Observation System Experiments with the Hourly Updating Rapid Refresh Model Using GSI Hybrid Ensemble–Variational Data Assimilation

ERIC P. JAMES
Cooperative Institute for Research in Environmental Sciences, University of Colorado, and NOAA/OAR/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

STANLEY G. BENJAMIN
NOAA/OAR/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

(Manuscript received 18 October 2016, in final form 9 March 2017)

ABSTRACT

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 pseudo-innovations 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.

1. Introduction

Estimates of current atmospheric conditions are based on observations, in situ and remotely sensed, from an always evolving range of platforms. Consistency with physical relationships for these estimates is added through a time-evolving model assimilation cycle in a numerical weather prediction (NWP) system. The volume of meteorological observations is constantly increasing both from remote sensing platforms [satellites (Le Marshall et al. 2007) and radar] and also from some in situ platforms such as commercial aircraft. The geographical and temporal distribution of observations has a major impact on the initialization of NWP systems. In regions with large data voids, such as the Southern Hemisphere and the global oceans, model forecast skill is significantly degraded (Shapiro and Thorpe 2004; McMurdie and Mass 2004) although decreasingly as data assimilation techniques improve (e.g., Cardinali 2009). NWP models, both in the United States and internationally, are constantly improving in a variety of ways, including the recent ubiquitous increases in spatial resolution, the incorporation of more advanced data assimilation techniques, and improved physical parameterizations for subgrid-scale processes (Kalnay et al. 1998; Benjamin et al. 2016). However, all NWP systems remain fundamentally tied to the number and quality of the observations being assimilated (Hollingsworth et al. 1986). In the United States, 2012’s Hurricane Sandy disaster prompted a large investment in research and development related to
improving forecasts; the largest fraction of this “Sandy Supplemental” funding is directed toward improving observing systems (NOAA 2013). Questions remain regarding the most cost-effective path toward improved forecasts: What balance of investment is needed between more observational coverage versus an investment in high-performance computing infrastructure for NWP versus improved data assimilation and model technique development? Which observing systems will best add forecast accuracy? Observation system experiments (OSEs) provide a means of gauging the relative impact of different existing observation types on NWP forecast skill (Arnold and Dey 1986). Although the recently developed forecast sensitivity approach to observation impact (FSOI) can also assess observation impacts (Langland and Baker 2004; Zhu and Gelaro 2008; Lupu et al. 2015), this technique requires the development of an adjoint model for the data assimilation (Langland et al. 2016), which does not yet exist for the Gridpoint Statistical Interpolation (GSI) analysis system used for RAP. The FSOI method is a promising avenue for assessing short-range observation impact, but was not used here.

In the modern heterogeneous observing system, most observation types are frequently updated (i.e., new observational data become available every ~1 h or less). Therefore, in order to make the best use of recent observations, there is a corresponding need to frequently update the analysis and forecast from an NWP system. In the United States, the operational rapidly updating NWP model was upgraded from the Rapid Update Cycle (RUC; Benjamin et al. 2004a) to the Rapid Refresh (RAP; Benjamin et al. 2016, hereafter B16) model in May 2012. Subsequent versions of the RAP were implemented in February 2014 and August 2016. This most recent version of the model [RAP version 3 (RAPv3)], including hybrid ensemble–variational assimilation (e.g., Wang et al. 2013), was tested in this study. OSEs were conducted for the previous RUC model (e.g., Benjamin et al. 2010, hereafter B10), but as yet no published studies have investigated the observation impacts within the RAP system. NWP data assimilation and model upgrades (such as the 2012, 2014, and 2016 upgrades to the RAP) necessitate new OSEs in order to determine the impact of observation types within a modified NWP system. The observation mix has also changed over time, including a significant increase in aircraft data over the United States in recent years (WMO 2016). In this study, we aim to determine the relative impact of many of the observation types assimilated within the RAP modeling system partly to guide possible decision-making regarding the expansion of networks. We want to know whether the relative importance of observation types has changed compared to earlier OSE studies related to rapidly updating NWP systems.

The next section details the RAP model configuration used for the multiseason OSEs, followed by a section describing the experimental design for the OSEs. We then present our results and end the paper with our key conclusions.

2. Rapid Refresh model configuration

This section provides an overview of the RAP NWP system (B16), including its data assimilation component, its model component, and details related to its real-time application within NOAA. The volume of observations being ingested with each hourly cycle is then briefly summarized, highlighting some of the more important observation types, and followed by a summary of the most important differences from the RUC system used for the B10 OSE study.

a. NWP system overview

The version of the RAP used in these experiments is RAPv3, run operationally at the National Centers for Environmental Prediction (NCEP) starting in August 2016 and as described in B16. This version uses version 3.6.1 of the community-supported Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008). The RAPv3 uses a 954 × 835 model domain covering all of North America with a horizontal grid spacing of 13 km near the center of the domain (Fig. 1). The model computational grid is a rotated latitude–longitude grid, which reduces the stretching of the horizontal grid near the edges of the domain (Côté et al. 1993). The RAP has 50 hybrid vertical levels, with a model top at 10 hPa.

The data assimilation system used in the RAP is GSI, a community-supported data assimilation system (Kleist et al. 2009; Shao et al. 2016). The data assimilation is carried out in a hybrid three-dimensional ensemble–variational configuration (Wang et al. 2013), wherein the data assimilation background forecast covariances are a weighted average of the traditional 3DVAR covariances (Wu et al. 2002; Whitaker and Hamill 2002; Whitaker et al. 2008) and flow-dependent covariances derived from the 80-member Global Forecast System (GFS) data assimilation ensemble (25% variational and 75% ensemble in this version of RAP; more details are provided in Table 2 of B16). This hybrid ensemble–variational GSI-based approach for RAP (Hu et al. 2017, manuscript submitted to Mon. Wea. Rev.) results in improved forecasts (particularly for upper-level winds; B16) compared to the 3DVAR data assimilation carried out in the RUC model. After the application of the primary hybrid data
assimilation, a nonvariational cloud and hydrometeor analysis is carried out based primarily upon satellite and surface-based ceilometer data (also described by B16).

