Forecast Impact of Assimilating Aircraft WVSS-II Water Vapor Mixing Ratio Observations in the Global Data Assimilation System (GDAS)

Brett T. Hoover[1], David A. Santek, Anne-Sophie Daloz, Yafang Zhong, Richard Dworak, Ralph A. Petersen

Cooperative Institute for Meteorological Satellite Studies, Space Science and Engineering Center, University of Wisconsin – Madison, Madison, WI

Andrew Collard

I.M. Systems Group at NOAA/NCEP/EMC, College Park, MD

PROJECT REPORT

February 2016

[1] Corresponding author information: Cooperative Institute for Meteorological Satellite Studies, Space Science and Engineering Center, University of Wisconsin – Madison, 1225 W Dayton St, Madison, WI 53706. Email: brett.hoover@ssec.wisc.edu

Abstract

Automated aircraft observations of wind and temperature have demonstrated positive impact on numerical weather prediction since the mid 1980s. With the advent of the WVSS-II humidity sensor, the expanding fleet of commercial aircraft with on-board automated sensors is also capable of delivering high-quality moisture observations, providing vertical profiles of moisture as aircraft ascend out of and descend into airports across the continental United States. Observations from the WVSS-II have to-date only been monitored within the Global Data Assimilation System (GDAS) without being assimilated.

In this study, aircraft moisture observations from the WVSS-II are assimilated in the GDAS, and their impact is assessed in the Global Forecast System (GFS). A two-season study is performed, demonstrating statistically significant positive impact on both the moisture forecast and the precipitation forecast at short-range (12-36 hours) in the warm season. No statistically significant impact is observed in the cold season.

An additional experiment is carried out to investigate if aircraft observations can completely replace rawinsonde observations where aircraft typically provide vertical profiles throughout the day. Results are mixed, leaving the impression that aircraft are currently not capable of routinely replacing rawinsonde observations, although available evidence suggests that profiles from aircraft observations can effectively remove the impact of a rawinsonde observation if aircraft observations are present in large enough numbers, leaving open the possibility for routine replacement of rawinsondes by aircraft observations in the future when aircraft observations become more numerous. These

results may also have relevance to the deployment of supplementary, off-time rawinsondes.

1. Introduction

Automated observations of wind and temperature from commercial aircraft have become a significant source of observations, especially since the establishment of the Meteorological Data Collection and Reporting System (MDCRS; Petersen et al. 1992). Today, 39 participating airlines deploy more than 3500 aircraft under the World Meteorological Organization's broader Aircraft Meteorological Data Relay (AMDAR) program, delivering more than 680,000 wind and temperature reports daily (Petersen et al. 2015).

Aircraft wind and temperature observations have demonstrated positive impact on numerical weather prediction since the mid 1980s when aircraft data became available in significant numbers (Moninger et al. 2003). Data denial experiments in the Rapid Update Cycle model (RUC, replaced by the Rapid Refresh (RAP) model in 2012; Benjamin et al. 2010) demonstrate that aircraft data is the most important dataset over the continental United States for 3-6 hour forecasts as well as 12-hour forecasts of upper tropospheric winds. Assimilation of wind, temperature, and moisture observations from Tropospheric AMDAR (TAMDAR) observations in the RUC demonstrate positive impact on wind, temperature, and moisture fields for the 3-hour forecast (Moninger et al. 2010). Impact tests in the European Centre for Medium-range Weather Forecasts (ECMWF) global forecast system demonstrate positive impact at 48 hours over the North Pacific, North America, North Atlantic, and Europe when assimilating aircraft wind and temperature observations (Andersson et al. 2005). An experimental ensemble-based observation impact system used with the NCEP GFS demonstrated that aircraft wind and temperature observations supply the largest per-observation impact of any in-situ observation type on the 24-hr forecast error, even surpassing rawinsonde observations (Ota et al. 2013).

Early attempts to derive automated moisture observations from aircraft sensors included the Water Vapor Sensing System (WVSS), which used a thin-film capacitor to measure relative humidity (RH; Fleming 1996). Tests of the device indicated a wet bias at high RH values and a dry bias at low RH values (Fleming 1998). In addition, biases in AMDAR temperature reports made it difficult to retrieve precise values of moisture variables, such as specific humidity (SH). The WVSS-II sensor was redesigned to use a tunable diode laser to measure water vapor content via infrared absorption spectroscopy, determining the water vapor content of sampled air from the measured transmittance of the laser across the air tube (Helms et al. 2010). Version 3 of the WVSS-II was developed in 2008, and performed well under most test conditions and eliminated technical issues with seals and thermal control that plagued the earlier versions of the design. The WVSS-II_(v3) is the device currently on board over 100 aircraft, routinely producing approximately 100,000 moisture observations daily over the continental US (Petersen et al. 2015).

WVSS-II moisture observations from AMDAR have been assimilated into the NDAS since the 18 October 2011 upgrade[2], which included substantial modification of the model grid, model physics, and data assimilation. However, these observations have only been monitored within the GDAS, passing through the data assimilation system and

[2] http://www.emc.ncep.noaa.gov/mmb/mmbpll/eric.html#TAB4

being assigned an interpolated model background moisture value, but not actually being assimilated. It is the goal of this study to assimilate these moisture observations in the GDAS, evaluate their impact on the GFS forecast, and determine what level of redundancy may exist between aircraft observations and rawinsonde observations in locations where airports provide ascending and descending aircraft observation profiles near rawinsonde launch sites. Section 2 outlines the model setup and experiment design; section 3 describes the methodology for assessing forecast impact; the results are presented in section 4, and conclusions are provided in section 5.

2. Model Setup and Experiment Design

Observations are assimilated using the operational, hybrid ensemble/3DVAR formulation of the GDAS to produce 6-hourly analyses. Analyses are produced at T670 resolution while using a set of 80 ensemble members at T254 resolution to define flow-dependent covariance (e.g. Wang et al. 2013). A 168-hr forecast is initiated from every 0000 UTC analysis for the purposes of assessing impact on the short- to medium-range forecast.

The GDAS was cycled for a warm season (01 April 2014 – 29 May 2014) and a cold season (01 December 2014 – 11 January 2015) to examine the impact of assimilated aircraft moisture observations across seasons. Moisture observations from AMDAR were switched from a monitoring-mode to an assimilation-mode at the script level, with an observation error profile copied from the NDAS. No change to quality control was made to account for the new moisture observations, allowing the GDAS to apply its existing quality control algorithm to these observations. Aircraft moisture observations are assimilated in SH space, rather than as a relative humidity measurement, which can be subject to significant effects from known biases in AMDAR temperature reports (Zhu et al. 2015). Each seasonal experiment was compared with a control that assimilated all observations except aircraft moisture observations. AMDAR wind and temperature observations are assimilated in both the experiment and control runs.

To test the redundancy of aircraft observations near rawinsondes, an additional warm season experiment was run. This experiment was identical to the moisture assimilation experiment already performed, except that rawinsondes at 10 selected sites in the continental US were deactivated in assimilation for all analysis periods, regardless of the aircraft observational coverage at any individual time. The choice of which sites to deactivate was based on the coverage of the rawinsonde site by aircraft moisture observations calculated from the original experiment (see Section 4c). This experiment was (referred to hereafter as the data-denial experiment) compared with the original experiment (referred to hereafter as the assimilation experiment) to determine the impact of the missing rawinsondes.

3. Methodology

Forecast impact was investigated in several ways. Observation-minus-background (OMB) statistics of assimilated moisture observations from both rawinsondes and nearby aircraft observations were compared to assess the data-quality of aircraft moisture observations relative to rawinsondes, an approach that is similar to that used in previous aircraft/rawinsonde collocation studies (e.g. Schwartz and Benjamin 1995). The

OMB statistics provide an evaluation of possible bias that may exist in the model background moisture field, as well as a measure of 6-hr forecast improvement.

Forecast performance in the shorter-range (1-2 days) is evaluated by examining the impact of assimilated observations on the Equitable Threat Score (ETS) and Bias Score[3] for precipitation in the 12-36 hour forecast. Forecast improvement or degradation is evaluated for statistical significance based on 10,000 Monte Carlo simulations. Although precipitation statistics are available for the 36-60 hour forecast range and the 60-84 hour forecast range, focus is maintained on the 12-36 hour forecast, because this is the time period over which aircraft moisture observations had the greatest impact on the forecast.

