

Evaluating Smartphone Pressure Observations for Mesoscale Analyses and Forecasts

LUKE E. MADAUS AND CLIFFORD F. MASS

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

(Manuscript received 19 July 2016, in final form 7 November 2016)

ABSTRACT

Smartphone pressure observations have the potential to greatly increase surface observation density on convection-resolving scales. Currently available smartphone pressure observations are tested through assimilation in a mesoscale ensemble for a 3-day, convectively active period in the eastern United States. Both raw pressure (altimeter) observations and 1-h pressure (altimeter) tendency observations are considered. The available observation density closely follows population density, but observations are also available in rural areas. The smartphone observations are found to contain significant noise, which can limit their effectiveness. The assimilated smartphone observations contribute to small improvements in 1-h forecasts of surface pressure and 10-m wind, but produce larger errors in 2-m temperature forecasts. Short-term (0–4 h) precipitation forecasts are improved when smartphone pressure and pressure tendency observations are assimilated as compared with an ensemble that assimilates no observations. However, these improvements are limited to broad, mesoscale features with minimal skill provided at convective scales using the current smartphone observation density. A specific mesoscale convective system (MCS) is examined in detail, and smartphone pressure observations captured the expected dynamic structures associated with this feature. Possibilities for further development of smartphone observations are discussed.

1. Background

A significant obstacle to producing skillful, short-term numerical weather forecasts of convection is the lack of high-density observations providing accurate, convective-scale initial conditions (e.g., Stensrud and Fritsch 1994; Roebber et al. 2002; Fowle and Roebber 2003; Dabberdt et al. 2005; Gallus et al. 2005; Snook et al. 2015; Sobash and Stensrud 2015). Several studies have suggested that forecast improvement associated with kilometer-scale numerical prediction is ultimately limited by a lack of high-density observations (Mass et al. 2002; Roebber et al. 2002, 2004). Current observation networks, particularly surface and radiosonde networks, were primarily designed for synoptic-scale forecasting and are ill-suited to constraining short-term, convective-scale forecasts (Sun et al. 2014).

Recent research has found that increased surface observation density can improve mesoscale forecast skill. For example, studies have shown that assimilating observations from regional surface observing networks can improve mesoscale forecasts of severe thunderstorm

development (e.g., Wheatley and Stensrud 2010; Sobash and Stensrud 2015; Xue and Martin 2006; Dong et al. 2011). However, these networks have often been temporary (available only for a particular field campaign) or exist at densities too low to resolve convective-scale features [e.g., the Oklahoma Mesonet with station separations of 30–70 km; McPherson et al. 2007; Sobash and Stensrud (2015)]. Alternatively, Madaus et al. (2014) describes, using a regional mesoscale ensemble, the impact of crowdsourced observations of surface pressure from networks such as the Weather Underground (<http://www.wunderground.com>) “backyard” weather stations. These pressure observations improved short-term (0–6 h) forecasts of mesoscale features like frontal passages and convergence zones in their experiments.

An increasing number of smartphones contain barometers, whose pressure observations can be used for meteorological data assimilation. Mass and Madaus (2014) describe these smartphone pressure observations (hereafter abbreviated SPOs) and provide a case study of a convective event in eastern Washington State in which assimilating SPOs improves the forecast placement of convective storms in an area of relatively sparse conventional observations. Since those experiments, the

Corresponding author e-mail: Luke E. Madaus, lmadaus@atmos.washington.edu

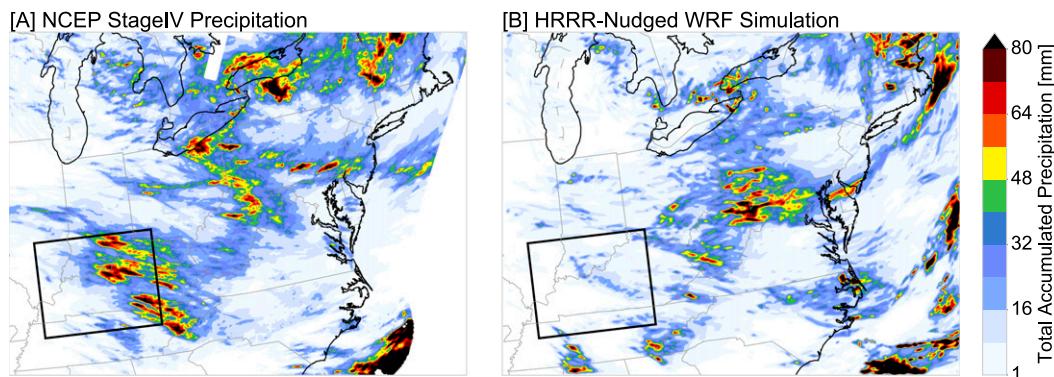


FIG. 1. Total accumulated precipitation (mm) from 1200 UTC 26 Jul 2014 through 1200 UTC 29 Jul 2014 as derived from (a) the NCEP stage IV hourly precipitation analyses and (b) a WRF simulation over the same period. The lower Ohio River valley is outlined in black.

number of available smartphone observations has grown considerably, and further work to provide quality control and characterize the errors of SPOs is ongoing. As SPOs could attain greater observational density than any extant surface observing network, this work investigates how currently available SPOs may contribute to improved numerical forecasts in a case study of a convectively active period.

2. Methods

a. Description of the study period

For this case study, we examine a 3-day period from 1200 UTC 26 July 2014 through 1200 UTC 29 July 2014 in the east-central region of the continental United States. Figure 1a shows the extent of the forecast domain, with total accumulated precipitation during the study period as estimated from the National Centers for Environment Prediction stage IV hourly precipitation analyses (Lin and Mitchell 2005). Figure 1b shows the accumulated precipitation over the same period as computed from a Weather Research and Forecasting Model (WRF; Skamarock et al. 2008) simulation nudged to hourly High Resolution Rapid Refresh (HRRR) analyses with 1-h HRRR forecasts as boundary conditions over this period. There is remarkably little agreement between the observed and simulated precipitation fields, indicating this was a challenging forecast period for operational mesoscale models.

The 500-hPa heights and winds as well as surface analyses from 0000 UTC 27 July 2014 and 0000 UTC 28 July 2014 are shown in Fig. 2. Initially, the synoptic-scale forcing is relatively weak, but there are several outbreaks of airmass-type convection, which is typical for this midsummer period. During the latter half of the

study period, a short-wave trough and associated surface cyclone bring organized forcing for several convective events (Fig. 2, right). One example is the initiation and evolution of a mesoscale convective system in the Ohio River valley, beginning at 0000 UTC 27 July. Though this MCS is prominent during the study period, it was completely absent from the HRRR-nudged simulation (Fig. 1, boxed region). While our analysis mostly evaluates forecast performance over the entire study region, this MCS event is examined in greater detail to examine how SPOs may contribute to improving highly sensitive convective forecasts.

b. Smartphone observations

For this case study, we collect SPOs from two smartphone apps: OpenSignal and PressureNet. Smartphone users download one of these apps and give it permission to run in the background and transmit pressure observations. The frequency of reporting varies, but most phones produce pressure observations at least once per hour. Given the small number of phones that transmit at higher frequencies, we limit ourselves to hourly observations in this study.

1) OBSERVATION QUALITY CONTROL

Quality control checks are applied to the observations prior to data assimilation. First, observations are limited to those within 15 min of each assimilation time. If there are multiple observations from the same smartphone within this window, the observation closest to the assimilation time is retained and others are discarded as duplicates. To remove potential mismatch between observation elevation and model terrain, the raw SPOs are converted to altimeter settings. For the remainder of this study, the terms smartphone pressure observation

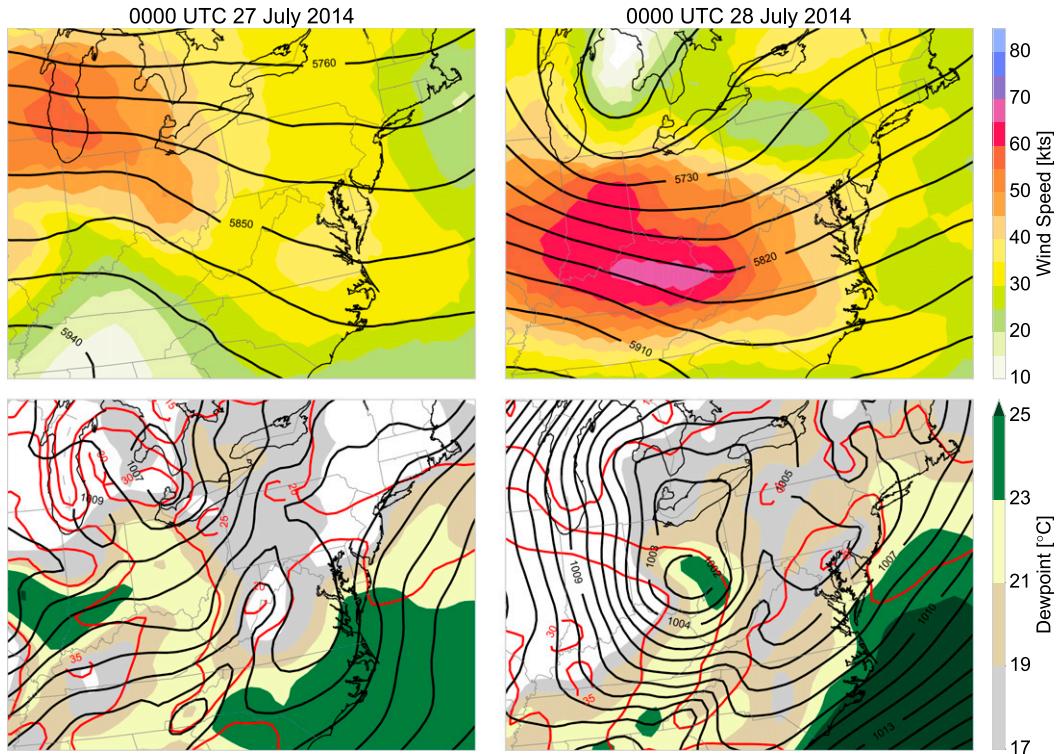


FIG. 2. NARR synoptic analyses from (left) 0000 UTC 27 Jul and (right) 0000 UTC 28 Jul 2014. Shown are (top) the 500-hPa geopotential height (m; contoured) and wind speed [in knots (kt), where 1 kt = 0.51 m s⁻¹; filled] and (bottom) the mean sea level pressure (hPa; black contours), surface temperature (°C; red contours), and surface dewpoint temperature (°C; filled).

and smartphone altimeter observation are used interchangeably. We compute the altimeter setting using the following formula:

$$P_{\text{alt}} = (p - 0.3) \left[1 + \frac{k_1 h}{(p - 0.3)^{k_2}} \right]^{1/k_2}, \quad (1)$$

where $k_1 = 8.4228807 \times 10^{-5}$ and $k_2 = 0.190284$ are constants derived from the expected vertical rate of pressure change in the *U.S. Standard Atmosphere, 1976* (Dubin et al. 1976), h is the phone elevation above sea level in meters, p is the raw smartphone pressure in hectopascals, and P_{alt} is the altimeter setting in hectopascals.

In general, the elevation retrieval from the smartphones is unreliable or not possible because of inaccuracies in the global positioning system (GPS) based elevation estimation method. As such, we compare the reported GPS location to Shuttle Radar Topography Mission (SRTM) 1-arc-s elevation measurements and assign each smartphone observation the SRTM elevation interpolated to the phone location, plus 1 m (as it is unlikely the phone is exactly on the ground). We

recognize that the GPS location may also contain errors, which could increase the uncertainty in the elevation estimate. Furthermore, there is no guarantee that the smartphone is close to ground level, particularly in urban areas with multistory buildings. Here, we do not attempt to estimate or correct for this possible error, but envision developing strategies for doing so in the future. This elevation uncertainty could be included in the observation error variance used for data assimilation, as near-sea level atmospheric pressure decreases by about 1 hPa for every 8 m of elevation increase. However, here all smartphone observations are assigned an observation error variance of 1 hPa², which is similar to the surface pressure observation errors used in operational data assimilation systems (e.g., Burton 2013; Hu et al. 2013).