After the data assimilation process, and before the WRF free-forecast integration, a diabatic digital filter initialization procedure (DFI; Peckham et al. 2016, Huang and Lynch 1993) is applied within WRF. This process acts to damp out the noise effects of spurious gravity waves during the first hour of the forecast. The diabatic portion of the filter introduces specified latent heating where available during the DFI period; this specified latent heating is derived from three-dimensional radar reflectivity observations and, to a much smaller extent, is augmented by lightning observations (B16). This procedure is a low-cost way of producing an initial field with existing mesoscale circulations related to precipitation processes. The RAP system uses a variety of physical parameterizations within the WRF framework, also described in detail in B16. To keep regional forecasts from drifting away from the truth, the RAP employs partial cycling (Rogers et al. 2009) to reintroduce the GFS atmospheric state twice a day, at 0900 and 2100 UTC. Figure 2 shows a schematic of the RAP partial cycling configuration; the GFS atmospheric state is “partially cycled” for 6 h before being injected into the RAP full cycle. Within the context of
TABLE 1. Observational data used in RAPv3 as of Sep 2015: $p$ is air pressure, $q_y$ is water vapor mixing ratio, $T_v$ is virtual temperature, RH is relative humidity with respect to water, $V$ refers to horizontal wind components, $T$ is temperature, $p_s$ is surface pressure, $T_d$ is dewpoint, and PW is precipitable water. Table is from B16.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Variables</th>
<th>No.</th>
<th>Frequency (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rawinsonde (including special observations)</td>
<td>$T$, RH, $V$, $z$</td>
<td>125</td>
<td>12</td>
</tr>
<tr>
<td>NOAA 405-MHz wind profiler</td>
<td>$V$</td>
<td>$-21$ (2013)</td>
<td>1</td>
</tr>
<tr>
<td>Boundary layer (915 MHz) profiler</td>
<td>$V$, $T_v$</td>
<td>$-30$</td>
<td>1</td>
</tr>
<tr>
<td>Radar-VAD winds (WSR-88D)</td>
<td>$V$</td>
<td>125</td>
<td>1</td>
</tr>
<tr>
<td>Radar</td>
<td>Reflectivity, radial wind</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Lightning</td>
<td>Flash rate converted to reflectivity</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Aircraft</td>
<td>$V$, $T$</td>
<td>$3000–25000$</td>
<td>1</td>
</tr>
<tr>
<td>Aircraft</td>
<td>$q_v$</td>
<td>$0–1000$</td>
<td>1</td>
</tr>
<tr>
<td>Surface: METAR, land</td>
<td>$V$, $p_s$, $T$, $T_d$</td>
<td>$2200–2500$</td>
<td>1</td>
</tr>
<tr>
<td>Surface: METAR, land</td>
<td>Ceiling/visibility</td>
<td>$2200–2500$</td>
<td>1</td>
</tr>
<tr>
<td>Surface: mesonet, land</td>
<td>$V$, $p_s$, $T$, $T_d$</td>
<td>$10000–16000$</td>
<td>1</td>
</tr>
<tr>
<td>Buoy/ship</td>
<td>$V$, $p_s$</td>
<td>$200–400$</td>
<td>1</td>
</tr>
<tr>
<td>GOES AMVs</td>
<td>$V$, $p$</td>
<td>$2000–4000$</td>
<td>1</td>
</tr>
<tr>
<td>GOES cloud top</td>
<td>$p$, $T$</td>
<td>$-10$-km resolution</td>
<td>1</td>
</tr>
<tr>
<td>AMSU-A/HIRS-4/Microwave Humidity Sounder (MHS)/GOES</td>
<td>Radiances</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPS precipitable water</td>
<td>PW</td>
<td>300</td>
<td>1</td>
</tr>
</tbody>
</table>

for NWP data assimilation. Near-real-time cloud properties suitable for data assimilation are derived hourly from Geostationary Operational Environmental Satellites (GOES; Minnis et al. 2008) using the methods of Minnis et al. (2011). In addition, atmospheric motion vectors (AMVs) are derived from GOES and other satellite observations (Velden et al. 2005; Bresky et al. 2012); these observations provide important tropospheric wind information in otherwise data-sparse regions. Vertically integrated water vapor observations can be derived from global positioning system (GPS) receivers; an extensive network of GPS receivers around the world provides precipitable water (PW) observations for NWP data assimilation (Wolfe and Gutman 2000). These observations, referred to as GPS-met PW observations, have been shown to play an important role in NWP forecast skill through their ability to depict the distribution of tropospheric water vapor (Gutman and Benjamin 2001; Smith et al. 2007). Integrated PW is dominated by changes in low-level water vapor, which is critical for convective weather forecasting (e.g., Weckwerth 2000; Weckwerth et al. 2004). The impact of satellite radiances, also assimilated in the RAP, is relatively small but significant and is described in separate articles by Lin et al. (2017) and Lin et al. (2017, manuscript submitted to Wea. Forecasting).

Radar observations from the Weather Surveillance Radar-1988 Doppler (WSR-88D) Next Generation Weather Radar (NEXRAD) network provide critical information about the distribution of hydrometeors in the atmosphere. Development of a real-time distribution system for these large datasets (e.g., Kelleher et al. 2007) has permitted their use in NWP initialization. In addition to the radar reflectivity and radial velocity
observations available from the WSR-88Ds, profiles of wind speed and direction can be derived using the velocity azimuth display (VAD; Lhermitte and Atlas 1960) technique. These wind profiles provide another dataset available for assimilation within NWP systems.

To analyze upcoming observation impact experiment results, it is important to identify the actual measurements made by different platforms as shown in Table 2. Aircraft, rawinsonde, and surface measurement platforms include separate instruments to determine wind, temperature, and relative humidity (RH) observations. Observation impact is across variables, not univariate; observations of a single variable can impact other variables through multivariate correlations in the data assimilation system.

c. Major changes from the RUC system in 2007 to the RAP system in 2016

The version of the RAP NWP system used in this study (RAPv3; B16) is a substantially more advanced modeling system than the 2007 RUC system used for the OSEs presented by B10. In this section, we summarize some of the key observation, data assimilation, and model improvements (from the perspective of reducing forecast error) from the March 2007 RUC system to the 2016 RAPv3 system.

