Impacts of Assimilating Smartphone Pressure Observations on Forecast Skill during Two Case Studies in the Pacific Northwest

CALLIE MCNICHOLAS AND CLIFFORD F. MASS

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

(Manuscript received 17 May 2018, in final form 3 August 2018)

ABSTRACT

Over a half-billion smartphones are now capable of measuring atmospheric pressure, potentially providing a global surface observing network of unprecedented density and coverage. An earlier study by the authors described an Android app, uWx, that served as a test bed for advanced quality control and bias correction strategies. To evaluate the utility and quality of the resulting smartphone pressure observations, ensemble data assimilation experiments were performed for two case studies over the Pacific Northwest. In both case studies, smartphone pressures improved the analyses and forecasts of assimilated and nonassimilated variables. In case I, which considered the passage of a front across the region, cycled smartphone pressure assimilation consistently improved 1-h forecasts of the altimeter setting, 2-m temperature, and 2-m dewpoint. During a postfrontal period, cycled smartphone pressure assimilation improved mesoscale forecasts of hourly precipitation accumulation. In case II, which considered a major coastal windstorm, cycling experiments assimilating smartphone pressures improved 10-m wind forecasts as well as the predicted track and intensity. For both cases, free-forecast experiments initialized with smartphone data produced forecast improvements extending several hours, suggesting the utility of crowdsourced smartphone pressures for short-term numerical weather prediction.

1. Introduction

Surface pressure observations can provide information on all scales of motion, ranging from convectively produced cold pools to midlatitude cyclones. Surface pressure is a particularly valuable surface parameter, since it reflects atmospheric structure through the full depth of the atmosphere and is less influenced by exposure and representation errors than surface temperature, moisture, and wind. These characteristics have motivated interest in evaluating the potential of surface pressure observations for improving data assimilation and numerical weather prediction.

On the synoptic scale, experiments assimilating only surface pressure observations have reproduced upper-tropospheric large-scale circulations (Compo et al. 2006) and generated realistic lower- and middle-tropospheric analyses (Whitaker et al. 2004; Dirren et al. 2007). Considering mesoscale simulations, Wheatley and Stensrud (2010) noted that the hourly assimilation of altimeter setting, and, to a limited degree, altimeter tendency reduced errors in mesohigh position and intensity, resulting in improved model depiction of cold pools. Madaus et al. (2014) assimilated 3-hourly altimeter and altimeter tendency observations from a high-density network of routine airport observations (METARs) and bias-corrected mesonet observations. A monotonic decrease in domain-averaged analysis error occurred as the number of assimilated pressure observations increased.

Since surface pressure alone can constrain model initializations at the surface and aloft, and model initializations are improved as observational density and frequency increases (Anderson et al. 2005; Lei and Anderson 2014; Madaus et al. 2014), large numbers of pressure observations from smartphones possess the potential for improving numerical weather prediction. Surface pressure observations from smartphones offer unparalleled density and can be collected at high temporal frequency (McNicholas and Mass 2018). Hanson (2016), using observation system simulation experiments with synthetic smartphone pressures, concluded that if observational uncertainty could be estimated, smartphones pressures could improve model forecasts.

The development of several crowdsourcing pressure applications, such as PressureNet and WeatherSignal, facilitated the initial evaluation of smartphone pressures for analysis and numerical weather prediction. Mass and Madaus (2014) described the potential of crowdsourcing

Corresponding author: Callie McNicholas, cmcnich@uw.edu

DOI: 10.1175/WAF-D-18-0085.1

smartphone pressures for mesoscale numerical weather prediction (NWP) and provided an example of smartphone pressure assimilation for a convective event in eastern Washington State. Madaus and Mass (2017) found that assimilating hourly smartphone pressures resulted in limited improvements to altimeter setting forecasts and a small reduction in forecast skill for other surface variables, such as 2-m temperature and 2-m humidity. The limited positive impact of smartphone pressure observations appeared to result from poor data quality. Madaus and Mass (2017) did not account for sensor bias and elevation uncertainty, undermining their ability to constrain model forecasts.