Forecast performance in the mid-range (2-3 days) is evaluated by comparing forecast total-column precipitable water (TPW) against TPW observed by Global Positioning Satellite (GPS) signals (e.g. Duan et al. 1996) available from Earth System Research Laboratory (ESRL). Unlike precipitation statistics that focus on the impact of moisture observations as the model approaches saturation, TPW comparisons provide a good means of evaluating the vertically integrated effect of AMDAR moisture reports throughout the full range of humidity. Errors from GPS-TPW have been shown to be less than 1 mm when compared to ground based Microwave Radiometer observations during the Measurements of Humidity in the Atmosphere and Validation Experiment (Leblanc et al. 2011) in California and at the Atmospheric Radiation Measurement (ARM) program (Dworak and Petersen, 2013) in Oklahoma. Furthermore, positive impacts on RUC

[3] Bias in the NCEP precipitation statistics is calculated as the ratio of the number of verification grid-boxes that are forecast to have precipitation in a given range (mm day^{-1}) to the number of grid-boxes where that amount of precipitation actually occurred.

forecasts out to 12 hours have been observed with the assimilation of GPS TPW data (Gutman et al 2004, Smith et al. 2007). Forecast errors relative to GPS observations were computed in this study for every 6-hour forecast period out to 72 hours, and deviations from the error in the control forecast are evaluated for statistical significance using a student's t-test for mean error and bias, and a chi-squared test for random error (see Section 4b).

Two differences between precipitation skill scores and TPW fit-to-observation scores must be considered. First, GPS observational coverage is more comprehensive spatially and temporally than precipitation data, which allows every forecast to be tested for accuracy at more locations than with the more sparse precipitation data. For example, the 12-36 hour forecast period over which the precipitation skill scores are presented is binned by precipitation amount, with the largest number of precipitation observations in the lowest-value bin. For the warm-season experiment, there are at least 42,057 data points used to determine ETS and Bias Scores when comparing the assimilation experiment and the data-denial experiment. By contrast, in the forecast fit-to-observations test, 69,071 observations were tested over the same forecast period, a 64% increase in available observations. This is due in part to the fact that TPW observations can exist where precipitation observations do not, sampling across the full spectrum of moisture values. There are over 400 active GPS-MET stations at any given observation period +/- 15 minutes from the hour, with high geographic density in California. Second, the forecast fit-to-observations test shows statistically significant degradation in the medium-range (72 hours), while statistically significant impact on precipitation scores only extends to the 12-36 hour forecast. For these reasons, one could argue that the GPS-TPW tests are more comprehensive.

4. Results

a) Impact of AMDAR moisture observations on rawinsonde moisture assimilation

A particular interest in this study is to investigate the relationship between rawinsonde moisture observations and AMDAR moisture observations during assimilation, when both are available in the same location. It is desirable to know, for example, how rawinsonde and AMDAR moisture observations compare to the model background derived from the 6-hour forecast – a closer fit of observations to the 6-hour forecast following assimilation can indicate that the observations are high quality and improve the initial (analysis) state. Likewise, an improved fit of *rawinsonde* observations to the 6-hour forecast as a result of assimilating *AMDAR* observations can indicate better model performance, as this can be equivalently expressed as a closer fit of the 6-hour forecast to trusted observations. Lower OMB in general implies greater consistency of the observations with other observation data sources as well as with information from observations assimilated previously, contributing to the model background.

Mean profiles of OMB values of specific humidity are produced for rawinsonde observations without AMDAR moisture assimilation (from the control), rawinsonde observations with AMDAR moisture assimilation (from the assimilation experiment), and for AMDAR observations when they are assimilated (Fig. 1). Profiles are produced at each rawinsonde site, averaging rawinsonde OMB scores within 25 equally-spaced pressure-layers between the highest recorded pressure and 300 hPa, which is the highest level where moisture observations are assimilated. AMDAR OMB scores are likewise averaged within these pressure-layers using all AMDAR moisture observation within 1 hour and 0.5 degrees of the rawinsonde location, representing a radius of 66 km to 77 km, depending on the latitude of the rawinsonde site. These profiles are then averaged across all rawinsonde sites in the continental US.

Profiles indicate that rawinsonde observations fit 6-hour forecasts better when AMDAR observations are assimilated during the warm-season experiment (Fig. 1a), signifying improved model performance in the warm-season. No clear change is observed in the cold-season experiment (Fig. 1b). Likewise, AMDAR moisture observations fit closer to the 6-hr forecast than rawinsonde observations at essentially all levels in the warm-season experiment, although this relationship does not exist in the cold-season. This indicates that AMDAR moisture observations are of high quality, even in comparison to rawinsonde observations. In the cold-season when SH patterns are more strongly organized by synoptic-scale weather systems and values are smaller due to colder temperatures, AMDAR and rawinsonde observations appear to have largely indistinguishable quality characteristics by this metric, except for perhaps the surface and near-surface levels where rawinsondes are more moist than the model background and AMDAR observations are not. In the warm-season, the OMB for rawinsondes is improved to statistical significance within the lower troposphere down to just above the surface. The difference in OMB performance between the warm and cold seasons may be a byproduct of the increased presence of smaller-scale moisture structures in the warm season.

b) Impact of AMDAR moisture observations on precipitation and TPW forecasts

To quantify the effect of the analysis changes on the forecast due to inclusion of AMDAR moisture observations, precipitation forecast skill was determined using the Equitable Threat Score (ETS) and Bias Score (Wilks 1995) over the continental United States, binned by precipitation thresholds per 24 hours. The inclusion of both AMDAR and RAOB moisture observations (from the assimilation experiment) improved the mean ETS to statistical significance for 12-36 hour precipitation forecasts of below 5 mm/day in the warm-season experiment (Fig. 2a). Bias is slightly improved for these categories as well. There is statistically significant ETS degradation for only the 10 mm/day category of the 60-84 hour forecast (not shown), while the ETS and bias are not significantly changed for any other category at any forecast lead-time. The cold-season experiment expresses no statistically significant improvement in ETS or bias for any category or forecast lead-time (Fig. 2b), with the exception of a degradation in bias for very high precipitation (50-75 mm/day) in 60-84 hour forecasts (not shown); these higher precipitation categories have very few observations from which to derive statistics, and are dominated by a single event, making the statistics less reliable. Since the GFS precipitation forecast is more accurate in the cold-season, due to more organized convection from synoptic-scale forcing, improvement of the cold-season precipitation forecast is expected to be smaller than improvement of the warm-season forecast.

These precipitation statistics demonstrate an improvement to short-range (12-36 hour) precipitation forecasts by assimilation of AMDAR moisture observations. An additional measure of forecast skill can be observed by computing the forecast fit-to-observations using GPS total-column precipitable water. For each 6-hour forecast period from the analysis time to 72-hours, the forecast TPW fields were interpolated to a database of GPS observations and the error was computed (Fig. 3). The error is divided into two components: the bias of the error, represented by the mean difference between observations and the forecast field, and the random error, represented by the standard deviation of the difference between observations and the forecast field. In general, the bias of the error is typically 10-20% of the magnitude of the random error, indicating that the random error is responsible for most of the error.

While bias in the error is slightly increased in the first 18 hours of the forecast in the warm season experiment (Fig. 3a), the total error is reduced, with random error improved to statistical significance from 0-36 hours into the forecast, with additional statistically significant improvement at 60-66 hours (Fig. 3b). The mean error, which is a combination of both the bias and the random error, is reduced to statistical significance in for 0-18 hours into the forecast (not shown). The impact of AMDAR moisture observations on the cold-season experiment is less significant, with no statistically significant change in bias of error (Fig. 3c) and statistically significant reduction in random error only in the first 0-6 hours of the forecast (Fig. 3d). The mean error is only reduced to statistical significance at the analysis time (not shown). The difference in impact between the warm and cold season experiments may be due to differences in precipitation regime between the two periods. In the warm season, precipitation often forms in small-scale features under weak synoptic forcing, while in the cold season precipitation is dominated by large-scale, strong synoptic forcing. The GFS has less skill in the warm season regime, leaving more room for improvement.

c) AMDAR/Rawinsonde redundancy experiment

1) VERTICAL AND TEMPORAL COVERAGE OF RAWINSONDE LAUNCH SITES BY AIRCRAFT OBSERVATIONS

In the data-denial experiment, the value of rawinsonde observations in regions best observed by aircraft observations was tested. This is a very gross test of the potential of AMDAR observations (u,v,T,q) to complely replace rawinsondes at sites where AMDAR observations provide the most consistent coverage. The availability of aircraft observations at US rawinsonde sites was determined at each six-hourly analysis period by collecting aircraft moisture observations available within varying spatial thresholds of 0.25 to 1 degree in latitude/longitude space and 0.75 to 1.25 hours in time of the rawinsonde launch (Table 1). These aircraft observations are defined as 'collocated' with the rawinsonde for the purposes of defining coverage of the site. Coverage of a rawinsonde by aircraft observations is determined through coverage by aircraft SH observations alone; many aircraft provide wind and temperature observations while relatively few aircraft provide WVSS-II moisture observations. Thus coverage of wind, temperature, and SH observations by aircraft is primarily determined by the coverage of SH observations.