Regarding observations, in this study, radar reflectivity observations are now assimilated using latent heating specification during a preforecast application of forward–backward digital filter initialization (B16, section 2d). Satellite radiance observations are assimilated in RAP but were not in RUC (B16), and the volume of aircraft data has increased globally by a factor of 4–5 (dominated by a U.S. increase) since 2007 (WMO 2016). The NOAA Profiler Network (Benjamin et al. 2004b) with about 30 wind profilers was shut down in 2013; however, other multiagency profilers continue to operate in 2016. The data assimilation for hourly updated models has been significantly improved since 2007 (RUC 3DVAR then), now using hybrid ensemble–variational assimilation with GSI and improved assimilation of cloud, surface, and radar observations in RAPv3. Observation errors (OEs) remain essentially unchanged, with both the RUC and the RAP using an OE specification from the North American Mesoscale Forecast System (NAM) that is appropriate for a regional NWP system. The dynamical core for the RAP is WRF-ARW, while the physics parameterizations used in RAP are improved versions of those used in RUC and developed for short-range effectiveness.

The RAPv3 has a greatly expanded domain from the previous RUC domain, comprising approximately 4 times the RUC horizontal area (Fig. 1). Most of the domain expansion was toward the north and west, such that the 2016 RAP domain covers all of the Canadian Arctic region, as well as the entire Aleutian Islands chain. This domain expansion reduces the influence of the GFS boundary conditions and farther removes much of the CONUS from domain boundary influences.

3. Observation impact experimental design

For this study, observation impact experiments were configured similarly to B10. All of the assessments were conducted for three 10-day periods over three seasons (Table 3). The spring period (15–25 May 2013) featured several major severe weather outbreaks in the

<table>
<thead>
<tr>
<th>Season</th>
<th>Period</th>
<th>Characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>15–25 May 2013</td>
<td>Frequent convective activity, dominated by large-scale forcing</td>
</tr>
<tr>
<td>Summer</td>
<td>15–25 Jul 2014</td>
<td>Frequent convective activity with less large-scale forcing and more frequent local-scale forcing</td>
</tr>
<tr>
<td>Winter</td>
<td>1–11 Jan 2015</td>
<td>Winter storms and severe cold in eastern North America</td>
</tr>
</tbody>
</table>
south-central United States, including significant tornadoes in Oklahoma on 19 and 20 May, and a high coverage of severe wind reports scattered throughout the southern and eastern CONUS during 21–22 May. Experiments were also carried out for summer and winter season periods (Table 3) to better assess the seasonal dependence of the observation impact and to obtain overall year-round estimates of impact.

a. Control and data-denial experiments

To configure each of the multiseason OSEs (Table 3), a control RAP cycle with all available observations was first conducted. After the completion of this control experiment, we conducted up to 14 separate data-denial experiments (9 for spring and 13 for winter), excluding various observation types from the data assimilation process. Table 4 presents the names and descriptions of the experiments. All of the 10-day experiments are begun in the same way, using surface fields from the 1-h forecast from the 0200 UTC real-time RAP cycle, and atmospheric fields from the GFS, to initialize a 6-h partial cycle (see Fig. 2). From 0300 UTC on the first day through 2300 UTC on the final day, the data-denial experiments are configured identically to the control experiment except for the exclusion of certain observation types, with 12-h forecasts initialized every hour. Boundary conditions (from the GFS) are the same for all of the experiments. Note that our experiments did not include any OSEs on satellite radiance assimilation; the impact of these observations is discussed by Lin et al. (2017) and Lin et al. (2017, manuscript submitted to Wea. Forecasting).

During the course of executing the OSEs summarized in Table 4, several additional experiments were carried out in order to investigate in more detail the impact of assimilating surface observations and satellite AMVs (Table 4). These experiments, discussed in sections 6 and 7, led to the development of revised RAP configurations, which then were used as control experiments for examining the “no surface” and “no AMV” OSEs. Each of the experimental pair error differences shown herein is thus a controlled experiment, but modified control runs are used for calculating forecast degradation for the no-surface and no-AMV OSEs; for more details, the reader is referred to sections 6 and 7.

b. Verification

For all of the retrospective experiments, model forecasts were verified against conventional, twice-daily rawinsonde observations over the lower 48 United States (or CONUS) as well as over the entire RAP domain (Fig. 1). For each experiment, residuals (forecasts minus observations) were calculated for 3-, 6-, 9-, and 12-h forecasts at each rawinsonde location. The root-mean-square difference (RMSE; magnitude of the root-mean-square vector difference in the case of wind) between the forecasts and observations was then calculated for each 12-h rawinsonde verification time (0000 and 1200 UTC). All of the verification presented here uses observations...
and model forecasts interpolated to vertical resolution every 10 hPa from significant-level rawinsonde observations and native model data. For more detail on the verification procedure used here, the reader is referred to Moninger et al. (2010).

We estimated the forecast error reduction percentage for interpreting subsequent results in this study. Examining longer-term error statistics from the RAP (approximated as the difference in fit to rawinsonde observations between 6-h RAP forecasts and 0-h RAP analyses during 2015; Fig. 3) provides context for the forecast improvements described. This difference is used to estimate an approximate 25% reduction in forecast error for wind, temperature, and RH (Table 5). The analysis fit to rawinsonde observations is used as a proxy for truth, since that analysis (see B16) includes information on expected observation errors for rawinsondes and the other observation types also assimilated. From this 2015 result, we estimate 25% of the maximum possible improvement in 6-h forecast skill as 0.3 m s\(^{-1}\) for wind, 0.1 K for temperature, and 1% (on the 0–100 scale) for RH (included later in Fig. 7).