The results of Madaus and Mass (2017) motivated a follow-up study (McNicholas and Mass 2018, hereafter MM2018) in which smartphone pressure observations (SPOs) were collected from an Android app (uWx; www.cmetwx.com) that allowed the evaluation of pressure collection and quality control strategies. In uWx, sources of error were reduced, and observational uncertainty was quantified. A machine learning approach predicted and corrected smartphone pressure biases in real time, resulting in marked improvements in the quality of SPOs.

In this study, we evaluate the impacts of the quality control and bias-correction strategies of MM2018 on numerical weather prediction by performing ensemble data assimilation of SPOs, with and without bias correction/quality control, for two case studies. In the first case, an intensifying surface low and trailing cold front traversed the uWx SPO network. The second case study simulated a strong, compact midlatitude cyclone that formed from the remnants of Tropical Storm Songda. In this case, operational systems misplaced the location of landfall, resulting in poor surface wind forecasts.

The remainder of this paper is organized as follows. In section 2, the two events are reviewed. Section 3 describes the design and methodology of the data assimilation/forecasting experiments for both cases. The results of the experiments are examined in sections 4 and 5, respectively. Section 6 discusses the conclusions and implications of this study.

2. Case descriptions

The two cases selected for this study reflect two important types of events in the Pacific Northwest: 1) a typical surface low and frontal passage with postfrontal precipitation and 2) a major coastal windstorm.

a. Case I

This case represents a familiar scenario for operational forecasts in the Pacific Northwest: a surface low and cold frontal passage. Figure 1 provides a synoptic overview of this case. At 1200 UTC 15 November 2016, a surface low was positioned over western Washington, with a weak pressure trough and associated cold front to the south. Aloft (500 hPa), southwesterly flow dominated the region, with a jet streak extending off the Pacific Ocean into northern Oregon. The 15-h forecast from the operational High Resolution Rapid Refresh (HRRR; Blaylock et al. 2017) overestimated the east–west pressure gradient in western Washington, with positive errors (~2 hPa) along the Oregon and southwest Washington coasts and excessively low pressure over the Cascades, eastern Washington, and west of Vancouver Island. The surface temperature errors had less structure, with the low and trough generally being modestly cooler than observed.

b. Case II

In case II, a coastal cyclone developed from the remnants of Tropical Storm Songda (Fig. 2). At 0300 UTC 16 October 2016, a deep surface low was centered over Vancouver Island beneath a negatively tilted 500-hPa trough. The tight pressure gradient around the low produced strong near-surface winds (>25–30 kt, where $1\,\text{kt} \approx 0.51\,\text{m s}^{-1}$) over the waters surrounding Vancouver Island. Over the Puget Sound region, the observed near-surface wind speeds were relatively modest (10–15 kt); however, short-range forecasts from multiple operational systems such as the NOAA/NWS Global Forecasting System (GFS), the University of Washington's WRF Model (UW-WRF), and the NOAA/NWS HRRR moved the surface low over the Olympic Peninsula, bringing gale-force wind gusts to the western Washington interior. The 15-h HRRR forecasts misplaced the location of landfall, bringing the surface low approximately 100 km too far east, with large pressure errors (too low) over southern Vancouver Island. Consequently, there were significant near-surface wind forecast errors, most notably in northwest Washington and the eastern Strait of Juan De Fuca, where the predicted winds were too strong. The potential for SPOs to constrain pressure forecasts, especially errors in the intensity and track of the surface low, motivated this case.