The vertical profile of the rawinsonde launch site is divided into 25 equally-spaced pressure layers between the surface and 300 hPa (which is the lowest allowable pressure for assimilation of rawinsonde and aircraft moisture observations). The vertical coverage of the site by aircraft observations (C_{Vertical}) is defined as the percentage of these layers that contain at least one aircraft moisture observation, computed every 0000 UTC and 1200 UTC analysis period and averaged to produce a final score. The choice of 25 layers was made because 25 layers appears to provide the most contrast between rawinsondes with high vertical coverage and rawinsondes with low vertical coverage. Likewise, the temporal coverage of the site by aircraft observations (C_{Temporal}) is defined as the percentage of 0000 UTC and 1200 UTC analysis-periods where at least one collocated aircraft moisture observation is available, such that an aircraft observation profile can be produced. The total coverage score (C_{total}) for a rawinsonde launch site is the product of these two coverage statistics, varying between zero and one:

$$C_{total} = C_{Vertical} * C_{Temporal} \qquad (1)$$

Rawinsonde launch sites are ranked by coverage, and the most well covered sites are used for the data-denial experiment. Table 1 lists the coverage statistics for three spatial/temporal thresholds of coverage. For example: The highest C_{total} value for a 1.0 hour and 0.5 degree threshold around rawinsonde sites is Fort Worth, TX, with a value of C_{total}=0.603. This can indicate, in the limiting cases, a situation where C_{vertical}=1.0 (perfect vertical coverage) and C_{temporal}=0.603 (profiles available in 60.3% of the 0000 UTC and 1200 UTC analysis periods), or alternatively C_{vertical}=0.603 (an average of 60.3% of vertical layers are covered by AMDAR moisture observations) and C_{temporal}=1.0 (profiles available in all 0000 UTC and 1200 UTC analysis periods). The reality exists in between these limiting cases, with C_{vertical}=0.675 and C_{temporal}=0.893. The ten sites chosen for the experiment include sites that appear in the top-10 for at least two thresholds, except for Las Vegas, NV, which is in the top-11 for two thresholds and in the top-3 for the strictest (smallest space/time) threshold. These sites are spread across the continental US, which allows for the assumption that they impact the forecast largely independent of one another.

In the data-denial experiment, the GDAS was run on a six-hourly cycle over the warm season period from 01 April – 29 May 2014, following a spin-up period of one week. All routine observations plus aircraft moisture observations were assimilated, but the entire rawinsonde (wind, temperature, and moisture observations) at each of the ten chosen sites was excluded for the full period of the experiment, regardless of the AMDAR coverage at any particular time. The purpose of this experiment is to determine if forecasts are significantly impacted by eliminating rawinsonde data where AMDAR observations are known to provide their most substantial coverage.

2) PRECIPITATION EQUITABLE THREAT SCORE (ETS) AND BIAS SCORE

The changes in ETS and Bias Score in the data-denial experiment are similar in form to the impact from the assimilation experiment that included both the aircraft moisture observations and the ten selected rawinsondes (Fig. 4). The 12-36 hour ETS score is improved to statistical significance for low precipitation amounts (0.2-2.0 mm/day) and bias is improved for precipitation amounts less than 10 mm/day. When compared to the same portion of the experiment with and without AMDAR moisture data, there is a notably larger positive impact on both ETS and bias scores when the ten selected rawinsondes have been removed, with statistical significance over these same precipitation thresholds (Fig. 5). Impacts on longer-range forecasts do not reach statistical significance (not shown). These results demonstrate that the ten selected rawinsondes are reducing precipitation skill rather than improving it; as shown in Fig. 1, OMB is smaller for aircraft moisture observations than for rawinsondes on average.

Aircraft moisture observations may be higher quality than rawinsonde moisture observations, consistent with other studies (Petersen et al. 2016), and that coverage by AMDAR moisture observations provides more information than rawinsondes launched twice daily. Thus it is possible that removing rawinsondes in regions of dense aircraft observational coverage could yield a positive impact at short-range (12-36 hours) in the warm season. The availability of aircraft observations at locations and times other than the twice-daily, point-specific rawinsonde launches may also allow aircraft observations to provide more information on moisture variability than is capable with rawinsondes.

3) FORECAST FIT-TO-OBSERVATIONS: GPS TOTAL-COLUMN PRECIPITABLE WATER

The fit of GPS-TPW observations to forecasts was calculated at all forecast times out to 72 hours for the data-denial experiment in the same manner that was applied to the assimilation experiment. The negative (dry) bias of forecast error in the data-denial experiment is more pronounced than in the original assimilation experiment (Fig. 6a), with a pronounced, statistically significant increase in bias through 0-48 hours into the forecast as well as at 66 hours. Random error is reduced in the data-denial experiment at a statistically significant magnitude on par with the original assimilation experiment up to 30 hours into the forecast, after which the random error of the data-denial experiment begins to become larger than the control, and exhibits statistically significant degradation from 60-72 hours (Fig. 6b).

4) OBSERVATION-MINUS-ANALYSIS (OMA) STATISTICS

As a final test of the impact of denying the selected rawinsondes, the observation-minus-analysis (OMA) statistics of aircraft moisture observations assimilated near the missing rawinsonde sites was compared with and without the rawinsondes present, and the difference was plotted against the density of aircraft observations present within a pressure layer and within 0.5 degrees and one hour of the rawinsonde (Fig. 7). While there is no correlation (i.e. linear relationship) between these two statistics (r = -0.0144), a relationship becomes clear when the points are plotted on a phase-space. The more aircraft observations nearby (higher values along the abscissa), the less the OMA statistic for aircraft moisture observations is capable of changing when the rawinsonde is denied (lower values along the ordinate). Thus the relationship between these two statistics is represented by an upper-bound on the ordinate as a function of the abscissa, which appears to obey an exponential-decay-like form.

5) EVALUATION OF RAWINSONDE DATA-DENIAL

The scores presented for this admittedly extreme test of the redundancy of rawinsondes at sites well covered by AMDAR observations do not reach a clear conclusion. While precipitation skill can be improved in the short range (12-36 hours) by their exclusion, forecast fit-to-observation against GPS total-column precipitable water suggests that denying the rawinsondes increases error, even to statistically significant degradation in random error at 60-72 hours against a control that contains no aircraft moisture observations.

In reconciling these results, one must consider the relative impact of moisture versus temperature and wind observations from aircraft. As shown previously, aircraft moisture observations near rawinsondes exhibit a lower OMB than rawinsonde observations (Fig. 1), which implies that aircraft moisture observations may be of higher quality. However, temperature observations from aircraft have been shown to suffer biases that can vary by individual aircraft as well as by whether the aircraft is ascending or descending (e.g. Ballish and Kumar 2008). While efforts to address these biases are currently being investigated (Isaksen et al. 2012, Zhu et al. 2015), NCEP does not currently employ a bias correction mechanism for these observations. It is possible that higher-quality moisture observations from aircraft improve the short-range precipitation forecast, when more accurate estimation of the humidity field may be most important, while denying rawinsondes may introduce errors into the wind and temperature fields that grow over time to degrade the later forecast, explaining the results from both the precipitation skill score test and the forecast fit-to-observations test.

The impact of the missing rawinsondes, measured as the change in OMA of the aircraft observations when the rawinsondes are excluded, demonstrates a relationship to the number of aircraft observations present; the more aircraft observations present (meaning the more redundancy in aircraft observational coverage at a particular location and pressure level), the smaller the upper-bound on the expected impact of denying the rawinsondes. Based on the best-fit curve describing the upper-bound, the expected OMA-impact of denying the 10 selected rawinsondes on a single, lone aircraft moisture observation is 1.47×10^{-3} kg/kg. To reduce this upper-bound by 50%, roughly 20 aircraft observations need to be present to reduce the impact of the missing rawinsondes. To reduce the upper-bound by another 50%, roughly 40 aircraft observations must be present. Given a threshold maximum allowable impact from denied rawinsondes, a minimum number of aircraft observations must be present.

Since the amount of aircraft observational coverage is highly variable, even for the most well-covered rawinsonde sites, permanent exclusion of these sondes in favor of aircraft observations, representing an extreme and permanent departure from reliance on the rawinsonde network, does not seem plausible. However, the opposite case could be considered: "Where would an *additional* rawinsonde provide the *most* impact, based on aircraft observational coverage?" This scenario occurs during off-time rawinsonde launches, which have become part of the adaptive observation network, especially during the Atlantic hurricane season when a significant hurricane threatens to make landfall on the east coast of the United States. Under extreme scenarios rawinsondes can be launched at 0600 and 1800 UTC from all operating sites in the continental US, as was the case with the days leading up to landfall of Hurricane Sandy (2012).