As mentioned in section 2c, the RAP system uses partial cycling whereas the previous RUC (used for the OSEs of B10) did not. Since the verified forecasts are valid only at 0000 and 1200 UTC, the timing of the partial cycle introduction into the full cycle affects the influence of certain observation types. For example, for 3-h forecasts valid at 1200 UTC (initialized at 0900 UTC), only 7 h of RAP data assimilation have occurred (six during the partial cycle from 0300 to 0800 UTC, and once for the full cycle at 0900 UTC; cf. Fig. 2). Older observations have an influence only through the GFS initial conditions introduced at 0300 UTC; thus, we would not expect to see any impact of denying (within the RAP) the assimilation of observations before 0300 UTC. Since rawinsondes are primarily available only at 0000 and 1200 UTC, we would not expect to see any impact of rawinsondes in 3-h forecasts in our experiments. In general, 6-h forecasts are expected to show the strongest impact of observations since they reflect the longest period of RAP data assimilation, and impacts should also generally decrease with increasing forecast lead time as the model initial conditions become less important. All experiments carried the same specified set of observations (or denial) through the partial cycling and the primary cycle with the RAP forecasts.

Table 5. Normalizing percentage difference for 6-h forecast minus 0-h analysis for wind, temperature, and RH. Based on Benjamin et al. [2004a], their Figs. 9 and 10 and Eq. (11)]. A simple level-independent approximation is made for all levels, and values are averaged over all levels to determine an approximation for a 25% error reduction level.

<table>
<thead>
<tr>
<th>Approx mean fit to rawinsonde observations</th>
<th>Wind (1000–100 hPa, m s(^{-1}))</th>
<th>Temp (1000–100 hPa, K)</th>
<th>RH (1000–400 hPa, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-h forecast</td>
<td>-4.2</td>
<td>-1.4</td>
<td>-16</td>
</tr>
<tr>
<td>0-h analysis (best fit to obs)</td>
<td>-3.0</td>
<td>-1.0</td>
<td>-12</td>
</tr>
<tr>
<td>6-h forecast minus 0-h analysis</td>
<td>-1.2</td>
<td>-0.4</td>
<td>-4</td>
</tr>
<tr>
<td>(max possible improvement)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25% of max possible improvement</td>
<td>-0.3</td>
<td>-0.1</td>
<td>-1</td>
</tr>
</tbody>
</table>
3. Experiment results

In this section, we present our verification results for the OSEs in a similar format as B10, verifying upper-air RAP forecasts against 12-h rawinsonde observations. Results are presented in the form of “candlestick plots,” wherein each data-denial experiment is assigned a bar height associated with the degradation in RMSE (averaged against all observations within the region, vertical layer, and time of interest) as compared with the control run. Within this framework, the data-denial experiments are anticipated to generally have positive bars, indicating an increase in RMSE (or a degradation in forecast skill) as compared to the control run. Negative bars indicate that denying that particular dataset actually results in a reduction in RMSE (or a forecast improvement). Error bars are added to figures in this study to show the range of one standard error (i.e., the 67% confidence interval) calculated using the method of Weatherhead et al. (1998), as applied and described by B10. The standard error is calculated from the sample standard deviation and the lag-1 autocorrelation for the time series of RMSE differences as verified against rawinsonde observations (following the method of B10).

a. Overview results for the full depth (1000–100 hPa) over North America

As in B10, we begin with a broad view of the results, considering first the impact of the different observation types for each of the three experiment seasons within a vertically integrated column: 1000–100 hPa for wind (Fig. 4) and temperature (Fig. 5), and 1000–400 hPa for RH (Fig. 6), through the entire RAP domain (see Fig. 1).

For wind forecasts in this full tropospheric (and lower stratospheric) 1000–100-hPa layer, out of the 10 observation types, aircraft clearly have the largest impact and similarly over all three seasons: 0.3 m s$^{-1}$ for 6-h forecasts (Fig. 4). Aircraft data reduce tropospheric wind error by 25% over the RAP domain (per Table 5). GOES and raobs gave a positive but far smaller impact of 0.05–0.10 m s$^{-1}$ during each of the three seasons. Surface observations gave similar impacts for wind forecasts over this full 1000–100-hPa layer in the spring and summer seasons, with less impact in winter. All of the cited observation types (except for mesonet and wind profiler observations) showed significant impact at the 67% level during the spring and summer seasons; winter observation impacts are much less for most observation types.

For temperatures through the deep 1000–100-hPa layer (Fig. 5), aircraft observations had the largest impact: 0.10–0.14 K at 6 h for all seasons. (A value of 0.1 K corresponds to a $\sim 25\%$ error reduction.) Rawinsonde and surface observations had a somewhat reduced impact: 0.02–0.06 K, which is about a 5%–12% error reduction. The impact of many of the other observation types is quite muted in all seasons.

For RH through the 1000–400-hPa layer, a broader combination of observation types contributes to improved...
forecast skill, led by aircraft observations at 0.5%-0.7% RH for 6-h forecasts, but joined by GOES (with both cloud-clear data and AMVs; Table 2). Rawinsonde and surface observations both contribute by 0.2%-0.5% RH for 6-h forecasts, and even GPS PW and radar reflectivity results show a positive impact to RH forecast skill through this relatively deep (1000–400 hPa) layer (Fig. 6).

Results from the multiseason periods were combined to produce the summary statistics given in Fig. 7. For all variables, aircraft observations gave the largest impact at most forecast hours, contributing to a 25% reduction (calibrated from Table 5) in wind and temperature forecast error at 3 and 6 h, and about a 10%-15% error reduction for RH forecasts. Surface observations emerge as a leading contributor to forecast skill for all three variables, with an impact comparable to that of rawinsondes. An additional experiment (not shown) denying the assimilation of only surface pressure observations revealed that nearly all of the positive impacts of surface observations come from temperature, moisture, and wind observations (with the exception of 1000–800-hPa wind at night). GOES satellite observations have an RH impact similar to that from aircraft, with a reduced impact on winds (relative to aircraft) and negligible change for temperature. Minor but significant positive wind and RH impacts are evident for all other observation types, with the possible exception of wind profiler observations and VAD winds (see below).