Outlier observations are identified as follows. SPOs with an altimeter setting outside of an 890–1100-hPa range are discarded. This typically removes less than 1% of the observations. The SPO altimeter settings throughout the study domain are then sorted by elevation and observation value and an exponential fit is applied to these observations. This statistical consistency check discards as outliers any observations that lay more than three standard deviations from this fit.

TABLE 1. Average fraction of all available smartphone observations removed by the quality control (QC) checks described in section 2a during each assimilation cycle. The numbers in boldface highlight the percentage of observations remaining after the QC check.

QC check	Percent rejected
Duplicate obs	9
Valid range	<1
Statistical consistency	2
Spatial consistency	52
Remaining observations	36
1-h altimeter tendencies	
Valid range	<1
Spatial consistency	60
Remaining observations	40

Finally, a gross spatial consistency check is performed to remove observations that are in substantial disagreement with their surroundings. Altimeter observations from the Meteorological Atmospheric Data Ingest System (MADIS), which include airport observations and other “mesonet” surface-observing networks, are compared to the smartphone altimeter observations. Specifically, the nearest eight MADIS altimeter observations to each smartphone observation are distance-weighted interpolated to the smartphone observation to produce a “synthetic” observation to compare with the smartphone altimeter. The standard deviation of the altimeter settings from those nearest eight stations is also computed. If the difference between the smartphone altimeter observation and the interpolated synthetic observation is greater than twice this standard deviation, then the smartphone altimeter observation is rejected as an outlier. This tolerance is specifically generous to avoid rejecting observations of localized mesoscale features or observations in complex terrain. However, this must be balanced with the desire to reduce noise in these observations (section 3b). Even with the tolerances used here, this spatial consistency check removes over half of the candidate smartphone observations. The average fraction of observations removed by each of these quality control checks is summarized in Table 1. The set of smartphone altimeter observations remaining after these checks is used for data assimilation, with no further attempt to correct for bias or other errors.

These altimeter observations are then compared with altimeter observations obtained 1 h prior. Any altimeter observations from the same device at both the current time and 1 h prior that are separated in distance by less than 13 m are used for computing the altimeter tendency at that location. The 13-m threshold represents the approximate distance error associated with truncation in the latitude and longitude values received from the

smartphone. While it is possible for vertical motions of smartphones at the same horizontal location to introduce nonmeteorological signal into these tendency calculations, the randomness of this vertical motion (some phones going up while others going down) and the limited number of buildings greater than two or three stories outside of dense urban cores limits this potential error to within the prescribed $1 \text{ hPa}^2 \text{ h}^{-2}$ observation error variance. The 1-h change in altimeter setting is computed, using the observation times closest to the top of the hour if the observation reports frequently. The 1-h altimeter tendencies of greater than 7 hPa are discarded as unrealistic. A gross spatial consistency check is applied, described above for the smartphone altimeter observations, but instead comparing the 1-h smartphone altimeter tendencies to surrounding MADIS 1-h altimeter tendencies. This check rejects the majority of 1-h tendency observations (60% on average; Table 1). The remaining tendency observations are used for data assimilation.

2) SUMMARY OF AVAILABLE SMARTPHONE OBSERVATIONS

Figure 3 shows the average number of SPOs within $30 \text{ km} \times 30 \text{ km}$ bins that are available every hour during the case study period and that pass the quality control checks described above. Because only two small apps that require user permission to transmit pressure observations are considered here, the number of available observations acquired is from only a very small percentage of the total number of smartphones able to collect pressure observations. The availability of hourly pressure observations (Fig. 3a) and 1-h pressure tendency observations (Fig. 3b) largely follow the population density, which is expected of crowdsourced observations. Requiring smartphones to be nearly stationary for 1 h to compute pressure tendency and employing the spatial consistency check limits the number of smartphone pressure tendencies to about 20% of the number of pressure observations, on average. Figures 3c and 3d show the standard deviations of the number of observations in each $30 \text{ km} \times 30 \text{ km}$ bin over the duration of the case study for pressure observations and pressure tendency observations, respectively. Most land areas of the domain have nonzero standard deviations, indicating that transient SPOs are able to sample the domain even outside the urban centers. Additional aspects of the SPO network are discussed in section 3a.

3) OTHER OBSERVATIONS

To verify precipitation forecasts, we use the National Centers for Environment Prediction (NCEP) stage IV hourly precipitation analyses (Lin and Mitchell 2005). In

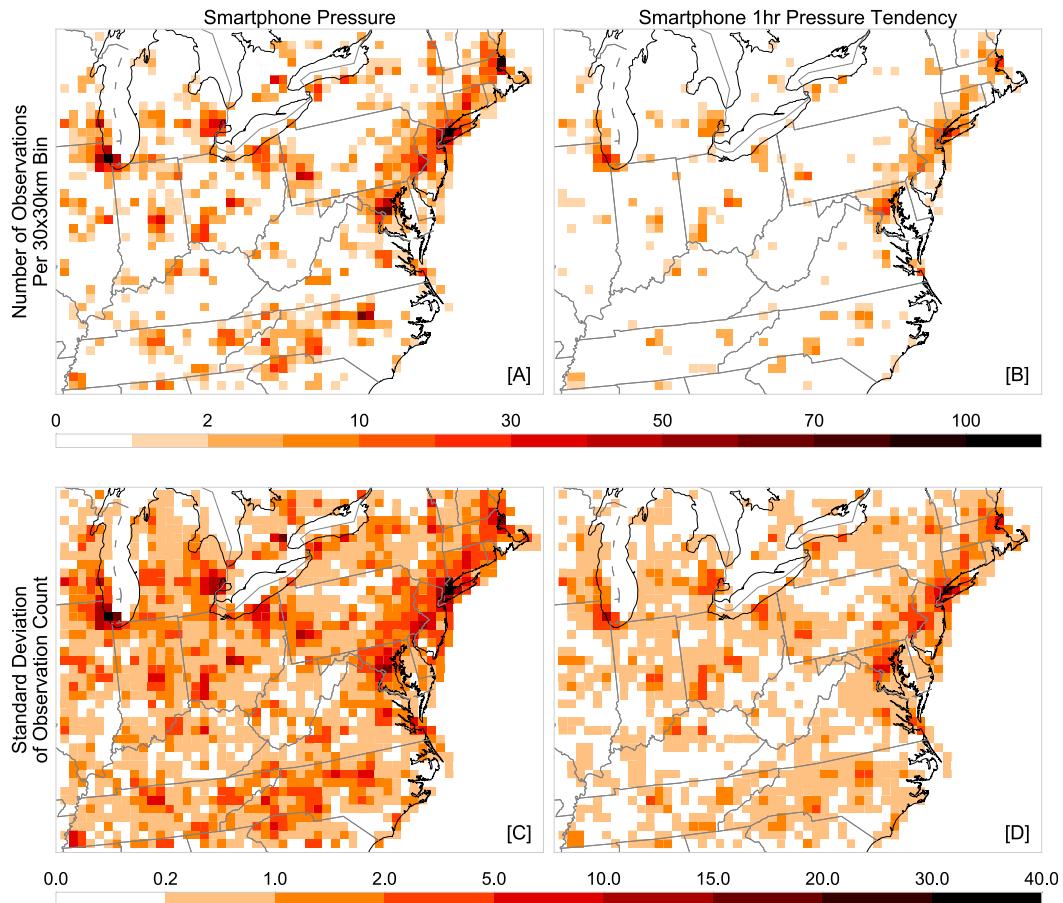


FIG. 3. (a) Average number of SPOs and (b) 1-h smartphone pressure tendency observations available at each assimilation time during the case study period. Observation counts are binned into $30\text{ km} \times 30\text{ km}$ bins. The standard deviations of the observation counts during the case study period are shown for (c) SPOs and (d) 1-h smartphone pressure tendency observations.

addition, we use archived Next Generation Weather Radar (NEXRAD) level II radar reflectivity data to composite radar reflectivity observations from all radars within the study domain at 1-km elevation every hour during the study period using the Python-ARM Radar Toolkit (PyART; <https://github.com/ARM-DOE/pyart>). These composites facilitate evaluation of simulated radar reflectivity fields. We also use Automated Surface Observing System (ASOS) aviation routine weather reports (METAR) observations for verification of skill and to compare the analysis quality from SPOs to standard METAR altimeter observations.

c. Forecast model

The forecast model in these simulations is the Advanced Research core of the WRF (ARW), version 3.6.1 (Skamarock et al. 2008). The model grid structure and parameterizations are set to match the configuration of the operational National Centers for Environmental

Prediction (NCEP) HRRR forecasting system as of March 2015 (Benjamin et al. 2015). The horizontal grid spacing is 3 km with 51 vertical levels. Parameterizations include Thompson cloud microphysics (Thompson et al. 2008), a nine-level RUC/Smirnova land surface model with Moderate Resolution Imaging Spectrometer (MODIS) derived land surface characteristics including fractional coverage (Smirnova et al. 2015), and the Mellor–Yamada–Nakanishi–Niino (MYNN)–Olson boundary layer scheme (Nakanishi and Niino 2009; Olson and Brown 2012). No convective parameterization is used. We apply gravity wave damping in the upper 5 km of the model domain.

Initial and boundary conditions are derived from HRRR forecasts archived at the National Center for Atmospheric Research (NCAR). Though many aspects of our simulation design mimic the forecasts from the operational HRRR, in our cycling data assimilation experiments we do not assimilate the full complement of

observations that are used in the operational HRRR (e.g., GOES cloud observations, lightning, and radar data) to specifically highlight the contributions of SPOs. As such, we do not anticipate that the performance of our forecasts will match that of the operational HRRR. However, an ensemble system based on the HRRR physics offers a reasonable convection-allowing platform for testing various observational datasets.

For these experiments, a 50-member ensemble is generated at 0000 UTC 26 July 2014 with initial conditions derived from the HRRR analysis at that time. Figure 1a shows the geographical extent of the simulation domain. To generate ensemble diversity, we employ the stochastic kinetic energy backscatter scheme (SKEBS) available in WRF (Berner et al. 2011), which provides good background error representation in mesoscale models (e.g., Ha et al. 2015). We integrate the ensemble members using 1-h HRRR forecasts as boundary conditions, which are also perturbed using SKEBS. Here, the maximum amplitude of both streamfunction (ψ) and temperature (t) perturbations in SKEBS is set to 1×10^{-4} and the decorrelation time scale to 1 h. After allowing perturbations to spin up, we found these values produced an ensemble with 1-h forecast variance in surface temperature, wind, and pressure fields that approximated the mean squared error in the ensemble mean measured against METAR surface observations (not shown). The ensemble members integrate without data assimilation for 12 h to allow SKEBS to spin up ensemble perturbations. Ensemble cycling with data assimilation begins at 1200 UTC 26 July 2014. The ensemble is cycled hourly, but 6-h forecasts are made after every ensemble cycle, with boundary conditions provided by the 0–6-h HRRR forecast initiated at the same time. In these forecasts, only the surface fields (2-m temperature, 10-m u and v winds, surface pressure, and 2-m specific humidity), the simulated reflectivity, and the accumulated rainfall at the surface are retained for analysis.

d. Data assimilation

For data assimilation, we use the ensemble adjustment variation of the ensemble Kalman filter (EAKF; Anderson 2001) as implemented in the Data Assimilation Research Testbed (DART; Anderson et al. 2009). To maintain sufficient ensemble spread, both spatially and temporally adaptive covariance inflation (Anderson 2009) and sampling error correction for a finite ensemble size (Anderson 2012) are applied. In addition, we apply horizontal spatial localization to all assimilated observations, with covariance weights defined using a Gaspari–Cohn function (Gaspari and Cohn 2006) with a half-width of 500 km. This radius is similar to or slightly

smaller than radii used in other recent studies that assimilate mesoscale surface pressure observations (e.g., Wheatley and Stensrud 2010; Madaus et al. 2014; Sobash and Stensrud 2015) to reflect the higher spatial density of the smartphone observations. No vertical localization is used.