3. Methodology

a. Model setup

For all ensemble data assimilation (DA) experiments, simulations were performed with the WRF Model (Skamarock et al. 2008). WRF was run with 38 vertical levels, a horizontal grid spacing of 4 km, and a domain

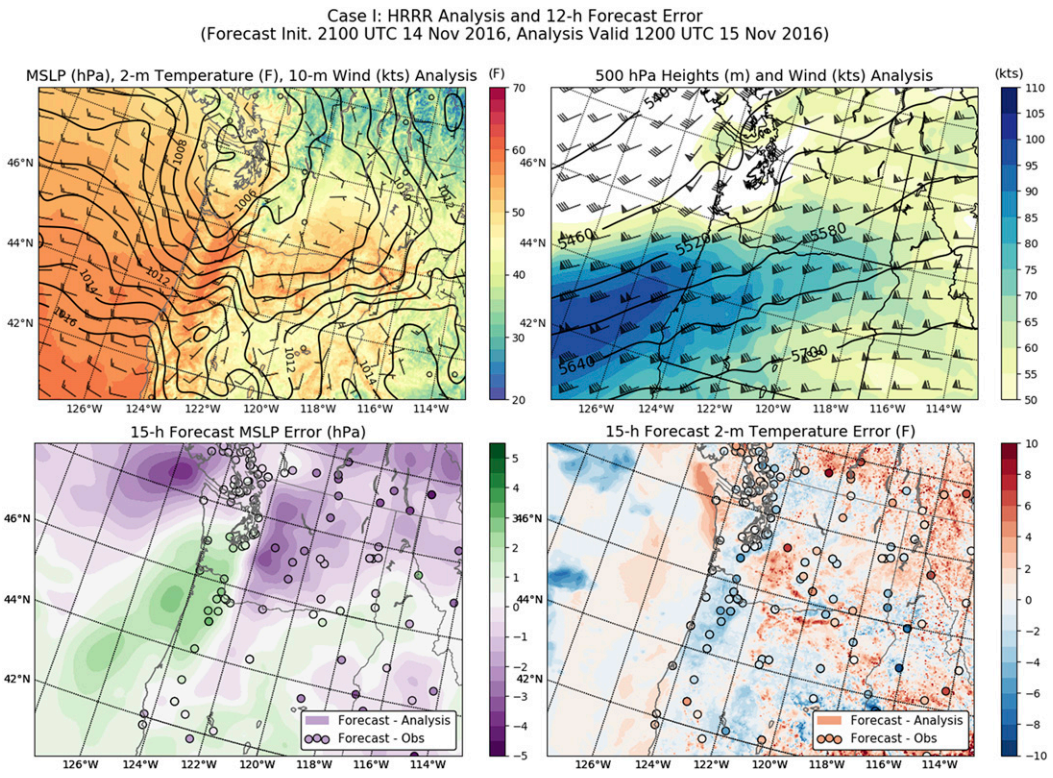

FIG. 1. HRRR analysis and 15-h forecast errors at 1200 UTC 15 Nov 2016. (top left) MSLP, 2-m temperature, and 10-m wind analysis. (top right) The 500-hPa heights and wind analysis. (bottom left) The 15-h HRRR forecast error for MSLP. (bottom right) The 15-h forecast error for 2-m temperature.

encompassing most of the Pacific Northwest. The model domain was centered at (46°N, 122°W) and had dimensions of 1200 km × 900 km. Physics parameterizations (Table 1) reflect those used in the operational National Centers for Environmental Prediction (NCEP) HRRR model (Benjamin et al. 2016). A total of 48 WRF ensemble members were produced using the stochastic kinetic-energy backscatter scheme (SKEBS) to perturb WRF initial and boundary conditions (Berner et al. 2011). SKEBS parameter values are listed in Table 2. Initial conditions at the beginning of a 12-h spinup period were provided by the NOAA/ESRL Rapid Refresh model analysis (RAP; Benjamin et al. 2016), with hourly boundary conditions generated with RAP 1-h forecasts to emulate a real-time cycled DA system in which RAP forecasts are available approximately 1 h after nominal time.