These integrated statistics on forecast impact were also examined for temporal and vertical consistency, as shown in Fig. 8. The impact of denying aircraft data was approximately similar for each 12-h rawinsonde observation time...
through the summer July 2014 period, and the vertical profile for the overall RMS vector error and aircraft impact is similar to that shown in B10. All of the results shown in this study have been examined similarly (not shown), and the temporal consistency over each seasonal test period as in Fig. 8 demonstrates that the 10-day test periods are representative for each of the seasonal evaluations.

b. Vertical variation of observation forecast impact

This section examines the vertical variation of the forecast impacts described in section 4a. The vertical variation of the wind forecast impact is shown in Fig. 9. For the summer period (Fig. 9c), aircraft observations once again dominate for mid- and upper-level wind forecasts, with an impact exceeding 0.4 m s$^{-1}$ at 6 h in the 400–100-hPa layer. This represents a reduction in forecast error of approximately 1/3. Surface observations contribute the most strongly for near-surface wind forecasts (with aircraft observations taking over above 900 hPa; not shown); however, the impact of surface observations extends into the 800–400-hPa layer and, surprisingly, even above 400 hPa (Figs. 9b,c; see section 6). Above 800 hPa, surface observations exhibit their peak impact at somewhat longer lead times (6–9-h forecasts). Mesonet observations produce a minor positive impact below 400 hPa. GOES satellite observations represent the second-most important observation type for winds in the 400–100-hPa layer, with an impact of 0.18 m s$^{-1}$ (or about 15%) at 6 h (Fig. 9c). AMVs contribute a 0.05 m s$^{-1}$ impact, statistically significant but relatively small compared to the overall GOES impact.
(combined AMV and cloud). Note that the small impacts seen for some other observation types are also nevertheless often statistically significant, as a result of the large sample size. These small but significant impacts for summer upper-level winds from surface, GPS PW, and radar reflectivity data are attributed to their contributions toward improved convection forecasts.

The majority of the GOES satellite observation impact on upper-level wind apparently comes from non-AMV observations (i.e., GOES cloud observations; see Table 2). One possibility is that this effect on wind is coming from GOES cloud (and indirectly RH) influences on the convective environment, since this is a summertime period. However, Figs. 9f,i reveal that non-AMV GOES observations continue to have a strong impact on upper-level winds in the spring and even winter seasons. The vertical profile of GOES observation (cloud and AMV) impacts (Fig. 10) shows the difference in the RH bias profile from the control experiment versus the no-GOES experiment for both summer and winter. It is apparent that the GOES cloud observations are acting to dry the column, particularly in the 600–100-hPa layer (through cloud analysis “clearing” where GOES observations indicate no cloud cover). Evaluation of precipitation forecast bias verified against Stage IV precipitation estimates (not shown) shows that for both summer and winter seasons, GOES satellite observations act to reduce the high precipitation bias in the RAP forecasts. Such changes to precipitation systems within the model atmosphere also affect the wind forecast skill in all seasons (e.g., Fig. 9).
Figure 11 shows a vertical layer-by-layer breakdown for temperature forecasts. The observation impact distribution for temperature forecasts is relatively simple, with only three major observational contributors (all in situ) to forecast skill. Aircraft observations again provide the strongest impact overall and are second only to surface observations in the 1000–800-hPa layer (Fig. 11a). Surface observations contribute to about a 0.1-K reduction of forecast RMSE at 3 h. Aircraft observations reach this level of impact above 800 hPa (Fig. 11b), with rawinsondes contributing to about a 0.05-K (relatively strong) reduction in forecast RMSEs for 6-, 9-, and 12-h forecasts above 400 hPa in the RAP domain (Fig. 11c).

Figure 12 shows additional layer-by-layer results, now for RH RMSE. In the near-surface 1000–800-hPa layer (Fig. 12a), surface and aircraft observations have an approximately equal impact, contributing to about a 0.3% RH RMSE reduction. Rawinsondes and GPS-met PW observations represent a secondary, but still significant, importance for the 1000–800-hPa layer. Within the 800–400-hPa layer, aircraft and GOES satellite (cloud) observations are the strongest contributors to RH forecast skill (Fig. 12b), each standing at 0.75% for 6-h forecasts (about a 15%–20% reduction in RMSE). VAD wind observations have a negative impact on RH forecast skill, particularly in the low to midlevels; we hypothesize that this is due to
contamination of horizontal wind observations by bird migration (mainly at night; see Fig. 13), which then has a subsequent effect on RH forecasts through increased error in the vertical motion. RH forecast errors could also result from horizontal wind errors leading to errors in the transport of features, particularly in regions of strong horizontal RH gradients. In the 400–100-hPa (essentially cirrus clouds, largely nonprecipitating) layer, GOES satellite observations overtake aircraft observations as the most important observation type (Fig. 12c), contributing to a 25% reduction in RMSE for RH, attributable to cloud (and related water vapor) data assimilation in the RAP (see B16, section 2e).

FIG. 12. As in Fig. 9, but for RH RMS error (% from 0 to 100) and for summer only.

FIG. 13. As in Fig. 7, but for summer only, and for the 1000–600-hPa layer (lower troposphere) results only, and for verification times (a)–(c) daytime (0000 UTC) and (d)–(f) nighttime (1200 UTC) for (a),(d) wind, (b),(e) temperature, and (c),(f) RH.
c. Diurnal variation of observation forecast impact

Observation impacts also exhibit diurnal variability (with the 0000 UTC raob verification time roughly representing afternoon/daytime atmospheric evolution across North America, and the 1200 UTC raob time representing largely nighttime evolution). To highlight diurnal variation occurring near the surface, we focus on the 1000–600-hPa layer for the summer test period (Fig. 13). During the daytime, there is a stronger impact for 3-h forecasts (2100 UTC initialization time) from surface observations, with a lower impact for 9–12 h (Figs. 13a–c). This is attributable to the planetary boundary layer (PBL) extension of surface observation assimilation in GSI enabled for RAP (B16, section 2f), with the deepest PBL typically occurring near 2100 UTC on average over the CONUS. Errors in forecast cloudiness could also contribute to obscuring the impact of the initial conditions on longer-lead-time low-level temperature forecasts. Neither of these factors apply at night, when stronger surface effects are seen from the 0000 UTC (12-h forecast) and 0300 UTC (9-h forecast) initializations (Figs. 13d–f). Most notable in the diurnal breakdown is the negative impact from VAD wind observations. The strong negative impact overnight on lower-tropospheric wind forecasts from VAD is likely attributable to bird and other biological activity (e.g., Lakshmanan et al. 2010; Wilczak et al. 1995), which may also be affecting overnight lower-tropospheric RH forecasts especially at 3 h, presumably through contamination of the vertical motion fields.