In practice, the smartphone pressure observations are assimilated as altimeter setting observations (and, likewise, pressure tendency as altimeter tendency) as noted above. Observation error variances are set to 1 hPa^2 for the altimeter setting and $1 \text{ hPa}^2 \text{ h}^{-2}$ for the altimeter tendency, following Madaus et al. (2014). To reduce the negative impact of outlier observations, prior to the assimilation of each observation, the innovation (the difference between the observation value and the ensemble mean estimate of the observation) is compared with the standard deviation in the ensemble estimate of the observation. Any observations where the innovation magnitude is greater than 3 times the spread in the ensemble observation estimate are rejected as outliers and are not assimilated.

e. Experiments

We perform two kinds of experiments to analyze the impact of smartphone pressure and pressure tendency observations on the analyses and forecasts. In both sets of experiments, the reference is a control experiment (control) that assimilates no observations and is entirely driven by the HRRR boundary conditions. We perform a set of three no-cycling experiments (e.g., Madaus et al. 2014) to specifically focus on contributions of SPOs to analysis quality. For these experiments, the same ensemble state at each hour of the control experiment is used as the background state for assimilation. We then assimilate either METAR altimeter (METAR_only_nocy), smartphone altimeter (phone_only_nocy), or 1-h smartphone altimeter tendency (phone_tend_only_nocy) observations to produce an analysis. In these no-cycling experiments, the ensemble is not cycled between assimilation times: at each assimilation time, the background state from the control experiment is used. Analyses of surface fields are then verified against METAR observations. This allows the analysis quality from each observation type to be compared given the same set of background (prior) states.

Second, we perform two experiments to evaluate the impacts of the smartphone observing network on mesoscale forecasts using fully cycled ensembles as described above. We assimilate each observation type individually in these experiments to clearly identify the strengths and weaknesses of each type of smartphone observation. The phone_only experiment assimilates only smartphone altimeter setting observations, while

the phone_tend_only experiment only assimilates 1-h smartphone altimeter tendency observations.

f. Verification methods

A variety of metrics are considered for evaluating the performance of the ensemble forecasts. For verification of ensemble forecasts against independent observations, the root-mean-square error is computed as

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^n (f_i - o_i)^2 \right]^{1/2}, \tag{2}$$

where n is the total number of grid points, f_i is the forecast value at each verification observation, and o_i is the value of the verifying observation.

For computing probabilistic forecast skill from the ensembles, the Brier score (BS) is employed (Brier 1950), which is similar to RMSE but for probabilistic forecasts:

$$BS = \frac{1}{n} \sum_{i=1}^n (P_i - o_i)^2, \tag{3}$$

where n is the total number of locations (or times) where forecasts are produced, P_i is the forecast probability of occurrence of a specific event at each location, and o_i is a value of 1 or 0, depending on whether an event occurred at that location or not, respectively. In our analysis, P_i is determined by the fraction of ensemble members that forecast the event to occur. We specifically use the Brier score within the context of the Brier skill score (BSS) to compare the skill of one forecast against another:

$$BSS = 1 - \frac{BS_{fcst}}{BS_{ref}}, \tag{4}$$

where BS_{fcst} is the Brier score from the test forecast and BS_{ref} is the Brier score from some reference forecast. A negative BSS indicates the test forecast has less skill than the reference forecast (here, forecasts from the control experiment), while a positive BSS indicates greater skill than the reference. A perfect forecast would have a BSS of 1.

To evaluate the skill of these simulations across different spatial scales, we employ the fractions skill score (FSS; Roberts and Lean 2008), which extends the Brier skill score to evaluations of fractional coverage across different length scales. The FSS is computed as

$$FSS = 1 - \frac{FBS}{FBS_{worst}}, \tag{5}$$

where FBS is the fractions Brier score:

$$FBS = \frac{1}{N} \sum_{j=1}^N (O_j - M_j)^2. \tag{6}$$

At each grid point j , the fractional coverage of a binary metric (e.g., precipitation greater than 1 mm) is evaluated within a specified spatial neighborhood in both the model forecast M_j and the gridded observation data O_j . The squared difference in these coverages is averaged over all N points in the domain. Where both the model and the observations have the same fractional coverage within the neighborhood, $FBS = 0$. If there is mismatch, $FBS > 0$. To complete the FSS, the FBS is compared with FBS_{worst} , which is the expected score when there is no overlap in the fractional coverage between model and observations:

$$FBS_{worst} = \frac{1}{N} \left(\sum_{j=1}^N O_j^2 + \sum_{j=1}^N M_j^2 \right). \tag{7}$$

The FSS varies from 0 for a forecast with no skill to 1 for a perfect forecast. FSS may be evaluated using different spatial neighborhoods over which fractional coverage surrounding each grid point is computed. This indicates how the skill in the forecast varies as a function of spatial scale. For instance, the forecast may correctly indicate the location and extent of a broad line of convective storms but have errors in the exact placement of individual convective elements within that line. In this case, the FSS would be higher at larger spatial scales (the scale of the convective line as a whole), but lower at smaller spatial scales (the scale of individual convective storms).

Similarly, for some forecast evaluations we consider neighborhood probability forecasts instead of raw gridpoint-based probabilities. Specifically, we evaluate 30-km neighborhood probabilities as the fraction of ensemble members that meet some criteria at each grid point or within 30 km of that grid point. Neighborhood probabilities reduce penalties for small errors in the exact location of predicted features and are commonly used for precipitation evaluation (e.g., Schwartz et al. 2010).

3. Results

a. Assimilation performance

Figure 4 shows the number of smartphone altimeter and altimeter tendency observations available and assimilated in the fully cycled experiments (phone_only and phone_tend_only). Recall that the number of available observations for this study represents only a small fraction of the total number of smartphones that

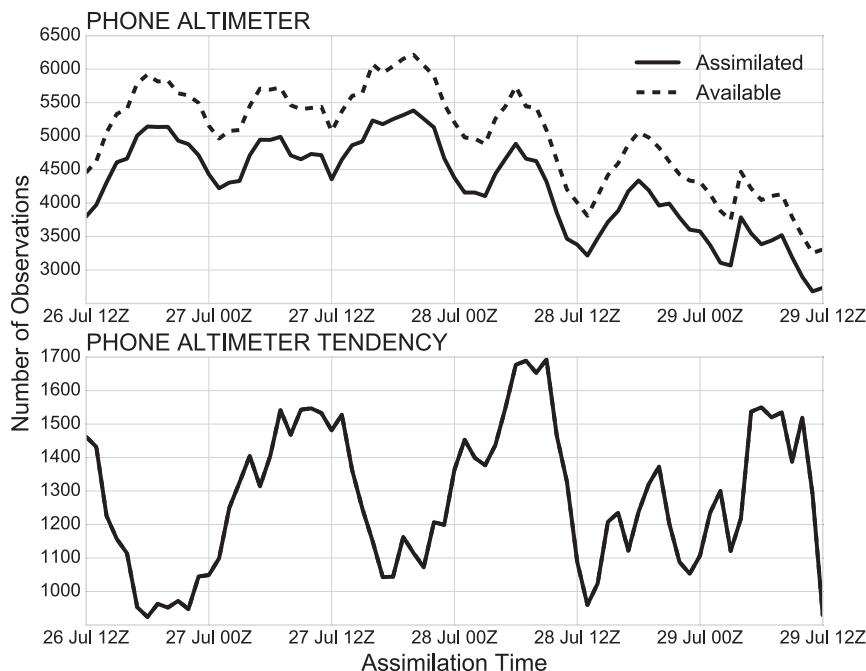


FIG. 4. Number of observations available (dashed) and assimilated (solid) as a function of analysis time during the cycling experiments. Shown are the (top) phone altimeter observations (from `phone_only`) and (bottom) phone altimeter tendencies (from `phone_tend_only`). Here, the available observations are those that have passed all quality control checks described in [section 2a](#).

are capable of observing surface pressure. Here, the available observations are those that have passed all the quality control checks described in [section 2a](#). The rejection of available observations occurs when the observed value is more than three standard deviations from the ensemble mean estimate of the observation, as described above. At any given assimilation time, an average of 15% of the smartphone altimeter observations are rejected as a result of this constraint. However, none of the altimeter tendency observations are rejected. The lack of rejected tendency observations highlights one benefit of this observation type, as tendencies are insensitive to observation biases that are constant in time ([Madaus et al. 2014](#)).

The number of available SPOs varies in time. On all days there is a peak in the number of observations during the local daytime (between 1500 and 2100 UTC) with secondary peaks during evening hours (between 0300 and 0900 UTC). The diurnal cycle is more pronounced in the smartphone altimeter tendency observations, where distinct peaks are observed during the local overnight hours (between 0300 and 1200 UTC). These peaks are explained by the smartphone altimeter tendency algorithm, which requires the phone to be stationary for the duration of the tendency (here, 1 h) to

produce an observation. This becomes more likely when the phone is not in motion, hence the greater number of observations during the overnight hours.

b. Observation consistency

As noted in [section 2b](#), the smartphone pressures were subject to quality control to remove outlier observations. As noted above, some 15% of the remaining observations are rejected as outliers by the assimilation system. To evaluate the consistency between the observations that are assimilated (i.e., how well nearby observations agree on the sign of the assimilation increment), we compare the analysis error magnitude (the analysis ensemble mean at the observation location minus the observation value) to the background (prior) error magnitude (the background ensemble mean at the observation location minus the observation value) for all assimilated smartphone observations in the no-cycling experiments. In an ideal case, it is expected that the analysis error is reduced from the background error for the vast majority of all observations assimilated. In other words, the assimilation process should, on average, bring the ensemble mean closer to the observations to produce an analysis with lower error. [Figure 5](#) breaks down the difference in analysis error and background error

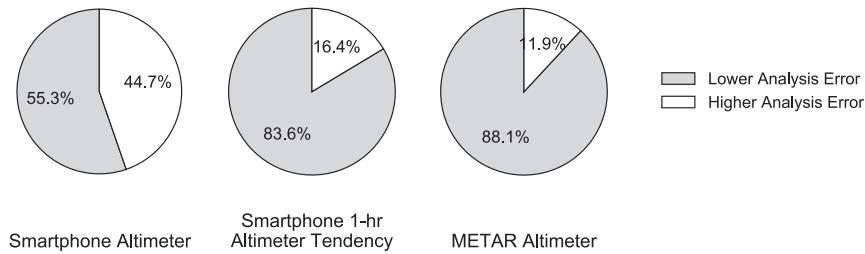


FIG. 5. Fraction of assimilated smartphone observations in the no-cycling experiments where the analysis ensemble mean error magnitude is less than the background ensemble mean error magnitude (gray) and where it is greater (white). Shown are (left) the smartphone altimeter observations (`phone_only_nocy`), (center) the smartphone 1-h altimeter tendency observations (`phone_tend_only_nocy`), and (right) the METAR altimeter observations (`METAR_only_nocy`).

magnitudes for all observations assimilated in `phone_only_nocy`, `phone_tend_only_nocy`, and `METAR_only_nocy`.