b. Data assimilation

Assimilation experiments were conducted on the Microsoft Azure Cloud using the Data Assimilation Research Testbed (DART; Anderson et al. 2009) ensemble square root adjustment filter. Table 3 lists WRF state variables updated by DART during assimilation experiments. Spatially and temporally varying adaptive covariance inflation was employed to promote and maintain ensemble spread (Anderson et al. 2009). Sampling error correction was applied to help maintain ensemble spread and constrain sampling errors associated with limited ensemble size (Anderson 2012). Gaspari–Cohn localization was used in the horizontal, with a half-width of 500 km (Gaspari and Cohn 2006). Adaptive localization applied a threshold of 500 observations to decrease the localization cutoff in regions of dense observations (Anderson and Collins 2007). For this study, this procedure effectively reduced the localization radius for SPOs to approximately 330 km. The DART system includes quality control (QC) checks on observations to improve assimilation quality. Specifically, when the difference between an observation and the ensemble-mean estimate of that observation exceeded 3 times the ensemble spread, the observation is rejected as an outlier. Surface observations whose elevation deviated from the model elevation by more than 200 m were not assimilated.

c. Experimental design

In each case study, a control (CNTRL) ensemble was generated using the approach outlined in Fig. 3a. For the

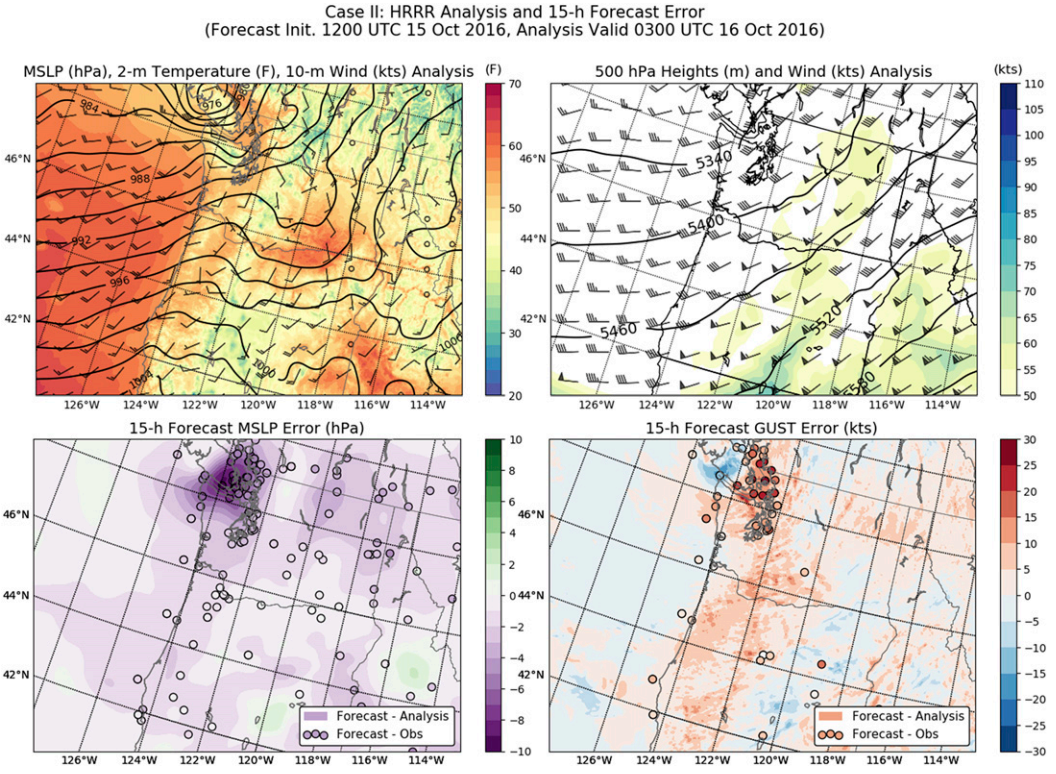

FIG. 2. HRRR analysis and 15-h forecast error at 0300 UTC 16 Oct 2016. (top left) MSLP, 2-m temperature, and 10-m wind analysis. (top right) The 500-hPa heights and wind analysis. (bottom left) The 15-h forecast error for MSLP. (bottom right) The 15-h forecast error for surface wind gusts.