5. Aircraft observation impact breakdown for wind, temperature, and moisture

With the dominant impact from commercial aircraft observations shown in previous sections, a series of additional experiments were carried out to examine the contributions from different components of aircraft observations as shown in Table 4: (i) en route (defined here as pressure less than 350 hPa) versus ascent–descent (pressure greater than 350 hPa) and (ii) the specific contribution from aircraft moisture and temperature observations. We examine results for the full RAP domain. Given that the overall 1000–100-hPa aircraft impact was similar during the summer, spring, and winter seasons (Figs. 4–6), we focus here solely on the summer season.

Considering the tropospheric profile as a whole (Fig. 14), observations from en route aircraft and those from ascending–descending aircraft represent approximately equal contributors to the overall 1000–100-hPa wind and temperature forecast skill. Aircraft temperature observations evidently contribute indirectly to wind forecast skill (through multivariate background error covariances in GSI; Fig. 14a). Aircraft temperature observations also do not represent the only aircraft contribution to temperature forecast skill; around 0.3 K of the total 1.1-K positive impact for 6-h forecasts comes indirectly through aircraft wind observations, again through mass–wind background error covariances (Fig. 14b). The total aircraft observation impact for 1000–400-hPa RH is about 0.5% (a major contributor to overall observation impact; cf. Fig. 7). This entire impact comes from ascending–descending aircraft, and is apparently not caused in these experiments by the presence of abundant aircraft jet-level (<350 hPa) wind observations. Direct RH aircraft observations represent only about 1/3 of the total aircraft observation impact on RH forecasts. A vertical profile of RH forecast impact from aircraft observations overall versus aircraft RH observations only (Fig. 15) also confirms this...
1/3 fraction overall and up to 50%–60% in the 600–500-hPa layer. It is important to note that direct aircraft moisture observations reach this level of impact despite being available in only about 1/8 of the total number of aircraft observations assimilated within the RAP system. [These results included the Water Vapor Sensing System (WVSS; Petersen et al. 2016) aircraft water vapor observations, but not Tropospheric Airborne Meteorological Data Reporting (TAMDAR; Daniels et al. 2006) aircraft water-vapor observations; TAMMDAR data (Moninger et al. 2010) were not available to NOAA in real time during these test periods.]

Looking specifically at wind forecasts in the 400–100-hPa layer (where wind forecasts are particularly important for flight planning; Fig. 16), it is seen that approximately 2/3 of the total aircraft observation impact comes from aircraft above 350 hPa, with the remaining 1/3 from ascending–descending aircraft. A significant positive impact on upper-level wind accuracy (about 25% of the overall aircraft impact) also comes from aircraft temperature observations partially through the multivariate background error covariances mentioned above but also possibly from improved convection and general precipitation forecasts as surmised for moisture-related observations and discussed concerning Fig. 9c.

6. Surface observation impact and related sensitivity tests

The assimilation of surface observations is an important component of the RAP system. Results shown above in this study use an improved treatment of surface-based observations with assimilation refinements determined through earlier experiments performed for this study. These refinements and related sensitivity tests are now described in this section.

Surface observations influence the RAP analysis primarily through their use within the GSI data assimilation framework (for pressure, temperature, water vapor, and wind observations), but surface ceilometer observations are one of two cloud observation sources (along with GOES) for the nonvariational 3D cloud analysis (B16; section 2e). Unexpected results from preliminary

---

**Fig. 15.** Vertical forecast impact (similar to Fig. 8b) for 6-h RH forecasts valid at 0000 UTC over the CONUS from denial of aircraft observations overall (red) and denial of aircraft moisture observations only (blue). Results are from the July 2014 test period. Rectangular boxes containing values significant at the 95% level are also shown at each level.

---

**Fig. 16.** As in Fig. 14, but for 400–100-hPa (top level) vector wind RMSE.
OSEs conducted within this study led to refinements to the treatment of surface observations for improving forecast skill impacts. These new experiments (Table 4) are not data-denial experiments, as shown previously in this study, but are data assimilation sensitivity tests for refinements designed to improve the impact of surface observations. Differences in these pairs of experiments are shown with candlestick plots as previously. However, for tests of these potential data assimilation improvements, a negative impact bar indicates better forecast skill for that test relative to the control experiment. Figure 17 shows forecast skill differences in the 1000–600-hPa layer for these tests. Experiment H (in red) represents the impact from the denial of surface observations using the original configuration of surface assimilation. For wind forecasts, surface observations originally had a near-neutral effect (depending on the forecast length considered). A positive impact (25% of total error for 3 h) occurred for daytime 1000–600-hPa temperature forecasts valid at 0000 UTC, but RH forecasts gave a consistent negative impact of surface observations using that original design (experiment H).