In the `phone_only_nocy` experiment (Fig. 5, left) the analysis error is greater than the background error for almost 45% of the assimilated observations. The Kalman filter guarantees an analysis that is between the observation value and the background estimate when assimilating each observation individually. Thus, an increase in analysis error indicates that covariances from neighboring observations are introducing local increments in conflict with what is suggested by that observation itself. It is possible that the ensemble covariances are suboptimal or the localization radius used is inappropriate. However, this seems unlikely as the localization half-width radius used here (500 km) agrees well with computed correlation length scales for surface pressure from this ensemble (Fig. 7). In contrast, for the METAR altimeter observations (Fig. 5, right), only about 12% of the observations have an analysis error worse than the background error. With such a large percentage of SPOs where the analysis is worse, it is likely there are significant inconsistencies within the SPO network, which may limit the ability of these observations to inform the state in a useful way.

The 1-h altimeter tendency observations have better agreement (Fig. 5, center), with almost 84% of the observations having an analysis pressure tendency error that is smaller in magnitude than the background pressure tendency error. While there still appear to be some inconsistent observations, the analysis error is smaller for most altimeter tendency observations, suggesting that observations based on relative changes in the pressure observations are more consistent than those using the potentially biased raw pressure values. This agrees with the findings of Madaus et al. (2014), who showed that altimeter tendency observations could

alleviate the bias problem in crowdsourced pressure observations.

c. Surface analysis quality

To evaluate the quality of the surface analyses from assimilating SPOs, we examine the ensemble mean RMSE of the analyses evaluated against METAR observations in the no-cycling experiments. Recall that in these experiments at each assimilation time the same background ensemble state is used (from the control experiment) as the prior state, allowing a fair comparison of assimilation effectiveness. Figure 6 shows the RMSEs in these analyses, both as a function of assimilation time (left panels) and as differences in error between the control and `phone_only_nocy`, control and `phone_tend_only_nocy`, and control and `METAR_only_nocy` experiments (right panels). We note that in the `METAR_only_nocy` experiment, the METAR altimeter observations assimilated are the same as those used for verification, so this is not an independent verification of analysis quality for the `METAR_only_nocy` experiment.

Surface pressure analyses (METAR altimeter; Fig. 6, top) are improved over the control for the majority of the time in both `phone_only_nocy` and `phone_tend_only_nocy`. This affirms that, on average, the smartphone pressure observations are contributing positively to constraining the surface pressure field. As expected, the analysis error at METAR altimeter locations when those METAR altimeter observations are assimilated is significantly reduced from the control. Virtually no improvement is seen in the 10-m wind analyses (Fig. 6, middle) or 2-m temperature analyses (Fig. 6, bottom) when the smartphone altimeter observations or 1-h smartphone altimeter tendency observations are assimilated. However, improvements are noted in the analyses of these fields when the METAR altimeter observations are assimilated.

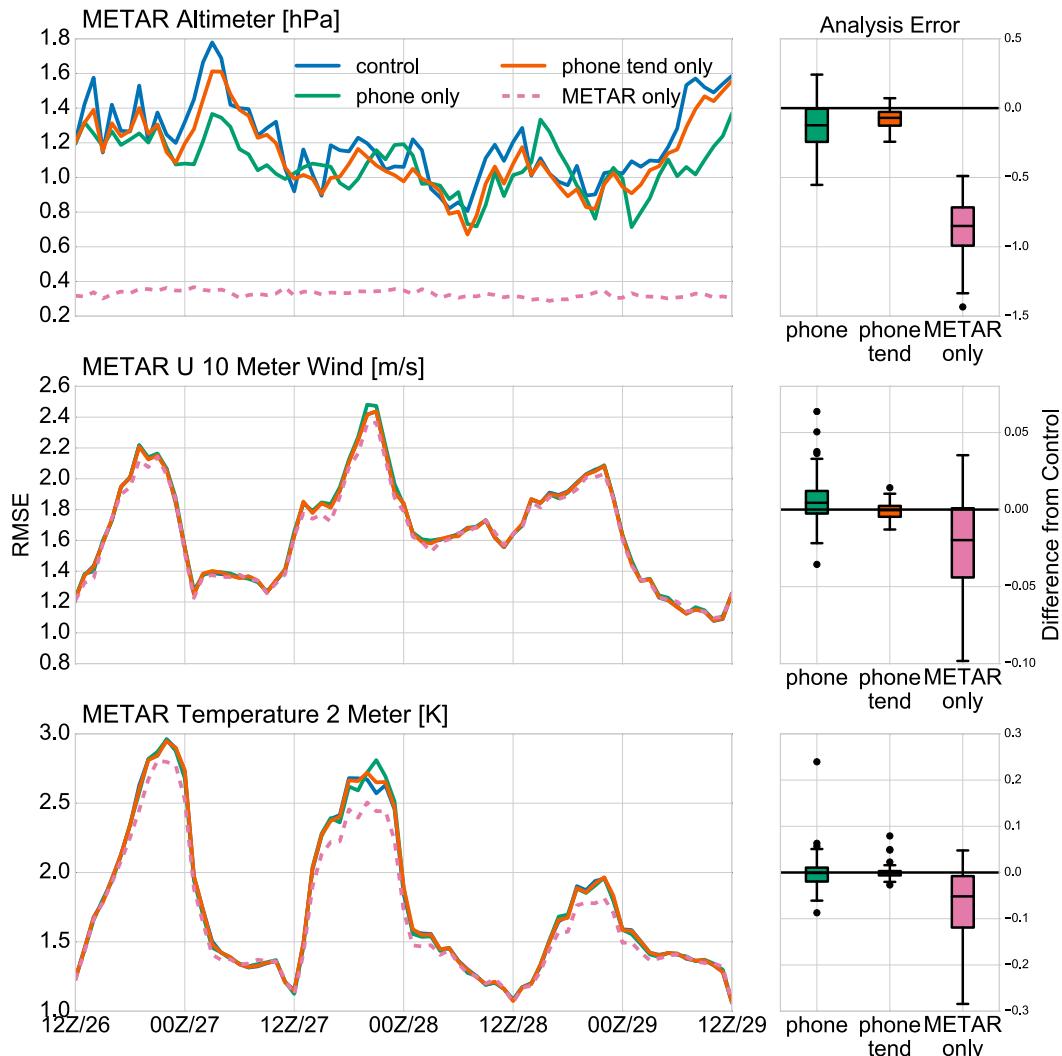


FIG. 6. RMSEs of the (top) altimeter setting, (middle) U10, and (bottom) T2 analyses in the no-cycling experiments as verified against ASOS/METAR observations. Statistics are given for control (blue), phone_only_nocy (green), phone_tend_only_nocy (orange), and METAR_only_nocy (pink). Shown are (left) RMSE time series and (right) box-and-whisker plots of the time-by-time differences between the control forecast RMSE and each of the other experiments. Note that for the METAR altimeter results in the top panel, the METAR_only_nocy experiment (pink; dashed) assimilates the METAR altimeter observations and therefore, for this experiment, the observations are not independent.

We explore two factors contributing to the lack of improvement in 10-m wind and 2-m temperature analyses when SPOs are assimilated: correlation length scales and noise in the observations. Figure 7 shows estimates of the average error correlation magnitude computed from the ensemble as a function of distance between the surface pressure and other variables for these simulations. We note that these length scales may vary depending on the local variability in the surface characteristics or topography; these represent domain-wide averages. In particular, correlation length scales may be smaller in areas of complex terrain, where increased observation density

may be critically important. Surface pressure correlated to itself decreases in magnitude to 0.2 around 500 km (Fig. 7, left). This indicates the localization radius used here is well calibrated for updating surface pressure. However, the correlation magnitudes between surface pressure and 2-m temperature (T2) or 10-m wind (U10) are much smaller in magnitude, and asymptote to a “noise” magnitude on much shorter length scales of 100–200 km. Thus, METAR altimeter observations—being collocated with the verifying 10-m wind and 2-m temperature observations—are more likely to contribute to improving the local temperature analysis.

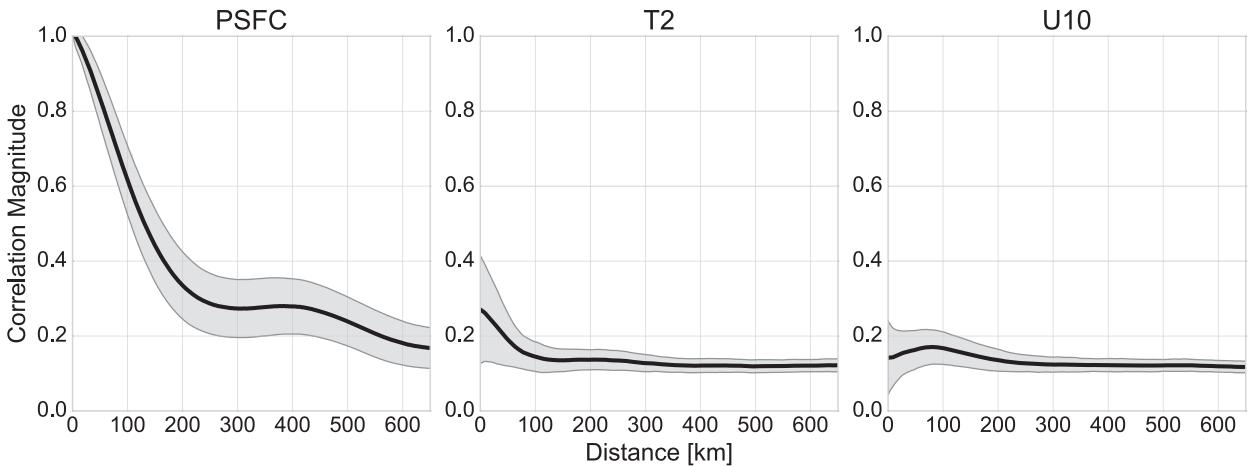


FIG. 7. Average correlation coefficient magnitude as a function of distance between (left) surface pressure (PSFC), (center) T2, and (right) U10. These plots are generated from a subsample of 10 grid points randomly drawn from background (prior) ensemble pressure fields at each assimilation time (total sample size of 730) in the control experiment. Gray shading indicates one standard deviation of correlation magnitude from the samples compared.

However, since many SPOs are clustered in urban areas where METAR observations are nearby (Fig. 3), correlation length scales alone likely do not explain why SPOs offer no improvement to 10-m wind and 2-m temperature analyses. We therefore investigate how observation noise may affect the quality of the analyses. With noisy pressure observations and relatively short correlation length scales for 2-m temperature and 10-m wind (Fig. 7), we expect the error in the temperature and wind analyses to decrease with locally increased observation density. This occurs as the analysis state is less influenced by any individual observation and opposing errors from multiple nearby observations may cancel out. Figure 8 examines the analysis error magnitude for

all (unassimilated) METAR observations as a function of the surrounding smartphone observation density in the phone_only experiment. At each assimilation time, the number of SPOs within 50 km of each METAR observation is computed, as well as the analysis error verified against that METAR observation. These values are used to construct the 2D histograms seen in Fig. 8. For METAR observations with zero SPOs within 50 km (left columns), the mean analysis errors are relatively low for all variables. However, when there are even a few smartphone observations within 50 km of a METAR observation, the range of analysis error broadens and the mean error increases. As the number of smartphone observations within a 50-km radius

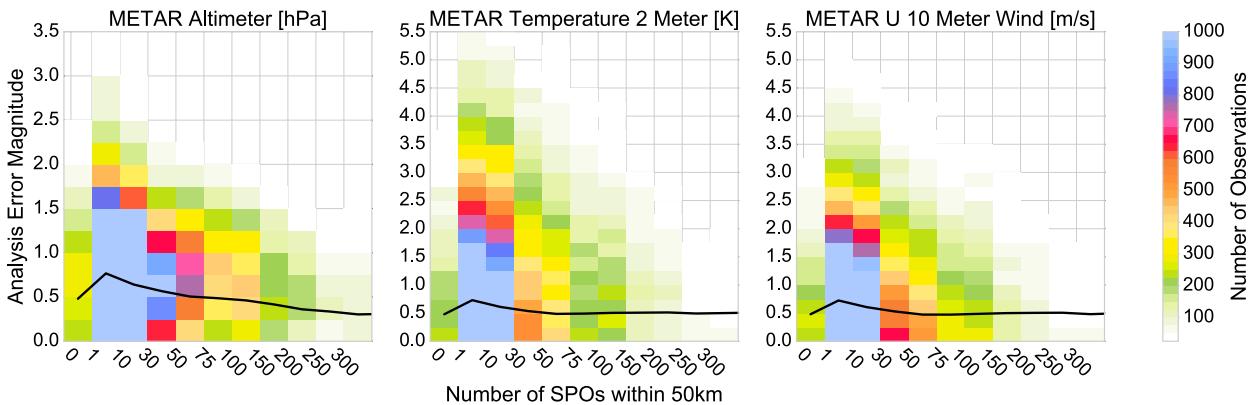


FIG. 8. Two-dimensional histograms of unassimilated (left) METAR altimeter, (center) T2, and (right) U10 observations from all assimilation times in the phone_only experiment binned by the analysis error magnitude (y axis) and the number of smartphone altimeter observations within 50 km of the evaluated observation (x axis). Bins are inclusive along the left edge, except for the first bin, which only includes METAR observations with zero smartphone observations within 50 km. The mean analysis error magnitude in each bin is denoted by the black line. Note the quasi-logarithmic scaling along the x axis.