In scenarios such as these, the goal may be to deploy a limited number of off-time rawinsondes with a goal to maximize the impact on the analysis and the forecast of an extreme weather event. One could then expect that rawinsondes deployed where there is an expectation of dense aircraft observations would have less impact than rawinsondes deployed where there is an expectation of sparse aircraft observational coverage. The decision to launch an off-time rawinsonde at a particular site could be aided by statistics on the aircraft observational coverage at existing rawinsonde sites for these times.

6. Conclusions

The impact of assimilated aircraft moisture observations from the WVSS-II was evaluated in the GDAS/GFS analysis-forecast system. Cycled experiments were carried out for a warm season (April – May 2014) and a cold season (December 2014 – January 2015). The warm season experiment demonstrated positive impact on ETS and bias scores for low-precipitation categories in the 12-36 hour forecast. Assimilation of aircraft moisture observations in the warm season also produced smaller OMB values for rawinsonde observations, implying that the 6-hour forecast had been improved, though the nearby aircraft moisture observations had even lower OMB values. When the total-column precipitable water forecast was compared to observations from GPS, the assimilation of aircraft moisture observations in the warm season improved random error in the forecast as far out as 66 hours. By contrast, the cold season experiment only demonstrated statistically significant positive impact on random error out to 6 hours.

The difference in impact between the warm season and cold season experiments may be partially attributable to different precipitation regimes in either season. Warm season precipitation is often defined by small-scale moisture structures and weak synoptic forcing, which is an ongoing challenge to forecast in global NWP. The most room for improvement in precipitation forecasting is in the warm season, which may allow assimilation of AMDAR moisture observations to express a larger impact. By contrast, precipitation in the cold season is dominated by strong, synoptic-scale forcing that is more accurately predicted in global NWP. Under these circumstances, there may be less importance from small-scale moisture structures observed by AMDAR, and the already accurate forecasts from the GFS are more difficult to improve upon.

Redundancy between rawinsondes and aircraft observations was investigated by assimilating aircraft moisture observations, but also denying rawinsonde observations at 10 selected sites that are considered well-covered by aircraft in an a posteriori analysis. Precipitation skill scores are improved when the rawinsondes are denied, while the total-column precipitable water forecast suffers statistically significant degradation by 60-72 hours. It is possible that both of these results can be reconciled by recognizing that regions with denied rawinsondes will rely more heavily not only on aircraft moisture observations (which may be as high or higher quality than rawinsonde moisture observations), but also aircraft temperature observations, which are known to suffer biases based partially on phase of flight (ascending, descending, or flight-level). Relying on superior moisture observations from aircraft may improve the short-range precipitation forecast, while relying on biased temperature observations from aircraft may create growing errors that degrade the later forecast.

When the impact of denying rawinsonde observations is plotted as a function of the number of aircraft observations present, an exponential-decay-like relationship appears. Based on the relationship observed, it may take dozens of aircraft observations to reduce the impact of a denied rawinsonde to near zero, and coverage of a rawinsonde launch site by aircraft observations isn't consistent enough in time to allow for deactivation of even a rawinsonde near a busy airport where aircraft observations are collected frequently. However, the results may be of relevance to adaptive, off-time rawinsonde deployment; if the aircraft coverage at rawinsonde sites can be anticipated with enough lead time, it seems possible to anticipate when redundancy may occur and adjust the adaptive deployment accordingly. The fleet of aircraft providing these observations is growing, allowing for a possible re-evaluation of these redundancy experiments in the future.

Based on the presented research, the decision was made to implement assimilation of aircraft moisture observations in the operational GDAS, as part of NCEP's next upgrade. Implementation is currently slated for May 2016.

Acknowledgements

The authors would like to acknowledge Dr. Steve Lord (NOAA) and Dr. John Derber (NOAA) for their advice and expertise in guiding this project, Dr. Jim Jung (CIMSS) and Kate Howard (NOAA) for their advice and help with running the GDAS, Dr. Dennis Keyser (NOAA) and Yangrong Ling (NOAA) for their help with providing special rawinsonde data sets for the data-denial experiment, and Jim Nelson (CIMSS) for his help with software and data for the GPS fit-to-observation tests. This study was funded by NOAA through grant NA13NWS4830022.

References

Andersson, E., C. Cardinali, B. Truscott, and T. Hovberg, 2005: High frequency AMDAR data – a European aircraft data collection trial and impact assessment. ECMWF Technical Memorandum 457, 15 pp.

Ballish, B. A., and K. Kumar, 2008: Systematic differences in aircraft and radiosonde temperatures. *Bull. Amer. Meteor. Soc.*, **89**, 1689-1708.

Benjamin, S. G., B. D. Jamison, W. R. Moninger, S. R. Sahm, B. E. Schwartz, and T. W. Schlatter, 2010: Relative short-range forecast impact from aircraft, profiler, radiosonde, VAD, GPS-PW, METAR, and mesonet observations via the RUC hourly assimilation cycle. *Mon. Wea. Rev.*, **138**, 1319-1343.

Duan, J., M. Bevis, P. Fang, Y. Bock, S. Chiswell, S. Businger, C. Rocken, F. Solheim, T. van Hove, R. Ware, S. McClusky, T. A. Herring, and R. W. King, 1996: GPS meteorology: Direct estimation of the absolute value of precipitable water. *J. Appl. Meteorology*, **35**, 830-838.

Dworak, R. and R. Petersen, 2013: The validation of GOES-Li and AIRS total precipitable water retrievals using ground based measurements. Joint EUMETSAT 2013 Meteorological Satellite Conference and 19th American Meteorological Society AMS Satellite Meteorology, Oceanography, and Climatology Conference, Vienna, Austria, 16-20 September 2013.

Fleming, R. J., 1996: The use of commercial aircraft as platforms for environmental measurements. *Bull. Amer. Meteor. Soc.*, **77**, 2229-2242.

Fleming, R. J., 1998: A note on temperature and relative humidity corrections for humidity sensors. *J. Atmos. and Oceanic Technology*, **15**, 1511-1515.

Gutman, S. I. and S. G. Benjamin, 2001: The Role of Ground-Based GPS Meteorological Observations in Numerical Weather Prediction. *GPS Solutions*, **4**, 16-24.

Helms, D., A. Hoff, H. G. J. Smit, S. Taylor, S. Carlberg, and M. Berechree, 2010: Advancements in the AMDAR Humidity Sensing. WMO Technical Conference on

Meteorological and Environmental Instruments and Methods of Observation, TECO-2010.

Isaksen, L., D. Vasiljevic, D. Dee, and S. Healy, 2012: Bias correction of aircraft data implemented in November 2011. *ECMWF Newsletter*, No. 131, ECMWF, Reading, United Kingdom, 6-6.

Leblanc, T. and Coauthors, 2011: Measurements of Humidity in the Atmosphere and Validation Experiments (MOHAVE)-2009: Overview of campaign operations and results. *Atmospheric Measurement Techniques*, **4**, 2579-2605

Moninger, W. R., R. D. Mamrosh, and P. M. Pauley, 2003: Automated meteorological reports from commercial aircraft. *Bull. Amer. Meteor. Soc.*, **84**, 203-216.

Moninger, W. R., S. G. Benjamin, B. D. Jamison, T. W. Schlatter, T. L. Smith, and E. J. Szoke, 2010: Evaluation of regional aircraft observations using TAMDAR. *Wea. Forecasting*, **25**, 627-645.

Ota, Y., J. C. Derber, E. Kalnay, and Miyoshi, 2013: Ensemble-based observation impact estimates using the NCEP GFS. *Tellus*, **65A**, 20038.

Petersen, R., C. Dey, R. C. Martin, R. D. Londot, and G. T. Ligler, 1992: The Meteorological Data Collection and Reporting System (MDCRS): System overview and benefits. *Proc. National Weather Service Aviation Workshop*, Kansas City, MO, National Weather Service, 251-255. [Also available as NOAA Tech. Memo. NWS Cr-102.]

Petersen, R., L. Cronce, R. Mamrosh, and R. Baker, 2015: Impact and benefits of AMDAR temperature, wind, and moisture observations in operational weather forecasting. WMO Technical Report 2015-01, 93 pp. http://library.wmo.int/pmb_ged/wigos-tr_2015-01_en.pdf

Petersen, R., 2016: On the impact and benefits of AMDAR observations in operational forecasting. Part I: A review of the impact of automated aircraft wind and temperature reports. *Bull. Amer. Meteor Soc.*, in press. doi:10.1175/BAMS-D-14-00055.1

Schwartz, B., and S. G. Benjamin, 1995: A comparison of temperature and wind measurements from ACARS-equipped aircraft and rawinsondes. *Wea. Forecasting.*, **10**, 528-544.