control simulations, the 48-member WRF ensemble was advanced hourly with SKEBS perturbed boundary conditions from 1-h RAP forecasts but with no DA. Conversely, in no-cycling DA experiments (Fig. 3b), the 1-h forecast (prior) from the CNTRL ensemble is updated with observations using the DART ensemble square root filter to create an analysis (a.k.a., the posterior). Since the model is not advanced from updated analyses, no-cycling DA experiments are designed to examine the impact of pressure assimilation on model analyses, given the same prior states. To examine the impact of pressure assimilation on forecasts, cycled DA is performed (Fig. 3c) wherein the model is advanced from analyses (posteriors) produced by assimilating surface pressure observations with DART.

For each case, DA experiments were performed over a 60-h period. For case I, no-cycling DA and cycling DA experiments were performed with SPOs, METARs, and mesonet surface pressure observations available from the Meteorological Assimilation Data Ingest System (MADIS) between 1200 UTC 14 November and 0000 UTC 17 November 2016. In case II, cycling DA experiments with SPOs were performed from 1200 UTC 14 October to 0000 UTC 17 October 2016. All DA experiments were verified with quality-controlled METAR observations.

In both cases, extended forecasts were initialized from the CNTRL ensemble and SPO cycled ensembles to evaluate the impact of SPO assimilation at lead times beyond 1 h. In case I, 6-h free forecasts were initialized every 6 h beginning at 1200 UTC 14 November and ending at 0000 UTC 17 November 2016. In case II, 5-h free forecasts were initialized at 2300 UTC 15 October 2016. In all free-forecast experiments, the full 48-member WRF ensemble was advanced with SKEBS perturbed boundary conditions from the RAP model.

TABLE 1. WRF physics parameterizations.

Physics	Parameterization
Microphysics	Thompson
Planetary boundary layer	Mellor–Yamada–Nakanishi–Niino (MYNN)
Cumulus	None
Shortwave radiation	RRTMG
Longwave radiation	RRTMG
Land surface	RUC land surface model
Surface layer	MYNN

TABLE 2. SKEBS parameter values.

SKEBS parameter	Value
Total backscattered dissipation rate for streamfunction	$5.0 \times 10^{-5}\,m^2\,s^{-3}$
Total backscattered dissipation rate for potential temperature	$1.0 \times 10^{-4}\,m^2\,s^{-3}$
Decorrelation time for streamfunction perturbations	3600 s
Decorrelation time for potential temperature perturbations	3600 s
Spectral slope for streamfunction perturbations	−1.83
Spectral slope for potential temperature perturbations	−1.83

TABLE 3. WRF state variables updated by DART.

WRF state variable	Description
U	*X*-wind component
V	*Y*-wind component
W	*Z*-wind component
PH	Perturbation geopotential
T	Perturbation potential temperature
MU	Perturbation dry air mass in column
QVAPOR	Water vapor mixing ratio
QCLOUD	Cloud water mixing ratio
QRAIN	Rain water mixing ratio
U10	*U* at 10 m
V10	*V* at 10 m
T2	Temperature at 2 m
Q2	QVAPOR at 2 m
PSFC	Surface pressure

4. Observation preprocessing

Both "corrected" and "uncorrected" SPOs are used in this study. Corrected SPOs are bias corrected and quality controlled using the approach outlined in MM2018. Uncorrected SPOs are retrieved prior to bias correction and quality control. Altimeter setting is used in all experiments, with SPOs reduced to sea level using the altimeter equation [Eq. (2)] in MM2018.

a. Observation bias correction

For both case studies, SPOs were corrected following the methodology of MM2018. Specifically, a random forest, machine-learning approach (Breiman 2001) was used with uWx data to predict and correct smartphone pressure bias. Random forests were generated using the Python Scikit-learn machine learning library (Pedregosa et al. 2011). For the first case, random forests were trained from 15 August to 9 November 2016. During and prior to the second case study, uWx was advertised to the public, resulting in a doubling of the number of uWx users from approximately 1000 to 2000. Since many SPOs collected during this case were retrieved from smartphones that had just joined the uWx network, bias correction of SPOs using past data was not possible. As a result, SPOs used in the second case study were bias corrected with random forests trained on data retrieved during the month *after* the event (19 October–23 November 2016).