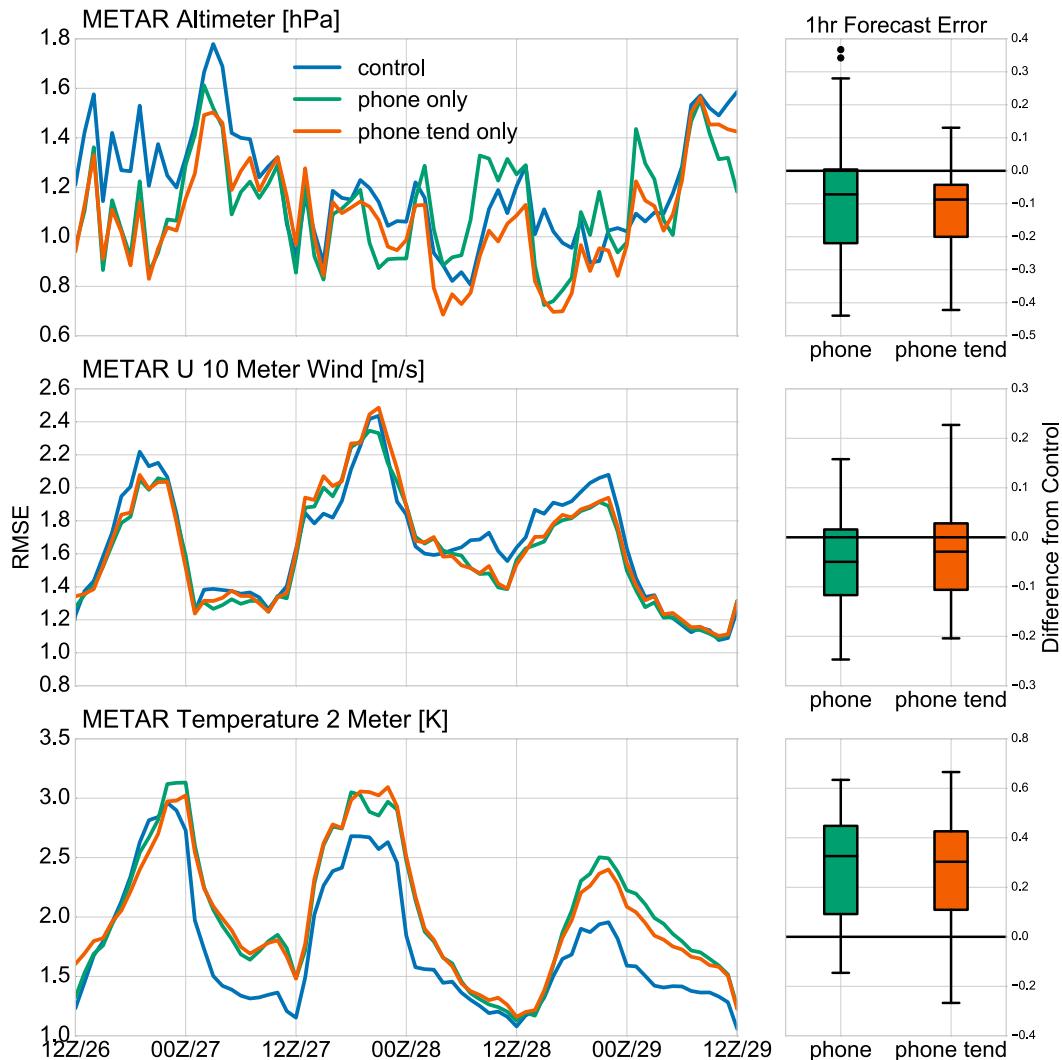


FIG. 9. As in Fig. 6, but here evaluating 1-h forecasts of surface fields against unassimilated METAR observations in the cycled `phone_only` and `phone_tend_only` experiments.

increases, the mean analysis error decreases along with the error variability. This finding is consistent with the expectation that, without better observation quality control, cancellation of errors from multiple smartphone observations is required to isolate a net weather signal.

d. Forecast errors in surface fields

We now turn to the fully cycled experiments to explore how assimilating SPOs affects the forecast. Figure 9 evaluates the short-term (1 h) forecasts of selected surface fields against unassimilated METAR observations in `phone_only` and `phone_tend_only`. Small improvements are noted in the 1-h 10-m wind and altimeter forecasts, where errors are less than the control over 60% of the time. However, the sample size of forecasts here is not enough for this improvement to be

considered statistically significant. In contrast, 1-h 2-m temperature forecasts are actually worsened when either type of SPO is assimilated. For all these variables, there is no significant difference between the improvements due to the assimilation of smartphone altimeter observations versus smartphone altimeter tendencies, though the errors in the altimeter forecasts are often lower in `phone_tend_only` than `phone_only`. The latter may reflect the insensitivity of altimeter tendency observations to observation bias, producing a more reliable estimate of the altimeter field.

e. Precipitation forecast skill

We evaluate precipitation forecast skill by comparing hourly precipitation accumulation forecasts to hourly NCEP stage IV 1-h precipitation accumulation analyses.

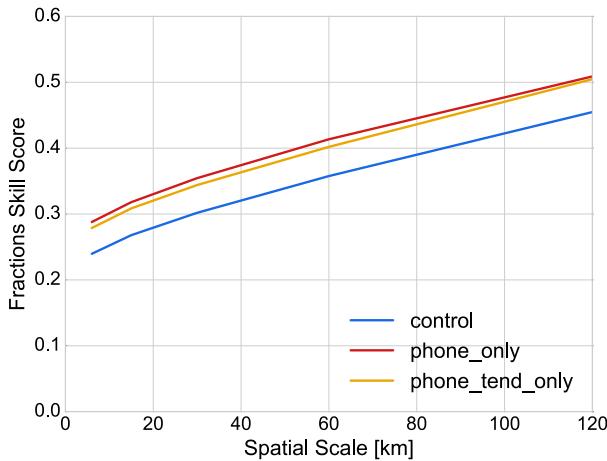


FIG. 10. FSSs for 0–1-h ensemble mean forecasts of precipitation ≥ 1 mm in the control (blue), phone_only (red), and phone_tend_only (gold) experiments.

We examine the FSSs for ensemble mean 1-h forecasts of precipitation ≥ 1 mm across a variety of spatial scales in Fig. 10. Both the phone_only and phone_tend_only experiments improve on the control forecasts across all spatial scales examined. Forecast skill increases as spatial scales increase, which is expected for the FSS. We note that Roberts and Lean (2008) suggest that forecasts with “useful” skill typically require $FSS \geq 0.5$. Here, this threshold is not reached until spatial scales of 120 km, which indicates that these precipitation forecasts could be considered useful for broad mesoscale precipitation forecasting.

The FSS at the 30-km spatial scale is computed as a function of forecast hour for all the experiments to see how forecast improvements persist in time. Figure 11 shows these FSS values for the phone_only and phone_tend_only experiments relative to the control FSS. The forecast improvement (higher skill) in the phone_only experiment decays with increasing forecast hour, with the advantage over the control being reduced by 80% at hour 4. The skill decay in the phone_tend_only experiment follows a similar pattern, with skills in the phone_only and phone_tend_only experiments being nearly identical by hour 4. Thus, the precipitation forecast improvements provided by the smartphone observations in either form can persist for several hours into the forecast.

Figure 12 shows a reliability diagram for 1-h ensemble probabilistic forecasts of 1-h precipitation ≥ 1 mm using a 30-km neighborhood probability. A perfectly calibrated probabilistic forecast would lie upon the diagonal 1–1 line. All three ensemble experiments, including the control, produce similarly reliable forecasts for probabilities

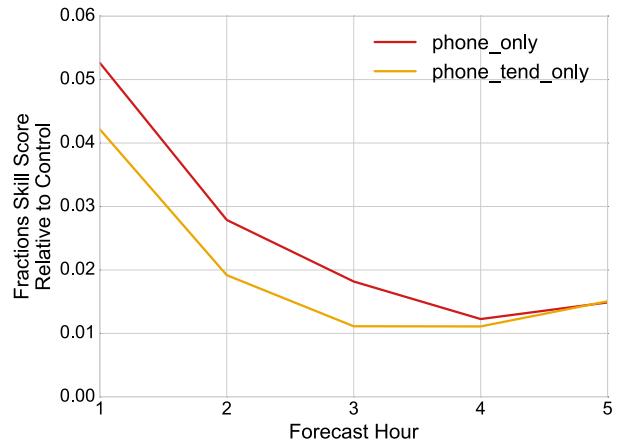


FIG. 11. FSSs relative to the control FSS at 30-km spatial scale for 1-h precipitation accumulation forecasts ≥ 1 mm as a function of forecast hour.

below 35%. There is a slight tendency for under-prediction in all experiments for these lower probabilities, but overall the reliability is good. For higher probabilities, the number of occurrences drops considerably (not shown). From 35% to 50% probability, those experiments assimilating smartphone observations are slightly more reliable than the control, though all begin to deviate from the 1–1 line at this point. The phone_only experiment produces higher

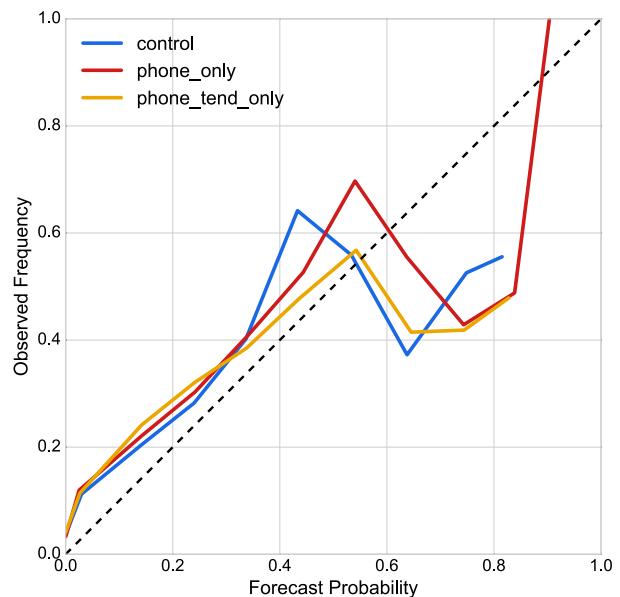


FIG. 12. Reliability diagrams for 0–1-h ensemble forecasts of precipitation ≥ 1 mm with a 30-km neighborhood probability filter applied for the control (blue), phone_only (red), and phone_tend_only (gold) experiments.