Smith, T. L., S. G. Benjamin and S. I. Gutman, 2007: Short-range forecast impact from assimilation of GPS-IPW observations into the Rapid Update Cycle. *Mon. Wea. Rev.*, **135**, 2914-2930.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098-4117.

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

Zhu, Y., J. C. Derber, R. J. Purser, B. A. Ballish, and J. Whiting, 2015: Variational correction of aircraft temperature bias in the NCEP's GSI analysis system. *Mon. Wea. Rev.*, **143**, 3774-3803.

Table and Figure Captions

Table 1. Total coverage (C_{total}) at each of ten rawinsonde launch sites considered for data denial experiment. Rankings of each site by coverage are provided for three thresholds defining collocation of aircraft observations to the rawinsonde: (left) observations within 0.75 hours and 0.25 degrees of the site, (middle) observations within one hour and 0.5 degrees of the site, and (right) observations within 1.25 hours and 0.75 degrees of the site. Rankings in the top-10 are highlighted in red. Rankings provided are ranks provided out of all US rawinsonde sites.

Figure 1. Mean profiles of specific humidity ob-minus-background (OMB) for the warm-season experiment (left) and cold-season experiment (right) at rawinsonde launch sites. The blue profile is the mean rawinsonde moisture OMB when AMDAR moisture observations are not assimilated. The red profile is the mean rawinsonde moisture OMB when AMDAR moisture observations are assimilated. The green profile is the mean AMDAR moisture OMB. The shading around each profile represents the 5% and 95% confidence limits around the mean, and pressure-levels where the rawinsonde OMB changes to statistical significance are highlighted with black squares along the ordinate.

Figure 2. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for (a) warm-season experiment, and (b) cold-season experiment. The left panel of each plot shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel of each plot shows the

precipitation bias score in the same bins. The black curve is for the control simulation, and the red curve is for the experiment. The bottom panels show the differences between the experiment and control, with bars indicating the minimum value necessary for 95% statistical significance.

Figure 3. Error in forecast fit-to-TPW observations from GPS for (top) April 2014 – May 2014 simulation and (bottom) December 2014 – January 2015 experiment. Statistics for the control simulation are provided in blue, and statistics for the experiment are provided in red. Error is computed as (left) bias of error, calculated as the mean error, and (right) random error, calculated as the standard deviation of the error. Thick contours represent the sample mean or standard deviation, and the shading represents the 5% and 95% confidence limits on the mean or standard deviation. Dots are placed on the red contour for all times where the difference between the experiment and control is statistically significant based on a student's t-test (for bias of error) or a chi-squared test on variance (for random error).

Figure 4. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for the warm-season assimilation experiment and data-denial experiment. The left panel shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel shows the precipitation bias score in the same bins. The black curve is for the control simulation, the red curve is for the assimilation experiment, and the green curve is for the data-denial experiment. The

bottom panels show the differences between each experiment and the control, with bars indicating the minimum value necessary for 95% statistical significance.

Figure 5. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for the warm-season assimilation experiment and data-denial experiment. The left panel shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel shows the precipitation bias score in the same bins. The black curve is for the assimilation experiment, and the red curve is for the data-denial experiment. The bottom panels show the differences between the two experiments (data-denial experiment minus assimilation experiment), with bars indicating the minimum value necessary for 95% statistical significance.

Figure 6. Error in forecast fit-to-TPW observations from GPS for April 2014 – May 2014 simulations. Statistics for the control simulation are provided in blue, statistics for the assimilation experiment (rawinsondes, aircraft moisture observations) are provided in red, and statistics for the data-denial experiment (selected rawinsondes removed, aircraft moisture observations) are provided in green. Error is computed as (top) bias of error, calculated as the mean error, and (bottom) random error, calculated as the standard deviation of the error. Thick contours represent the sample mean or standard deviation, and the shading represents the 5% and 95% confidence limits on the mean or standard deviation. Dots are placed on the red and green contours for all times where the difference between the experiment and control is statistically significant based on a student's t-test (for bias of error) or a chi-squared test on variance (for random error).

Figure 7. Phase-space diagram of relationship between (ordinate) how much denied rawinsondes impact the assimilation of nearby AMDAR moisture observations, and (abscissa) how many AMDAR moisture observations are nearby. Each dot (red or blue) represents an AMDAR moisture observation assimilated within 1 hour and 0.5 degrees of a denied rawinsonde, for all 0000 UTC and 1200 UTC analysis periods in the data-denial experiment. The ordinate measures the absolute value of the difference in Observation-Minus-Analysis (OMA) between the assimilation experiment (where AMDAR moisture observations are assimilated and all rawindsondes are maintained) and the data-denial experiment (where AMDAR moisture observations are assimilated and selected rawinsonde observations are denied). The abscissa measures the number of AMDAR moisture observations collocated to the same rawinsonde within the same vertical pressure layer. The red dots represent the 5 highest OMA differences for each unique value along the abscissa, identifying the upper bound of the phase-space that is sampled by the observations. The solid black line is an empirically-derived exponential best-fit to the red dots, representing a theoretical expected upper-bound on the potential impact of denied rawinsondes as a function of the density of AMDAR observational coverage. The dashed black lines represent the 5% and 95% confidence bounds on the solid line.

Site	$\Delta T = 0.75, \quad \Delta D = 0.25$	$\Delta T = 1.00, \quad \Delta D = 0.50$	$\Delta T = 1.25, \quad \Delta D = 0.75$
Miami, FL	(25th) 0.046	(5th) 0.446	(5th) 0.653
Tampa, FL	(7th) 0.166	(2nd) 0.569	(2nd) 0.861
Atlanta, GA	(21st) 0.076	(8th) 0.376	(7th) 0.517
Fort Worth, TX	(12th) 0.126	(1st) 0.603	(3rd) 0.739
Nashville, TN	(1st) 0.231	(4th) 0.524	(4th) 0.717
Las Vegas, NV	(3rd) 0.213	(11th) 0.318	(11th) 0.410
Sterling, VA	(2nd) 0.222	(3rd) 0.540	(1st) 0.864
Denver, CO	(5th) 0.199	(9th) 0.368	(10th) 0.446
Oakland, CA	(4th) 0.209	(6th) 0.394	(8th) 0.496
Upton, NY	(10th) 0.132	(7th) 0.379	(9th) 0.478

Table 1. Total coverage (C_{total}) at each of ten rawinsonde launch sites considered for data denial experiment. Rankings of each site by coverage are provided for three thresholds defining collocation of aircraft observations to the rawinsonde: (left) observations within 0.75 hours and 0.25 degrees of the site, (middle) observations within one hour and 0.5 degrees of the site, and (right) observations within 1.25 hours and 0.75 degrees of the site. Rankings in the top-10 are highlighted in red. Rankings provided are ranks provided out of all US rawinsonde sites.