Hypothesizing that the negative or small/neutral impacts of surface observations were due to the configuration of pseudoinnovations (a method of extending the influence of surface observations into the PBL in well-mixed environments; see section 2f in B16), an additional experiment was conducted in which all pseudoinnovations were withheld. Results from this experiment can be interpreted in the same way as a traditional OSE, with negative bars indicating forecast skill degradation coming from the pseudoinnovations. Figure 17 (experiment O, orange bar) confirmed that with the original configuration, the pseudoinnovations were having a negative forecast impact for all variables, particularly at shorter forecast lead times. This indicated that the earlier configuration applying pseudoinnovations was either wrong or incorrectly applied. Through a subsequent set of experiments, we determined that surface-based pseudoinnovations were applied too strongly and too high up into the PBL. An additional experiment with reduced pseudo-observation density within the PBL (from a spacing of 20–40 hPa) and a reduced maximum pseudoinnovation height [40% of the PBL height instead of 75%; PBL height calculated using virtual potential temperature as described by Benjamin et al. (2004c)] was conducted. Verification of this experiment (experiment P; Fig. 17) indicates that this configuration performs very similarly to experiment O (no pseudoinnovations), indicating that using a reduced pseudoinnovation density results in a superior forecast during the control experiment.

We also hypothesized that some of the originally neutral-to-negative surface observation impacts could be due to the use of its ceiling (ceilometer) data. Prior experiments not shown here had shown negative impacts from middle- to upper-atmosphere cloud building (but not clearing). In the configuration used in the control experiments, cloud building from GOES satellite observations was allowed only up to 1200 m above ground level (AGL), but cloud building from METARs was allowed at any level where METAR ceilometers report cloud cover [up to 12 000 ft (3657.6 m) AGL]. An additional experiment was carried out, now applying the same 1200 m AGL cloud-building height limit to METAR
observations as well as to satellite observations. Figure 17 shows that, indeed, RH forecast RMSEs and wind RMS vector errors are significantly reduced with this experiment (experiment Q, yellow). A slight degradation of temperature forecasts is seen at longer forecast lengths; this is likely related to the reduction of cloud cover during the daytime and the associated increase in solar irradiance reaching the ground.

To test these modified surface assimilation design features, one final experiment was conducted with both the reduced pseudo-observation density configuration and the cloud-building height limit for METARs. This experiment is shown in Fig. 17 (experiment R; purple); it is seen that this configuration results in significantly improved wind and RH forecasts in the 1000–600-hPa layer, with a near-neutral effect on temperature forecasts. This new configuration, with modified treatment of surface observations, is the control run used for reference for the “no surface” OSEs elsewhere in this paper (experiment Hi in Figs. 4–7, 9, and 11–13). Vertical profiles of surface observation impact before and after these two assimilation changes (Fig. 18) show the change in 6-h RAP forecast impacts from negative to generally positive from these two assimilation changes. These changes described in this section both reduced the vertical extension of these surface-based observations. This improved configuration was a beneficial consequence of these RAP OSEs and the investigation of initially counterintuitive results.

7. Satellite cloud-drift wind impact

The results from our initial OSE tests indicate only a very small impact coming from satellite cloud-drift winds (e.g., Fig. 7). This impact was considerably smaller than expected, motivating a more detailed examination of the use of cloud-drift winds in the RAP. Because of the dense observing network over the CONUS, and concerns regarding the quality control of satellite AMV observations over land, these data have historically been assimilated only over oceanic regions within the RUC and RAP systems. One additional experiment, allowing the assimilation of satellite AMV observations at least 100 hPa above the land surface (and requiring innovations for \( u \) and \( v \)-wind components to be less than 8 m s\(^{-1}\) at all levels) was conducted to determine whether additional AMV impacts could be achieved by assimilating these observations over land within the RAP system (e.g., Table 4).

Figure 19 shows vertical profiles of the multivariable impact on 6-h RMS error from these experiments. The original AMV OSE indicated a positive impact of approximately 0.04 m s\(^{-1}\) in the 400–100-hPa layer; we see that assimilating AMVs over land leads to an additional impact of approximately 0.01–0.02 m s\(^{-1}\), resulting in a net effect of about 0.05 m s\(^{-1}\) in the 400–100-hPa layer (Fig. 19a). Additional small increases in observation impact on wind RMSE are seen down to approximately 750 hPa. The impacts on temperature and RH forecasts are very small (Fig. 19). The experiment assimilating AMVs over land is used as the control experiment for the calculation of AMV impacts elsewhere in this study (experiment J in Figs. 4–7, 9, and 11–13), with a maximum AMV impact for 400–100-hPa wind in summer (Fig. 9c) of 0.05 m s\(^{-1}\) at 6-h duration, smaller than for some other observation types but statistically significant (per candlestick “wick”).

8. Verification of surface forecasts

While upper-level forecasts are important for aviation and other applications, surface weather forecasts are
needed by other user communities. Figure 20 shows several of the summer and winter experiments verified against METAR observations over the High-Resolution Rapid Refresh (HRRR) domain (see Fig. 1). Only the larger-impact experiments are shown in these plots. It is apparent that surface observations have a very strong impact on surface forecast skill, particularly for temperature and moisture forecasts. Surface observations have the strongest impact for forecasts initialized during the late afternoon and early evening during the summer. Aircraft observations have the second-largest impact, with the differences peaking for forecasts initialized during the daytime and exceeding those of surface observations during summer for 2-m temperature and 10-m winds from 1600 to 2000 UTC (daytime). GPS-met PW observations contribute modestly to dewpoint forecasts during the summer period, while rawinsondes contribute also only modestly to forecast skill with their

**Fig. 19.** As in Fig. 18, but showing the vertical forecast impact for the original no-AMV experiment (blue) and the experiment assimilating AMVs over land as well as over water (red).

**Fig. 20.** Average diurnal cycle of RAP 6-h forecast RMSEs vs METAR observations for (a),(d) temperature (K), (b),(e) dewpoint temperature (K), and (c),(f) vector wind (m s\(^{-1}\)) for the (a)–(c) summer and (d)–(f) winter periods, verified over the HRRR domain (see Fig. 1). Shown are the no-surface experiment (green), the no-raobs experiment (black), the no GPS-PW experiment (blue), and the no-aircraft experiment (red).
twice-daily launches. Other experiments not shown here generally exhibit smaller impacts.