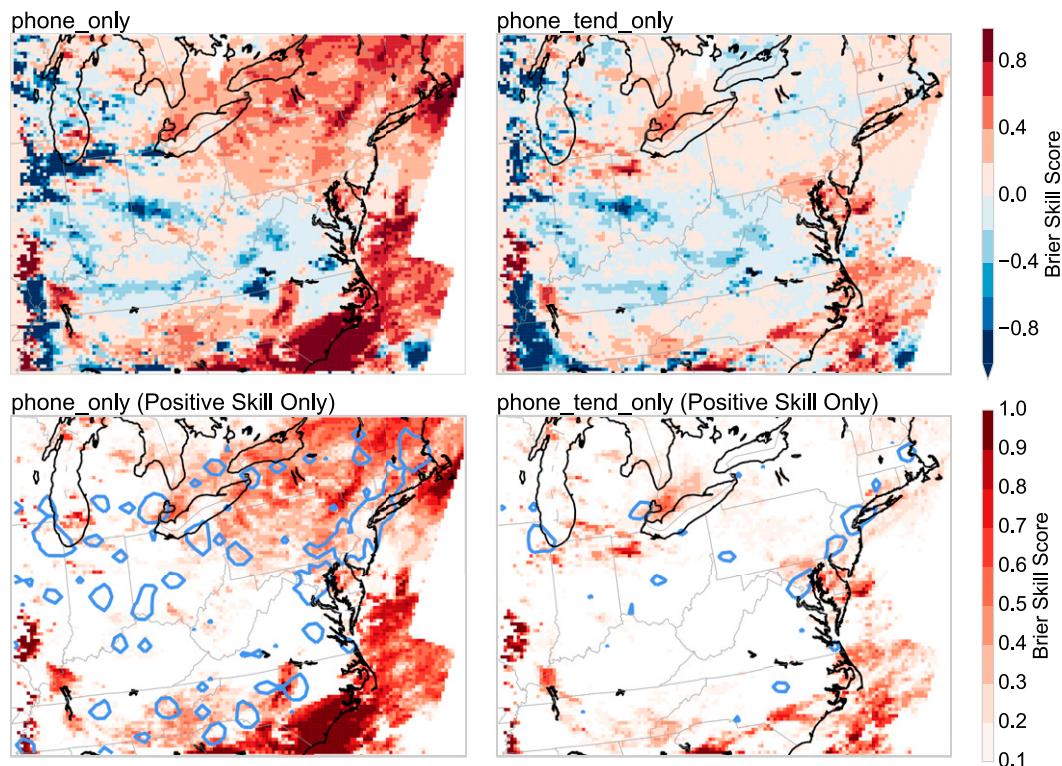


FIG. 13. BSSs of 1-h precipitation forecasts greater than 1 mm for the entire case study after assimilating (top left) only smartphone altimeter observations and (top right) 1-h smartphone altimeter tendency observations with control forecast skill as baseline. This evaluation uses 30-km neighborhood probabilities. (bottom) The areas of positive skill with regions with average observation densities greater than five observations per $30\text{ km} \times 30\text{ km}$ bin outlined in blue. Skill scores are evaluated with 30-km neighborhood probabilities.

probability magnitudes than either the control or the `phone_tend_only` experiments. This is likely related to the greater number of smartphone pressure observations assimilated, which produces lower analysis spread.

To evaluate precipitation forecasts geographically, we compute the BSSs of 1-h ensemble forecasts exceeding a threshold value (here, 1 mm) using the control forecasts (with no assimilation) as the baseline, again with 30-km neighborhood probabilities. Figure 13 shows maps of BSSs over the duration of the case study for the first hour of each forecast. Because of the relatively short time period of the case study, precipitation is not observed at all points and the resulting maps are somewhat noisy. However, distinct patterns are detectable.

For the 1-mm threshold in the `phone_only` experiment (Fig. 13, left panels), there are broad areas of notable improvement over the control ($\text{BSS} > 0$), particularly over the northeastern states and in eastern North Carolina. There are some areas where skill is reduced from the control, chiefly in the southern Great Lakes area near Chicago, Illinois, and into Indiana. There are several possible explanations for this. One is that the areas

of skill degradation are near the western edge of the domain, where there are relatively few upstream observations assimilated that could help constrain the flow.

A second explanation involves the structure of the precipitation. The precipitation in the southern Great Lakes area was mainly due to isolated convection initiated early in the study period in advance of stronger synoptic-scale forcing (Figs. 1 and 2). Madaus and Hakim (2016) find that the convective-scale pressure signals preceding isolated convective initiation are likely too small in magnitude and spatial scale to be detected except by an extremely dense network. While the Chicago metropolitan area has a high density of smartphone observations, this high density is limited to a small area, and the surrounding rural areas (where much of the convection occurred) have significantly lower densities (Fig. 3). When combined with the expected noise and error in SPOs, it is possible that local errors in nearby smartphone observations led to poorly initialized convective features, degrading skill. In contrast, the precipitation in the eastern part of the domain occurred in response to more organized meso- and synoptic-scale

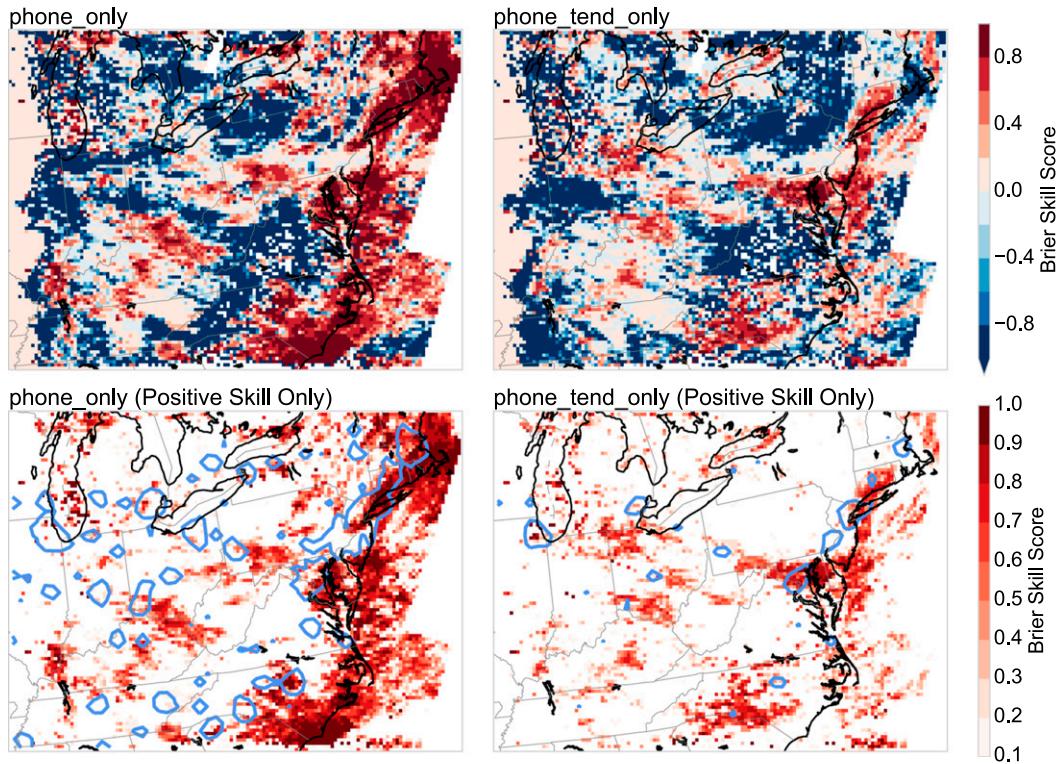


FIG. 14. As in Fig. 13, but for a precipitation threshold of 10 mm.

forcing in the latter half of the study period (Fig. 2). We expect that surface pressure observations are able to constrain these features better and, in turn, produce improved precipitation forecasts.

The *phone_tend_only* experiment does not produce as large a forecast improvement as the *phone_only* experiment, particularly over the northeastern United States. However, these observations do produce improvements in the mid-Atlantic and eastern North Carolina. In addition, the *phone_tend_only* observations do a better job with precipitation in the southern Great Lakes area, again suggesting that error (here, bias) in the raw phone altimeter observations may have contributed to poor convective initiation forecasts.

We repeat the analysis above, but for a threshold of 10 mm in Fig. 14 to examine if forecast skill can extend to more intense rain events where convective processes are more likely occurring. For both the *phone_only* and *phone_tend_only* experiments, the overall improvement of forecasts at this threshold is not as widespread, with large areas of skill degradation. In addition, sub-optimal adjustment of the ensemble state near the western boundary of the domain tended to produce spurious convective features that were quickly suppressed. This contributed to negative skills near the western boundary.

However, when the areas of improvement are compared with observation density (Fig. 14, bottom panels), it is evident that many of the regions of forecast improvement occur downstream from areas of high observation density. In particular, there are large areas of positive skill along the East Coast, downstream from the urban corridor from Washington, D.C., through Boston, Massachusetts. Additional areas of positive skill are downstream from urban central North Carolina and several interior cities such as Pittsburgh, Pennsylvania, and Columbus and Cincinnati/Dayton, Ohio. Interestingly, the area around and downstream from one area with high observation density—Chicago—does not show robust improvement in precipitation forecasts in Fig. 14. As noted above, we expect that the nature of the precipitation around Chicago—namely, isolated convective initiation without larger-scale organization—likely limits the effectiveness of smartphone observations.

The improvements in skill downstream from urban areas can be examined in light of the statistics shown in section 3a. There appear to be inconsistencies within the smartphone pressure network, with disagreement between nearby observations. This was also true to a more limited degree for the smartphone tendency observations. Although it would be tempting to suggest that the enhanced observation density in urban areas is better at

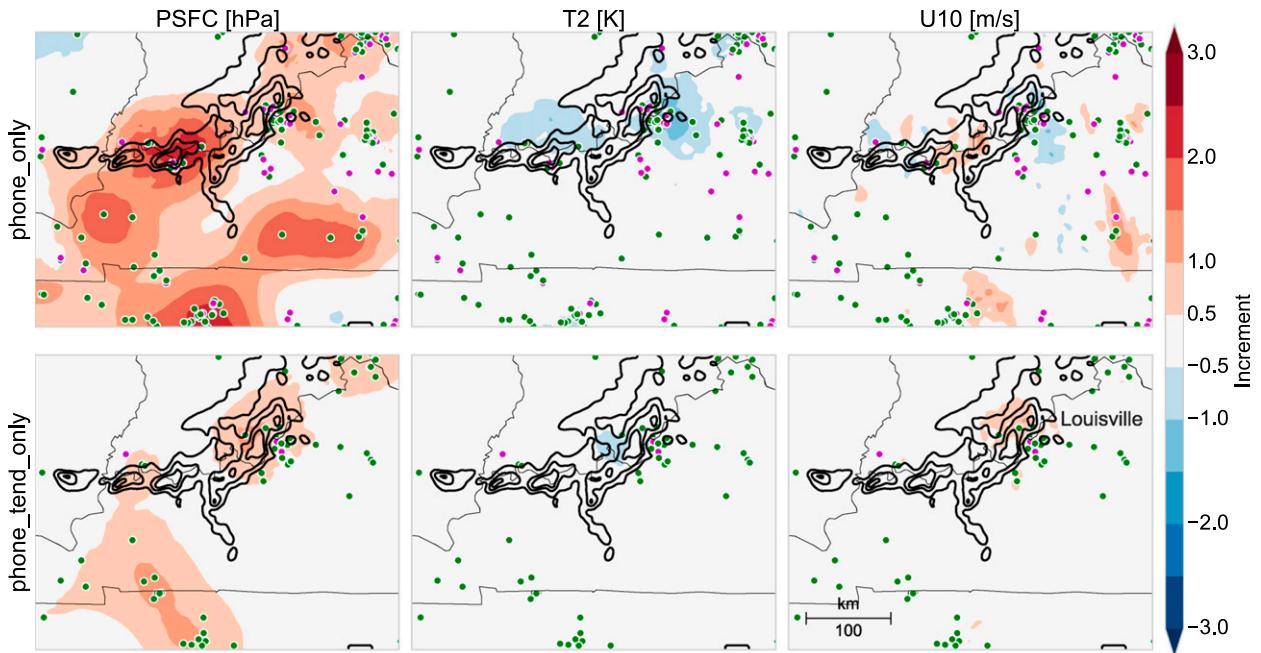


FIG. 15. Assimilation increments for (left) PSFC, (center) T2, and (right) U10 at 0200 UTC 27 Jul 2014 for the (top) smartphone altimeter assimilation experiment (phone_only) and (bottom) 1-h smartphone altimeter tendency experiment (phone_tend_only). Dots indicate the locations of smartphone altimeter observations in the top row and smartphone altimeter tendency observations in the bottom row assimilated at this time. Green dots indicate locations where the error magnitude was larger in the analysis (see section 3b), while purple dots are where the error magnitude was reduced in the analysis. Observed reflectivity at 1-km elevation composited from surrounding NEXRAD radar sites is shown for reference, contoured at 35, 40, and 45 dBZ (black contours). The area of Louisville is the concentration of observations to the left of the label in the bottom-right panel.