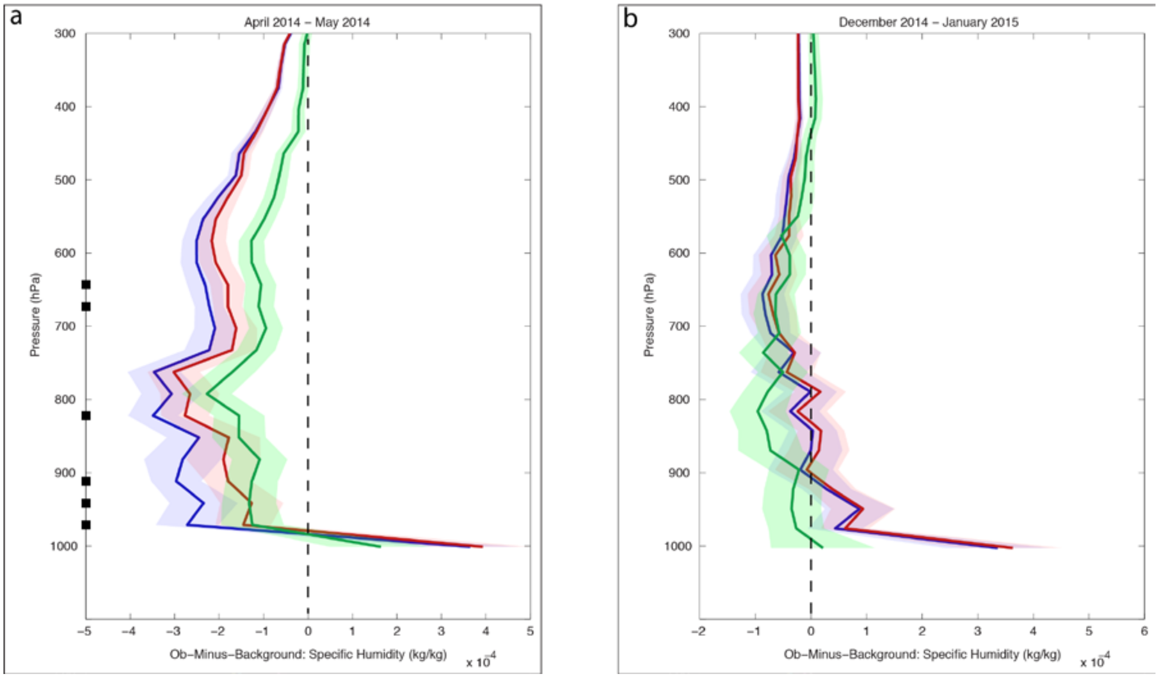

Figure 1. Mean profiles of specific humidity ob-minus-background (OMB) for the warm-season experiment (left) and cold-season experiment (right) at rawinsonde launch sites. The blue profile is the mean rawinsonde moisture OMB when AMDAR moisture observations are not assimilated. The red profile is the mean rawinsonde moisture OMB when AMDAR moisture observations are assimilated. The green profile is the mean AMDAR moisture OMB. The shading around each profile represents the 5% and 95% confidence limits around the mean, and pressure-levels where the rawinsonde OMB changes to statistical significance are highlighted with black squares along the ordinate.

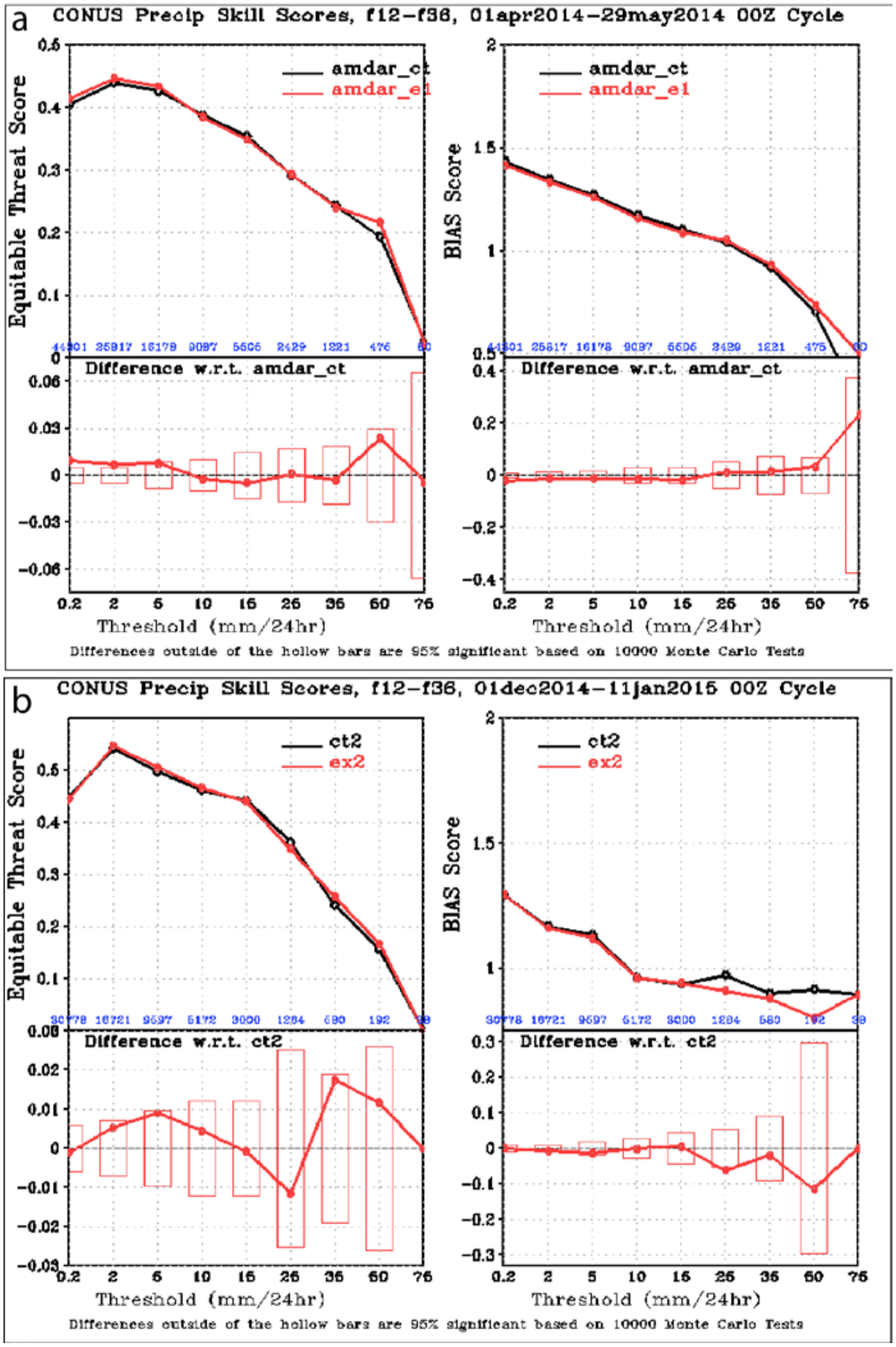

Figure 2. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for (a) warm-season experiment, and (b) cold-season experiment. The left panel of each plot shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel of each plot shows the precipitation bias score in the same bins. The black curve is for the control simulation, and the red curve is for the experiment. The bottom panels show the differences between the experiment and control, with bars indicating the minimum value necessary for 95% statistical significance.

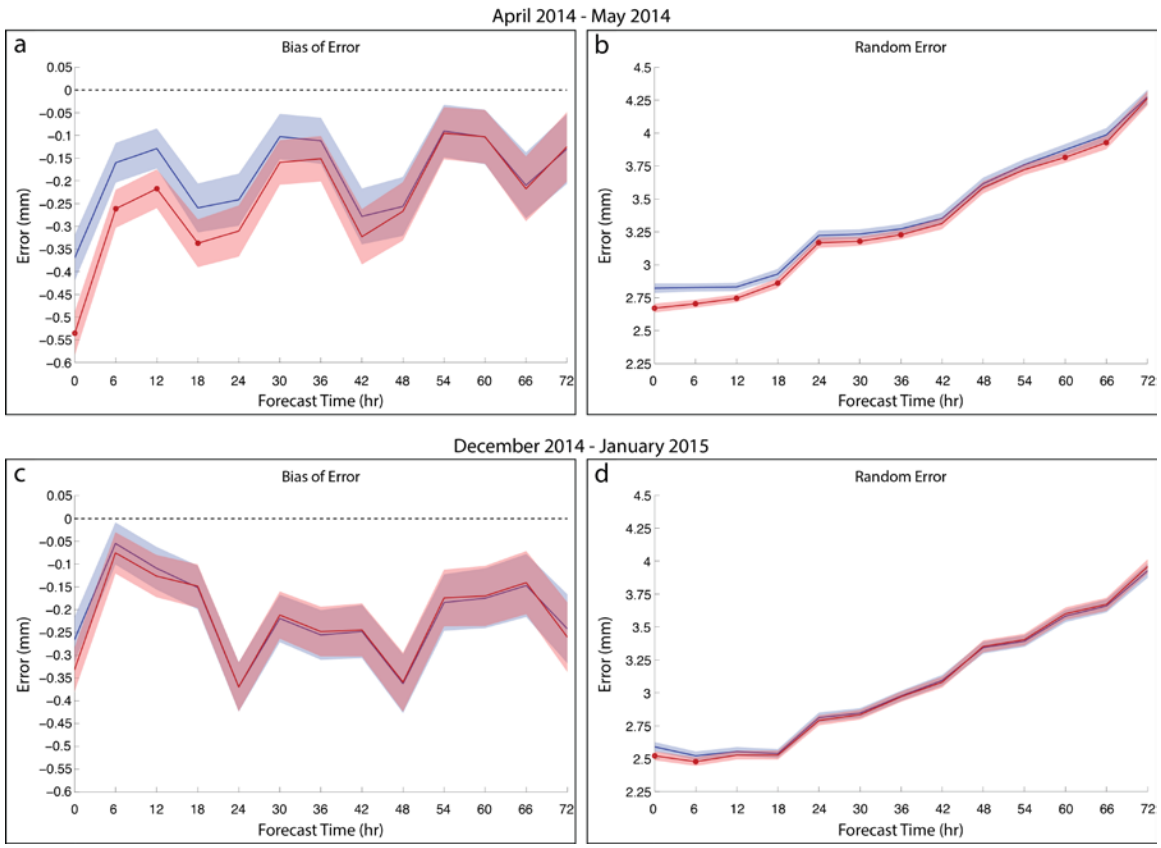

Figure 3. Error in forecast fit-to-TPW observations from GPS for (top) April 2014 – May 2014 simulation and (bottom) December 2014 – January 2015 experiment. Statistics for the control simulation are provided in blue, and statistics for the experiment are provided in red. Error is computed as (left) bias of error, calculated as the mean error, and (right) random error, calculated as the standard deviation of the error. Thick contours represent the sample mean or standard deviation, and the shading represents the 5% and 95% confidence limits on the mean or standard deviation. Dots are placed on the red contour for all times where the difference between the experiment and control is statistically significant based on a student's t-test (for bias of error) or a chi-squared test on variance (for random error).