Quality control of METAR and mesonet observations is performed within MADIS (Miller et al. 2005). Only METAR and mesonet observations that passed the first three stages of MADIS quality control were used in the DA experiments. Because DA was performed hourly, observations were binned by hour. If several observations from a specific METAR or mesonet station fell within 30 min of the hour, only the observation valid closest to the beginning of the hour was retained. This effectively reduced the observation window to 15 min for mesonet observations and 7 min for METAR observations. The same filtering was not performed for SPOs since a single smartphone can provide multiple observations, at unique locations, within a single assimilation cycle.

b. Observation uncertainty

Typically, observation error variances in data assimilation systems are set to a constant value for all altimeter setting observations (Wheatley and Stensrud 2010; Madaus et al. 2014; Madaus and Mass 2017). In this study, the error variances for METAR and mesonet altimeter setting observations were set to 1 and 1.5 hPa^2, respectively. SPO error variance was calculated as the square of the sum of SPO uncertainty, derived in MM2018 and listed in Table 4. This approach was used to calculate the error variance for both uncorrected and corrected SPOs.

The distribution of corrected/uncorrected SPO error variance for case I is displayed in Fig. 4. SPO error variance is right skewed by smartphones with larger bias correction/estimation uncertainty and at locations where the local terrain variance is large. In Table 4, the various contributions to error variance are different for each smartphone. Since individual smartphones contribute both uncorrected and corrected SPOs, the error variance distribution of uncorrected and corrected SPOs is similar. This suggests that uncorrected SPO error variance is underestimated using the approach outlined above.

c. Spatial and temporal characteristics of SPOs

Figure 5a displays the locations of corrected SPOs during the entire period of case I, as well as for a single time: 1200 UTC 14 November 2016. The distribution of mesonet and verification METAR observations is displayed in Fig. 5b. SPO density from the uWx app in the Seattle, Washington, metropolitan area far exceeds that