resolving convective features, the noise in these observations makes this unlikely. Complicating sensor error in urban areas is the number of high-rise buildings, which could contribute to greater variability in the pressures reported by smartphones at similar locations. Despite this observation noise, an alternative explanation for skill improvement is that within these dense urban areas, there are a sufficient number of observations such that inconsistent values are largely being canceled out during the assimilation. What emerges is a consensus on a true pressure signal in that area, which contributes to an improved forecast. This would also better explain the *reduction* in skill in areas away from dense observations. With fewer observations, error in any one observation is less likely to be canceled by surrounding observations. These errors are then able to propagate and magnify downstream.

f. Performance for the 27 July 2014 Ohio River valley MCS

1) BACKGROUND AND ASSIMILATION PERFORMANCE

Between 0000 and 0600 UTC, a mesoscale convective system formed in southeastern Illinois and propagated

southeastward through the southern Ohio River valley. This represented one of the most organized convective features seen during the case study and was not forecast well by the HRRR (Fig. 1, boxed region). Furthermore, well-developed MCSs have well-established cold pools with pronounced pressure signatures (Houze 2004), making them a promising target for mesoscale pressure observations. Thus, this poorly forecast MCS represents a good candidate for investigating forecast contributions from smartphone pressure observations.

Figure 15 shows assimilation increments in the phone_only and phone_tend_only experiments for a variety of surface fields at 0200 UTC 27 July 2014 as the developed MCS is crossing the Ohio River. Note that there are a large number of smartphone observations in the Louisville, Kentucky, area, which is about to be impacted by the MCS. For the phone_only experiment (Fig. 15, top), the pressure increments at this time are broad in scale, with positive pressure increments of 1–2.5 hPa along the southern end of the MCS where convective cells were more discrete. There are additional large areas of pressure increases well ahead of the MCS to the south. Temperature increments (Fig. 15, center) show some areas of negative adjustment to the north of the southwestern end of the convective line, which is an

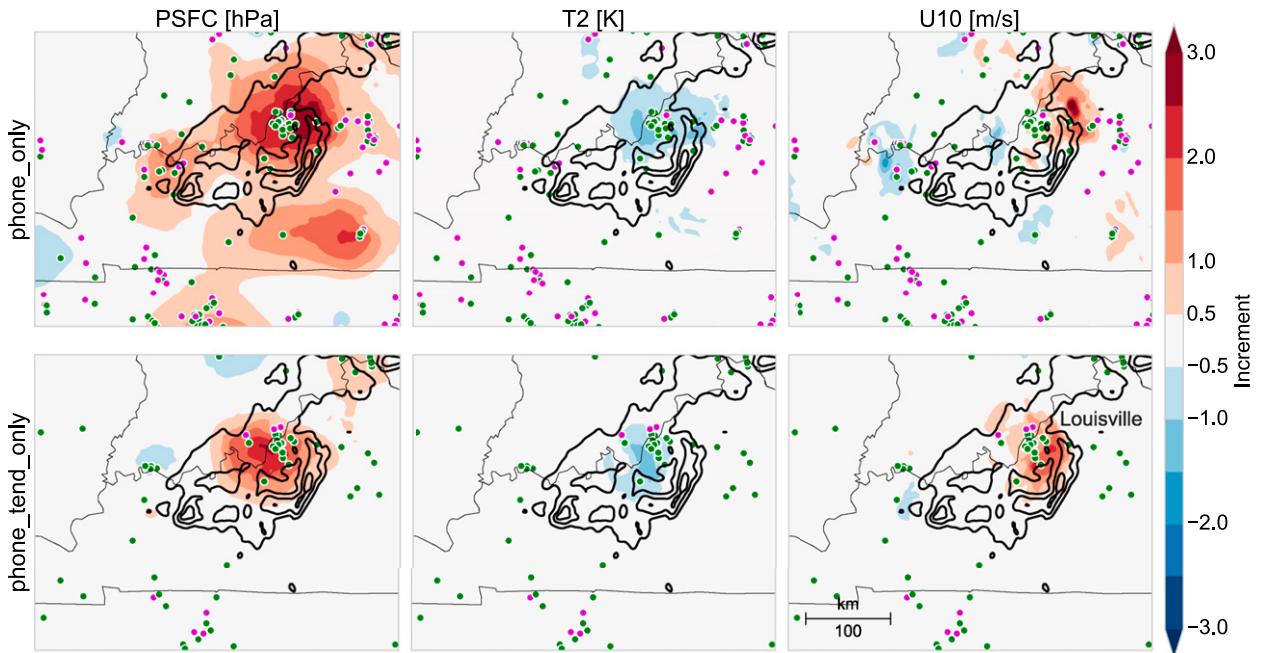


FIG. 16. As in Fig. 15, but for the assimilation at 0300 UTC 27 Jul 2014.

appropriate place for cold anomalies given the expected presence of a developing cold pool there. However, there are additional negative temperature increments of 1 K just ahead of the northeastern end of the MCS, which are more difficult to attribute. The U10 increments hint at increased convergence along the organized convective band that makes up the northeastern half of the MCS, which is broadly expected.

In the *phone_tend_only* experiment (Fig. 15, bottom), the increments are well aligned with the expected structure of the MCS. There are positive pressure increments and negative temperature increments immediately behind the developed convective line in the eastern half of the MCS. Additionally, increased low-level convergence is inferred within the convective line with positive U10 increments to the west. However, we also note that these increments are limited mostly to the vicinity of the dense concentration of observations around Louisville. As such, the alignment of the increments in relation to the MCS may be coincidental, but examining increments at other assimilation times suggests persistent bias is not involved (not shown).

The increments 1 h later (0300 UTC 27 July 2014) are shown in Fig. 16. By 0300 UTC, the organized convective line at the leading edge of the MCS has progressed southeastward into Kentucky and crossed the Ohio River. The Louisville area (with its higher density of smartphone observations) is now contained within the trailing cold pool of the MCS. The increments in both

the smartphone altimeter and smartphone altimeter tendency assimilation experiments reflect this. The positive increments produced by the altimeter observations do include the area where pressure rises are expected behind the convective line, but these pressure rises extend out in front of the line as well. There are additional large increments well out ahead of the line. The altimeter tendency observations produce positive pressure increments that are positioned along and behind the convective line and within the trailing stratiform region. Similarly, the temperature increments for both experiments show an anticipated negative adjustment in the area of the cold pool. For U10 increments the largest positive increments from the altimeter observations occur to the north and east (ahead) of the convective line. The altimeter tendency observations produce positive u -wind increments behind the leading convective line, contributing to increased convergence within the convective line (where it would be expected).

To summarize these assimilation increments, both the smartphone altimeter observations and the 1-h altimeter tendency observations appear to capture the signal of this MCS as it progresses through the area, and adjustments are made to the low-level temperature and wind fields that are consistent with the structure of an MCS and its associated cold pool. Figures 15 and 16 also show whether or not the assimilated observations (colored dots) reduce (green) or increase (purple) error in the analysis state. A proportionally larger number of the

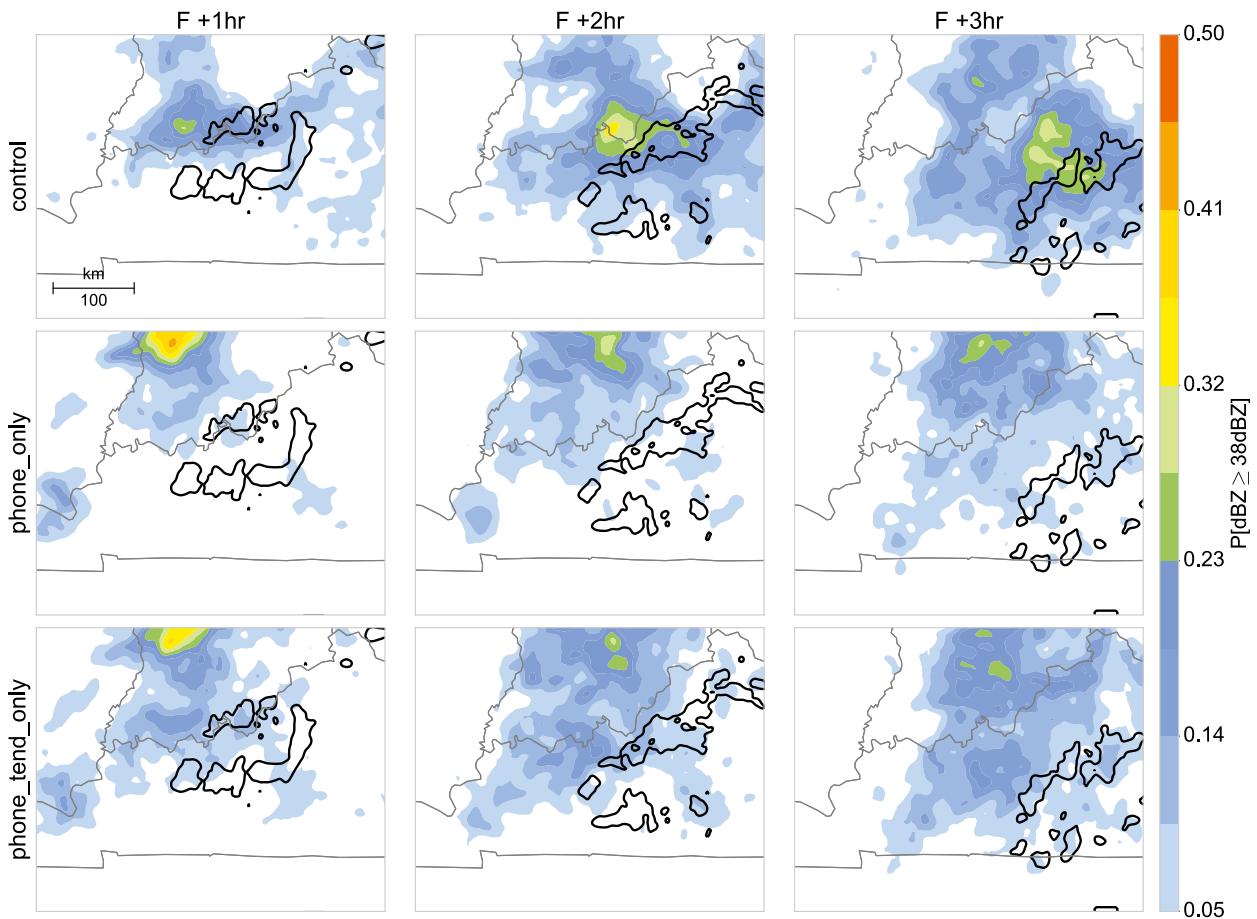


FIG. 17. Ensemble 1–3-h forecast probabilities of 1-km simulated reflectivity exceeding 38 dBZ for the (top) control, (middle) phone_only, and (bottom) phone_tend_only experiments initialized at 0200 UTC 27 Jul 2014. Ensemble probabilities are calculated with a 30-km neighborhood. The 38-dBZ observed 1-km compositing reflectivity from surrounding NEXRAD radars is shown by black contours.

altimeter observations showed increased analysis error for these assimilation times as compared to the altimeter tendency observations, reflecting the finding in section 3a that there is disagreement between observations.