9. Conclusions from new RAP OSEs in this study

An extensive set of observation system experiments for three seasons was conducted with the hourly updated Rapid Refresh model using hybrid ensemble–variational data assimilation. Experiments were conducted in each season for nine different observation types. It was found that the heterogeneous observation system over North America is effective for reducing 3–12-h tropospheric forecast errors, with contributions from all observing systems. Aircraft data were the most important observation type overall for short-range forecasts of wind, RH, and temperature with error reductions of 15%–30% for 6-h forecasts on average in the troposphere and lower stratosphere. Aircraft observations are followed by satellite (GOES) and surface observations. Also evident are strong cross-variable impacts between mass, wind, and moisture variables due, in part, to the GSI hybrid ensemble–variational data assimilation.

The greatest impact on wind forecasts comes from aircraft observations, followed by rawinsonde observations; this result over a North American domain is consistent with that found by B10 over a domain covering the lower 48 United States, except that rawinsondes have relatively less impact here. In addition, GOES satellite observations and surface observations contribute at a similar level as rawinsondes. The considerably lower relative impact of rawinsondes in this study is possibly due to the relatively steady volume of rawinsonde observations over the years, while other observation types have increased in coverage, quality, and inclusion via improved data assimilation.

For temperature forecasts, B10 had found that the impact of aircraft observations was approximately equal to that of rawinsondes, and that surface observations contributed at a comparable level during the summer. In this new study, while surface observations remain equally important to rawinsondes on overall temperature forecasts, aircraft observations now far outweigh either in importance.

For RH forecasts, B10 found that rawinsondes were the most important observation type, with aircraft, GPS PW, and surface observations playing a secondary role. In these new RAP OSEs, it is seen that aircraft, rawinsonde, and surface observations contribute nearly equally for 1000–400-hPa RH forecasts. GOES satellite observations also emerge as a major player for RH forecast skill, with lesser (but still significant) impacts coming from radar reflectivity and GPS PW observations. Radar reflectivity observations remain critical for the initialization of higher-resolution rapidly updated NWP systems but, evidently, also contribute to forecast skill at 13-km grid spacing.

A more detailed investigation of aircraft observation impact was made through additional experiments with subsets of aircraft data. For upper-level (400–100 hPa) wind forecast accuracy (important for aviation operations, for instance), about 1/3 of this impact was from ascent–descent aircraft reports below 350 hPa (with this fraction increasing with forecast lead time). Without temperature and RH aircraft observations, the upper-level short-range wind forecast skill impact overall from aircraft was decreased by about 1/3. Additional, more detailed, aircraft-data experiments revealed that about 2/3 of the significant aircraft impact on RH forecasts was from wind and temperature observations, but that this impact was increased by about another 50% with the inclusion of aircraft water vapor observations [Water Vapor Sensing System version two (WVSS-II); Petersen et al. (2016)] contained within approximately 1/8 of aircraft reports.

Significant cross-variable impacts from GOES satellite observations (atmospheric motion vectors and cloud-top retrievals) are evident within our experiments. This set of observations was not tested by B10; however, the results presented here suggest that GOES observations play an important role within the observation suite. GOES satellite observations influence the RAP NWP system through two avenues: AMVs derived from feature tracking of satellite imagery, and cloud information: the presence or absence and their height when present. It is evident from the results presented here that the greater influence by far comes from the GOES cloud information. Cloud building and clearing has a direct influence on RH forecasts through the cloud analysis (described in B16), but also apparently a fairly strong cross-variable influence on temperature and wind forecasts; we hypothesize that this is related to the improved evolution of precipitation systems in all seasons (see section 4b).

Initial results from the OSEs conducted in this study led directly to several improvements in the configuration of the surface observation assimilation within the RAP/GSI system. First, it was determined that the prior configuration of applying "pseudo-observations" within the model PBL was too heavy handed, exerting too strong an influence on the model initial conditions. A revised configuration was tested, in which pseudo-observations were inserted only every 40 hPa instead of every 20 hPa, and the pseudo-observations were only inserted up to 40% of the PBL height instead of 75%. In addition, we examined the impact of limiting the
METAR cloud building to 1200 m above ground level (the same limitation applied to cloud building based upon satellite observations). Both of these tests yielded favorable results, and future operational versions of the RAP and HRRR will take advantage of these developments.

Additional tests on AMV assimilation, motivated by the initially limited impact of these observations, revealed that only minimal additional impact is achieved by assimilating these observations over land. The magnitude of the additional impact appears to be unrelated to the paucity of observations in the region considered, with similar results over Alaska. Impact in this study was measured only over land (against rawinsondes) and the AMV impact is likely stronger over oceanic regions. Further work investigating the use of these observations, particularly height assignment in the tracking of optically thin features, would help to optimize their impact within rapidly updating NWP systems.

The only negative observation impact seen within this study was for radar-derived VAD wind observations during the nighttime in the warm season. The negative impact was seen chiefly for forecasts initialized during the night (i.e., valid at 1200 UTC), which is commonly recognized as the main time of “biological echoes” being observed by radar (i.e., Wilczak et al. 1995). This highlights the need for improved quality control (QC) of these observations [as expected with dual-polarization radar, e.g., Ryzhkov et al. (2005); Tang et al. (2014)].

VAD wind observations with an additional level of QC are now available operationally; future work will examine the impact of this additional QC step.

In summary, the current observation suite provides an invaluable heterogeneous dataset for the initialization of rapidly updating NWP systems across North America. The dominance of commercial aircraft data, whose availability provided the initial motivation for developing hourly cycling models, persists in present-day systems, and will likely continue for future versions of the RAP and other rapidly updated models including the 3-km HRRR. As regional NWP systems continue to shift toward an ensemble data assimilation and ensemble forecast framework over the next several years (e.g., Whitaker et al. 2008; Tollerud et al. 2013), rapidly updating models are poised to continue to excel by taking advantage of these asymptotic observations.

Acknowledgments. The RAP was developed under significant support from NOAA, the Federal Aviation Administration (FAA), and the Department of Energy. The authors thank Dr. Ming Hu of ESRL/GSD for his help in configuring the experiments presented here; Dr. John M. Brown, also of GSD, for an insightful review of this manuscript; and three anonymous reviewers for further helpful comments.

REFERENCES


