satellite and radar data simultaneously generates the best analysis and forecast for most cloud hydrometeor variables.

Xue, Y., Wen, C. H., Yang, X. S., Behringer, D., Kumar, A., Vecchi, G., . . . Gudgel, R. (2017). Evaluation of Tropical Pacific Observing Systems Using NCEP and GFDL Ocean Data Assimilation Systems. *Climate Dynamics, 49*(3), 843-868 https://doi.org/10.1007/s00382-015-2743-6

The TAO/TRITON array is the cornerstone of the tropical Pacific and ENSO observing system. Motivated by the recent RAPid decline of the TAO/TRITON array, the potential utility of TAO/TRITON was assessed for ENSO monitoring and prediction. The analysis focused on the period when observations from Argo floats were also available. We coordinated observing system experiments (OSEs) using the global ocean data assimilation system (GODAS) from the National Centers for Environmental Prediction and the

ensemble coupled data assimilation (ECDA) from the Geophysical Fluid Dynamics Laboratory for the period 2004-2011. Four OSE simulations were conducted with inclusion of different subsets of in situ profiles: all profiles (XBT, moorings, Argo), all except the moorings, all except the Argo and no profiles. For evaluation of the OSE simulations, we examined the mean bias, standard deviation difference, root-mean-square difference (RMSD) and anomaly correlation against observations and objective analyses. Without assimilation of in situ observations, both GODAS and ECDA had large mean biases and RMSD in all variables. Assimilation of all in situ data significantly reduced mean biases and RMSD in all variables except zonal current at the equator. For GODAS, the mooring data is critical in constraining temperature in the eastern and northwestern tropical Pacific, while for ECDA both the mooring and Argo data is needed in constraining temperature in the western tropical Pacific. The Argo data is critical in constraining temperature in off-equatorial regions for both GODAS and ECDA. For constraining salinity, sea surface height and surface current analysis, the influence of Argo data was more pronounced. In addition, the salinity data from the TRITON buoys played an important role in constraining salinity in the western Pacific. GODAS was more sensitive to withholding Argo data in off-equatorial regions than ECDA because it relied on local observations to correct model biases and there were few XBT profiles in those regions. The results suggest that multiple ocean data assimilation systems should be used to assess sensitivity of ocean analyses to changes in the distribution of ocean observations to get more robust results that can guide the design of future tropical Pacific observing systems.