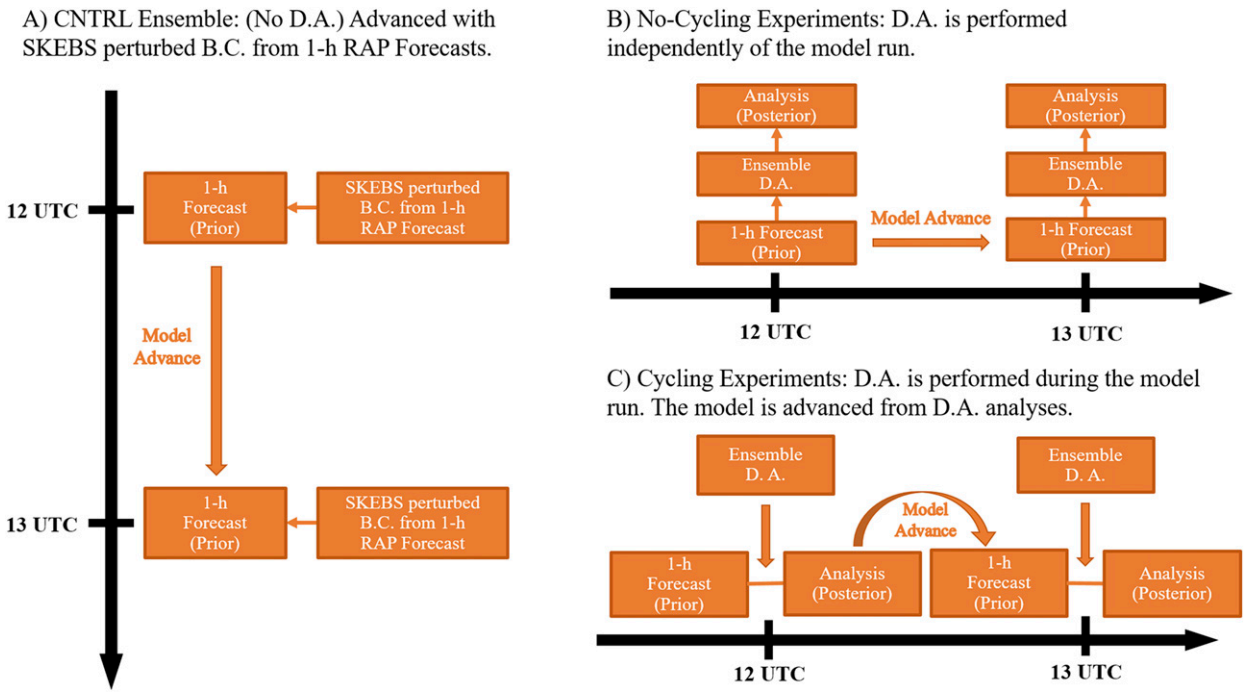

FIG. 3. Schematic illustrating the design and advance of (a) the CNTRL ensemble, (b) the no-cycling DA experiments, and (c) the cycling DA experiments. In this schematic DA refers to data assimilation, the process by which information from observations is incorporated into model analyses/forecasts. The acronym BC refers to boundary conditions, in this study provided by the operational RAP model.

of existing networks (Fig. 5b), while in rural eastern Washington the coverage is sparser.

The number of SPOs available during cases I and II is displayed in Figs. 5c and 5d. The number of available METAR and mesonet observations is also displayed in Fig. 5c, as these observations were assimilated in case I DA experiments. In contrast to mesonets and METARs, there is a substantial diurnal variation in SPO availability. Fewer SPOs are available overnight when smartphone use is reduced and the smartphone operating system is more likely to limit background tasks such as pressure retrieval. During case I, a small fraction of uncorrected SPOs fail DART's standard deviation checks, primarily during the day when more smartphones are in motion or located in urban areas where buildings are taller (Fig. 5b). During case II, the number of available SPOs increased as uWx was advertised to the public in the lead up to the windstorm. In case II, virtually all uncorrected SPOs passed DART's QC checks (Fig. 5c). This reflects the large uncertainty in

TABLE 4. Sources of uncertainty for SPOs. Smartphone pressure error variance is calculated as the square of the sum of the sources of uncertainty.

Source of uncertainty	Description	Type	Median magnitude (hPa)
Measurement noise	Standard deviation of 50 sample sensor time series averaged for pressure retrieval	Unique for all smartphones and SPOs	0.02
Sensor accuracy	Relative accuracy of a typical smartphone pressure sensor (0.17 hPa)	Constant for all smartphones and all SPOs	0.17 (constant)
Elevation uncertainty	Two standard deviations of the local elevation grid at the SPO location	Unique for all smartphones and SPOs	0.34
Pressure bias uncertainty	Uncertainty of the pressure bias estimate upon which uWx random forests are trained (unique for each smartphone)	Unique for each smartphone; constant for all SPOs of a given smartphone	0.42
Bias prediction uncertainty	RMSE of cross-validated random forest bias prediction	Unique for each smartphone; constant for all SPOs of a given smartphone	0.29