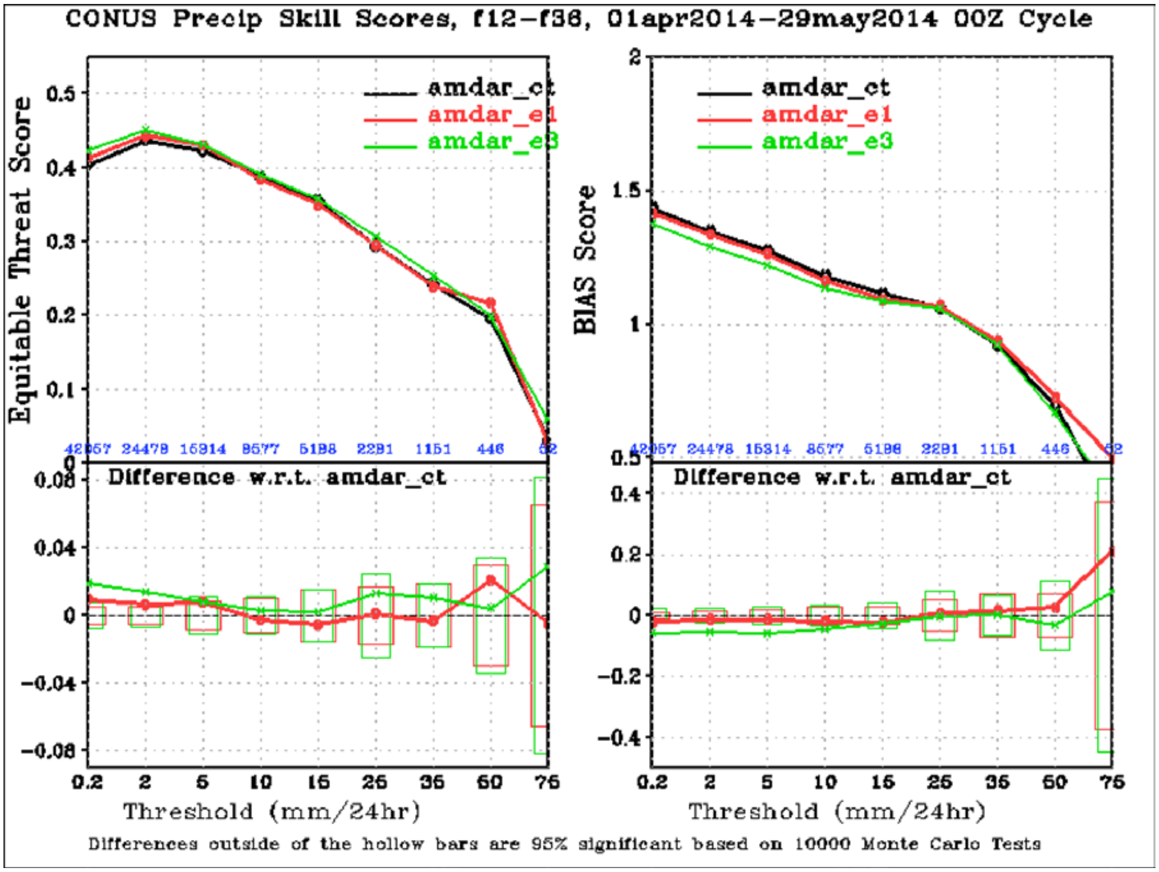

Figure 4. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for the warm-season assimilation experiment and data-denial experiment. The left panel shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel shows the precipitation bias score in the same bins. The black curve is for the control simulation, the red curve is for the assimilation experiment, and the green curve is for the data-denial experiment. The bottom panels show the differences between each experiment and the control, with bars indicating the minimum value necessary for 95% statistical significance.

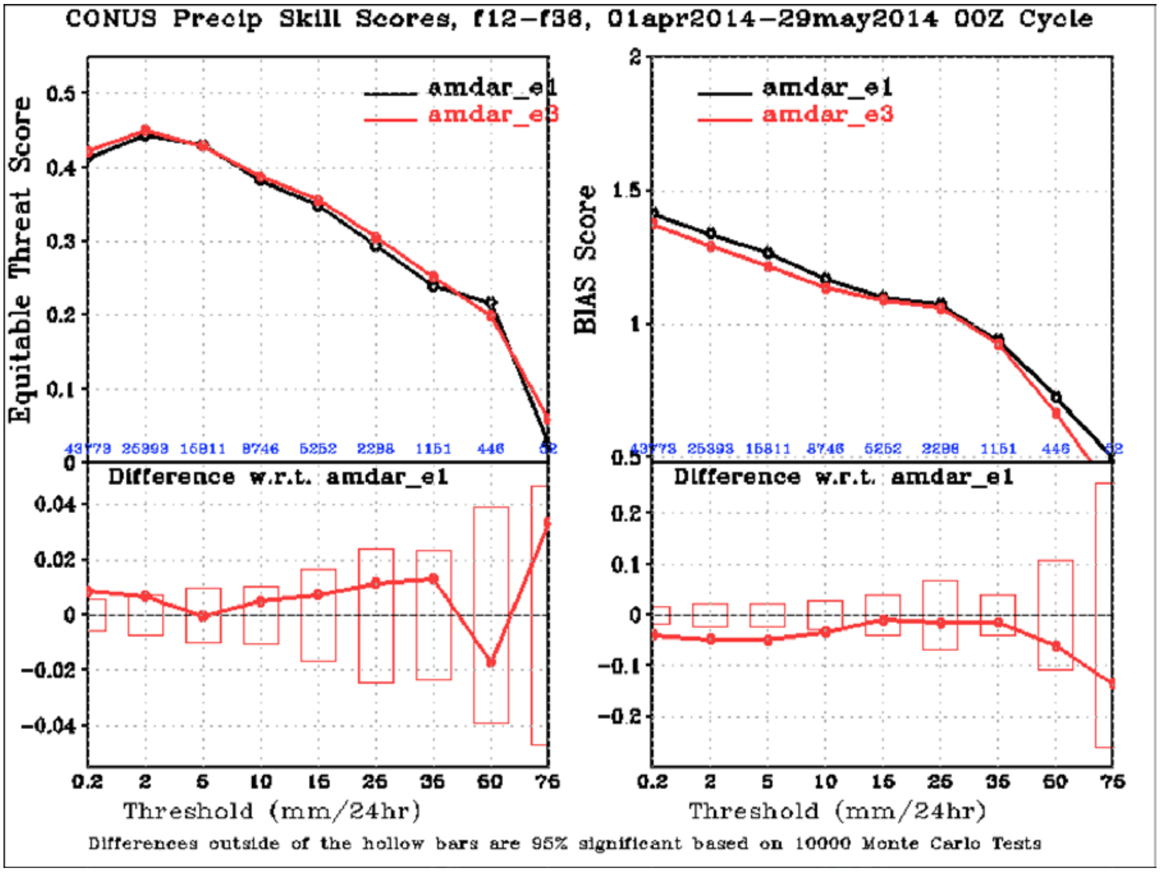

Figure 5. Precipitation skill and bias scores of 12-36 hour forecast over the continental United States for the warm-season assimilation experiment and data-denial experiment. The left panel shows the Equitable Threat Score (ETS) for precipitation binned by precipitation amounts in mm/24 hours. The right panel shows the precipitation bias score in the same bins. The black curve is for the assimilation experiment, and the red curve is for the data-denial experiment. The bottom panels show the differences between the two experiments (data-denial experiment minus assimilation experiment), with bars indicating the minimum value necessary for 95% statistical significance.