2) FORECAST EVALUATION FOR THE MCS CASE

The ensemble forecasts of convective evolution following the 0200 UTC assimilation cycle are shown in Fig. 17. In both the phone_only and phone_tend_only experiments, despite low-level increments in Figs. 15 and 16 consistent with the presence of the MCS, the adjustments are insufficient to promote the initiation of convection near the observed MCS. Higher ensemble probabilities for convective activity remain well to the north and west of where the MCS was observed at these times. The phone_tend_only experiment does maintain higher probabilities in the vicinity of the observed MCS, but the maximum probability area is clearly removed from this location and in an area where no significant convection is observed. In contrast, the control experiment produced a better forecast, with probability

maxima much closer to the observed MCS in the 0200 UTC (Fig. 17) forecasts. The MCS was also not initiated in the forecasts initiated at 0300 UTC in the phone_only and phone_tend_only experiments (not shown). Examinations of the full three-dimensional thermodynamic fields in both experiments indicated that, despite low-level increments that favor initiating convection, the assimilated SPOs did little to weaken the convective inhibition aloft at these assimilation times (not shown). This not only reflects the challenges of data assimilation in the presence of convective features with limited ensemble sampling, but also illustrates the limitations of SPOs for capturing the full thermodynamic state of the atmosphere. For convective events, a more comprehensive observing network is required.

4. Summary and future work

We evaluated the performance of a mesoscale, convection-allowing ensemble forecast system that makes use of pressure and pressure tendency observations from

smartphones. We examined a 3-day period marked by several rounds of convective activity in the east-central United States. Smartphone pressure observation density closely mirrors population density, resulting in an unevenly distributed surface observation network that varies in time. Quality control efforts to remove suspect observations reject, on average, 64% of possible smartphone observations, with the data assimilation system rejecting an additional 5%–6% as outliers. Assimilation statistics show that analysis error is actually increased at the locations of many of the assimilated pressure observations, reflecting poor agreement within the observing network. The 1-h pressure tendency observations, while only numbering about 20% of the total available pressure observations, appeared to show better agreement.

Analyses and forecasts after assimilating these observations had mixed levels of performance. Compared with the control experiment, analyses of surface pressure were improved, but no improvement was seen in 2-m temperature or 10-m wind analyses when evaluated with independent observations. One-hour forecasts of surface pressure and 10-m wind showed some improvement over the control, but errors were higher in 1-h 2-m temperature forecasts. Precipitation forecasts for hourly accumulations of ≥ 1 mm showed small, but broad, improvement across all spatial scales when either observation type was assimilated. Forecasts for 1-h precipitation accumulations ≥ 10 mm were improved in areas downstream from areas with high observation density, but there were skill reductions in areas where the observation density is sparse. We speculate that this is likely due to errors canceling in areas of high observation density.

A specific convective event—an MCS in the Ohio River valley early on 27 July 2014—was examined to evaluate the contributions of smartphone pressure and pressure tendency observations. Assimilation increments during this event agree well with expected features of MCS structure, suggesting that the smartphone pressure observations are able to capture the pressure signal from some mesoscale features. Despite these promising assimilation increments, the ensemble background states (which themselves were the product of cycling on the smartphone observations) were unable to initiate convection in response to these increments, resulting in poor forecast performance for the MCS.

In this work, we compared the impact of SPOs and SPO tendencies to a simple control experiment that assimilated no additional observations. The forecast improvements over this control were marginal and, while suggesting that SPOs can capture some meteorological signal, the impacts of the current SPO network appear small. Given these results, we suggest that further

research into smartphone pressure contributions to mesoscale forecasts must wait for better quality control methods, improved data assimilation techniques, or a large increase in available SPO density.

Research into these areas is ongoing. These results suggest that error within SPOs (or likely any crowd-sourced observation type) is a critical issue. There are current efforts under way to more robustly characterize the errors in the smartphone pressure network and correct for them. Data assimilation methods—such as combining observations (superobservations) in areas of dense observations or adapting localization to better suit correlation length scales—may also improve the SPO impact and could be investigated. Future work should also examine what role smartphone observations may play as a part of a more comprehensive mesoscale observation system. Finally, there has been a groundswell of interest in smartphone pressure observations, and a number of additional providers with substantially more access to the global smartphone network are working to collect smartphone pressure observations. We expect the number of available smartphone observations to rise substantially in the coming years and hope to revisit these observations at that time.

Acknowledgments. This work was supported through NOAA CSTAR Award NA13NWS4680006 and computing funding from NSF Grant AGS-1349847. We would also like to acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Greg Hakim also provided useful guidance and commentary throughout the development and execution of this study. Conor McNicholas has taken the lead in evaluating smartphone pressure observation quality and was responsible for developing and implementing the quality control algorithms used to clean up the smartphone pressure observations. We greatly appreciate his help in getting these observations into shape and look forward to his results as he explores smartphone pressure observation quality. Discussions with Glen Romine at NCAR were crucial to the development of the HRRR-based ensemble used here, specifically with respect to the use of the SKEB scheme to drive ensemble diversity.

REFERENCES

- Anderson, J., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi:[10.1175/1520-0493\(2001\)129<2884:AEAKFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2).
- , 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83, doi:[10.1111/j.1600-0870.2008.00361.x](https://doi.org/10.1111/j.1600-0870.2008.00361.x).

- , 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, doi:10.1175/MWR-D-11-00013.1.
- , T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, doi:10.1175/2009BAMS2618.1.
- Benjamin, S., and Coauthors, 2015: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.
- Berner, J., S. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, doi:10.1175/2010MWR3595.1.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- Burton, P., 2013: IFS documentation—Cy40r1. Part I: Observations. ECMWF Tech. Rep., 76 pp. [Available online at http://www.ecmwf.int/sites/default/files/IFS_CY40R1_Part1.pdf.]
- Dabberdt, W. F., and Coauthors, 2005: Multifunctional mesoscale observing networks. *Bull. Amer. Meteor. Soc.*, **86**, 961–982, doi:10.1175/BAMS-86-7-961.
- Dong, J., M. Xue, and K. Droegemeier, 2011: The analysis and impact of simulated high-resolution surface observations in addition to radar data for convective storms with an ensemble Kalman filter. *Meteor. Atmos. Phys.*, **112**, 41–61, doi:10.1007/s00703-011-0130-3.
- Dubin, M., A. R. Hull, and K. S. W. Champion, 1976: *U.S. Standard Atmosphere, 1976*. NOAA/NASA Tech. Rep., U.S. Government Printing Office, 227 pp. [Available online at http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19770009539_1977009539.pdf.]
- Fowle, M. A., and P. J. Roebber, 2003: Short-range (0–48 h) numerical prediction of convective occurrence, mode, and location. *Wea. Forecasting*, **18**, 782–794, doi:10.1175/1520-0434(2003)018<0782:SHNPOC>2.0.CO;2.
- Gallus, W. A., Jr., J. Correia, and I. Jankov, 2005: The 4 June 1999 derecho event: A particularly difficult challenge for numerical weather prediction. *Wea. Forecasting*, **20**, 705–728, doi:10.1175/WAF883.1.
- Gaspari, G., and S. Cohn, 2006: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **132**, 723–757, doi:10.1002/qj.49712555417.
- Ha, S., J. Berner, and C. Snyder, 2015: A comparison of model error representations in mesoscale ensemble data assimilation. *Mon. Wea. Rev.*, **143**, 3893–3911, doi:10.1175/MWR-D-14-00395.1.
- Houze, R. A., Jr., 2004: Mesoscale convective systems. *Rev. Geophys.*, **42**, RG4003, doi:10.1029/2004RG000150.
- Hu, M., H. Shao, D. Stark, and K. Newman, 2013: Gridpoint statistical interpolation (GSI) version 3.2 user's guide. DTC Tech. Rep., Developmental Testbed Center, 181 pp. [Available online at http://www.dtcenter.org/com-GSI/users/docs/users_guide/GSIUserGuide_v3.2.pdf.]
- Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.]
- Madaus, L. E., and G. J. Hakim, 2016: Observable surface anomalies preceding simulated isolated convective initiation. *Mon. Wea. Rev.*, **144**, 2265–2284, doi:10.1175/MWR-D-15-0332.1.
- , —, and C. F. Mass, 2014: Utility of dense pressure observations for improving mesoscale analyses and forecasts. *Mon. Wea. Rev.*, **142**, 2398–2413, doi:10.1175/MWR-D-13-00269.1.
- Mass, C. F., and L. E. Madaus, 2014: Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343–1349, doi:10.1175/BAMS-D-13-00188.1.
- , D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.
- McPherson, R. A., and Coauthors, 2007: Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321, doi:10.1175/JTECH1976.1.
- Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, doi:10.2151/jmsj.87.895.
- Olson, J. B., and J. M. Brown, 2012: Modifications to the MYNN PBL and surface layer scheme for WRF-ARW. *Proc. 2012 WRF Users Workshop*, Boulder, CO, NCAR, 3.3. [Available online at <http://www2.mmm.ucar.edu/wrf/users/workshops/WS2012/abstracts/3.3.htm>.]
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.
- Roebber, P. J., D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. *Wea. Forecasting*, **17**, 399–429, doi:10.1175/1520-0434(2002)017<0399:SROTM>2.0.CO;2.
- , —, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, doi:10.1175/1520-0434(2004)019<0936:TIPHA>2.0.CO;2.
- Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, doi:10.1175/2009WAF2222267.1.
- Skamarock, W., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH.
- Smirnova, T., J. Brown, S. Benjamin, and J. Kenyon, 2015: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the Weather Research and Forecast (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, doi:10.1175/MWR-D-15-0198.1.
- Snook, N., M. Xue, and Y. Jung, 2015: Multiscale EnKF assimilation of radar and conventional observations and ensemble forecasting for a tornadic mesoscale convective system. *Mon. Wea. Rev.*, **143**, 1035–1057, doi:10.1175/MWR-D-13-00262.1.
- Sobash, R. A., and D. J. Stensrud, 2015: Assimilating surface mesonet observations with the EnKF to improve ensemble forecasts of convection initiation on 29 May 2012. *Mon. Wea. Rev.*, **143**, 3700–3725, doi:10.1175/MWR-D-14-00126.1.
- Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.*, **122**, 2084–2104, doi:10.1175/1520-0493(1994)122<2084:MCSIWF>2.0.CO;2.

- Sun, J., and Coauthors, 2014: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bull. Amer. Meteor. Soc.*, **95**, 409–426, doi:[10.1175/BAMS-D-11-00263.1](https://doi.org/10.1175/BAMS-D-11-00263.1).
- Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:[10.1175/2008MWR2387.1](https://doi.org/10.1175/2008MWR2387.1).
- Wheatley, D. M., and D. J. Stensrud, 2010: The impact of assimilating surface pressure observations on severe weather events in a WRF mesoscale ensemble system. *Mon. Wea. Rev.*, **138**, 1673–1694, doi:[10.1175/2009MWR3042.1](https://doi.org/10.1175/2009MWR3042.1).
- Xue, M., and W. J. Martin, 2006: A high-resolution modeling study of the 24 May 2002 dryline case during IHOP. Part I: Numerical simulation and general evolution of the dryline and convection. *Mon. Wea. Rev.*, **134**, 149–171, doi:[10.1175/MWR3071.1](https://doi.org/10.1175/MWR3071.1).