689

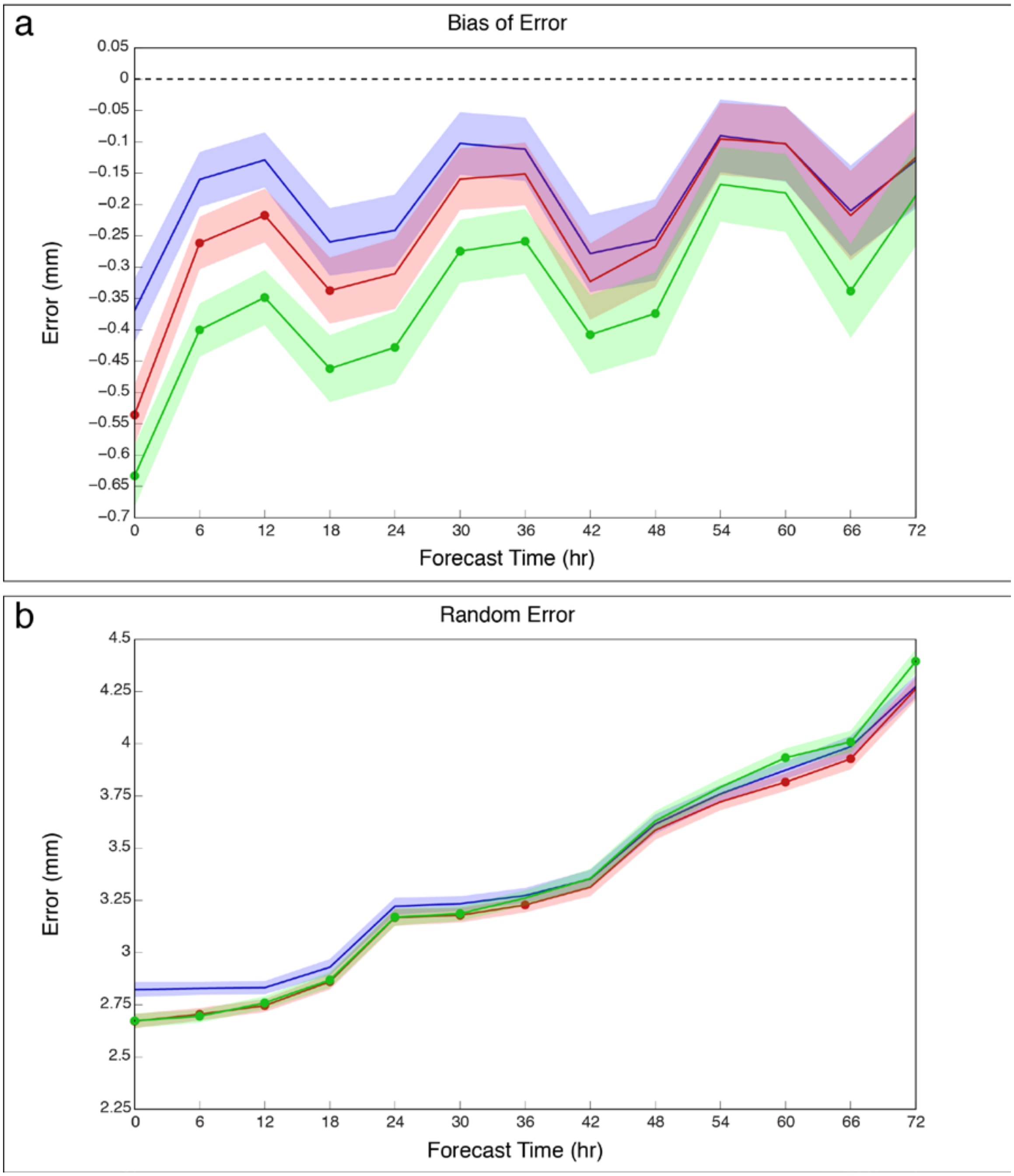

Figure 6. Error in forecast fit-to-TPW observations from GPS for April 2014 – May 2014 simulations. Statistics for the control simulation are provided in blue, statistics for the assimilation experiment (rawinsondes, aircraft moisture observations) are provided in red, and statistics for the data-denial experiment (selected rawinsondes removed, aircraft moisture observations) are provided in green. Error is computed as (top) bias of error, calculated as the mean error, and (bottom) random error, calculated as the standard deviation of the error. Thick contours represent the sample mean or standard deviation, and the shading represents the 5% and 95% confidence limits on the mean or standard deviation. Dots are placed on the red and green contours for all times where the difference between the experiment and control is statistically significant based on a student's t-test (for bias of error) or a chi-squared test on variance (for random error).

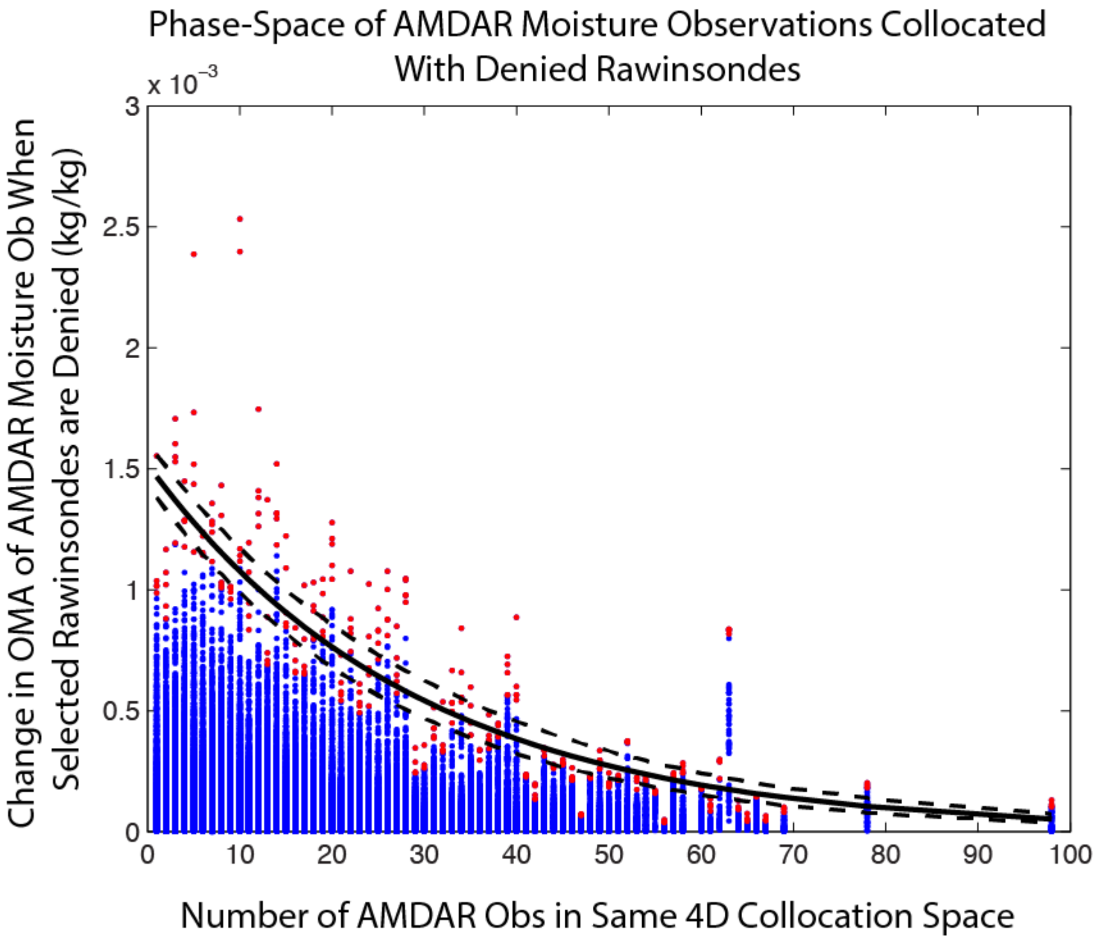

Figure 7. Phase-space diagram of relationship between (ordinate) how much denied rawinsondes impact the assimilation of nearby AMDAR moisture observations, and (abscissa) how many AMDAR moisture observations are nearby. Each dot (red or blue) represents an AMDAR moisture observation assimilated within 1 hour and 0.5 degrees of a denied rawinsonde, for all 0000 UTC and 1200 UTC analysis periods in the data-denial experiment. The ordinate measures the absolute value of the difference in Observation-Minus-Analysis (OMA) between the assimilation experiment (where AMDAR moisture observations are assimilated and all rawindsondes are maintained) and the data-denial experiment (where AMDAR moisture observations are assimilated and selected rawinsonde observations are denied). The abscissa measures the number of AMDAR moisture observations collocated to the same rawinsonde within the same vertical pressure layer. The red dots represent the 5 highest OMA differences for each unique value along the abscissa, identifying the upper bound of the phase-space that is sampled by the observations. The solid black line is an empirically-derived exponential best-fit to the red dots, representing a theoretical expected upper-bound on the potential impact of denied rawinsondes as a function of the density of AMDAR observational coverage. The dashed black lines represent the 5% and 95% confidence bounds on the solid line.