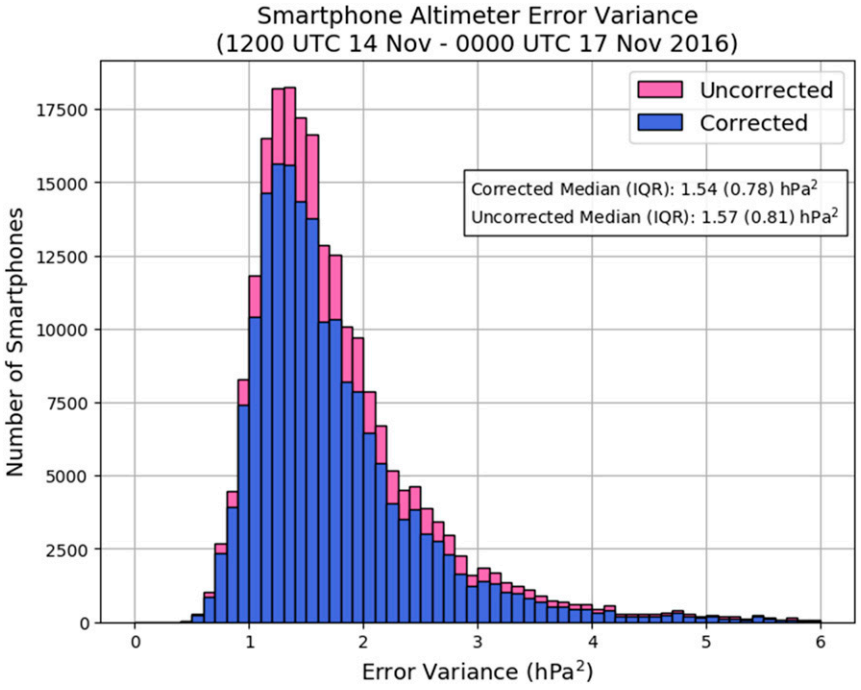

FIG. 4. Corrected and uncorrected smartphone altimeter error variances during case I (1200 UTC 14 Nov–0000 UTC 17 Nov 2016). Note that the right tail of the distribution extends beyond 6 hPa. The histogram is cut off at this value since approximately 99.5% of SPO error variances are less than 6 hPa^2.

the track and intensity of the windstorm in case II, which increased the background ensemble spread, resulting in more lenient DART QC.

5. Case I: Data assimilation and forecast results

a. No-cycling experiments

To evaluate the impact of assimilating SPOs on model analyses, four no-cycling DA experiments were performed. The METAR and MESONET experiments evaluated the impact of assimilating METARs and the mesonet altimeter setting. The PHONE and PHONE_NOQC experiments evaluated the performance of assimilating corrected and uncorrected SPOs, respectively. In all four DA experiments, analysis errors were computed by subtracting METAR observations from the ensemble mean analysis at the locations of all METARs in the model domain.

Figure 6a displays the domain-average altimeter bias for all four DA experiments and the CNTRL. Assimilating METAR, mesonet, and corrected SPOs nearly eliminates the positive pressure bias apparent in the CNTRL. Uncorrected SPOs, many of which were likely retrieved above ground level, introduced a systematic low pressure bias in no-cycling analyses. In the CNTRL, the domain-average pressure bias was a result of 2–3-hPa (positive) pressure biases throughout the Columbia River basin, in the lee of the Cascade Mountains. The CNTRL forecasts, in this region, were characterized by anomalously low temperature and anomalously high pressure throughout the case.

Domain-averaged RMSE was computed each hour for several variables from the analysis error at all METAR locations in the model domain (see the appendix for details). Figure 6b displays the domain-averaged time series of altimeter RMSE for the CNTRL and four no-cycling DA experiments. Period-averaged differences in RMSE between the CNTRL and the four experiments are displayed in the right panel. Relative to the prior (CNTRL), assimilating corrected SPOs consistently reduced the altimeter analysis error at METAR locations by approximately 0.5 hPa (~50%). Assimilating uncorrected SPOs proved nonbeneficial to altimeter setting analyses as the time-averaged altimeter RMSE in the PHONE_NOQC experiment was not significantly different from CNTRL. The assimilation of mesonet altimeter setting resulted in a median altimeter RMSE reduction of 0.6 hPa. The largest reduction in altimeter RMSE was achieved when METAR altimeter setting observations were assimilated. This result is expected as the assimilated observations were not independent of the verification.

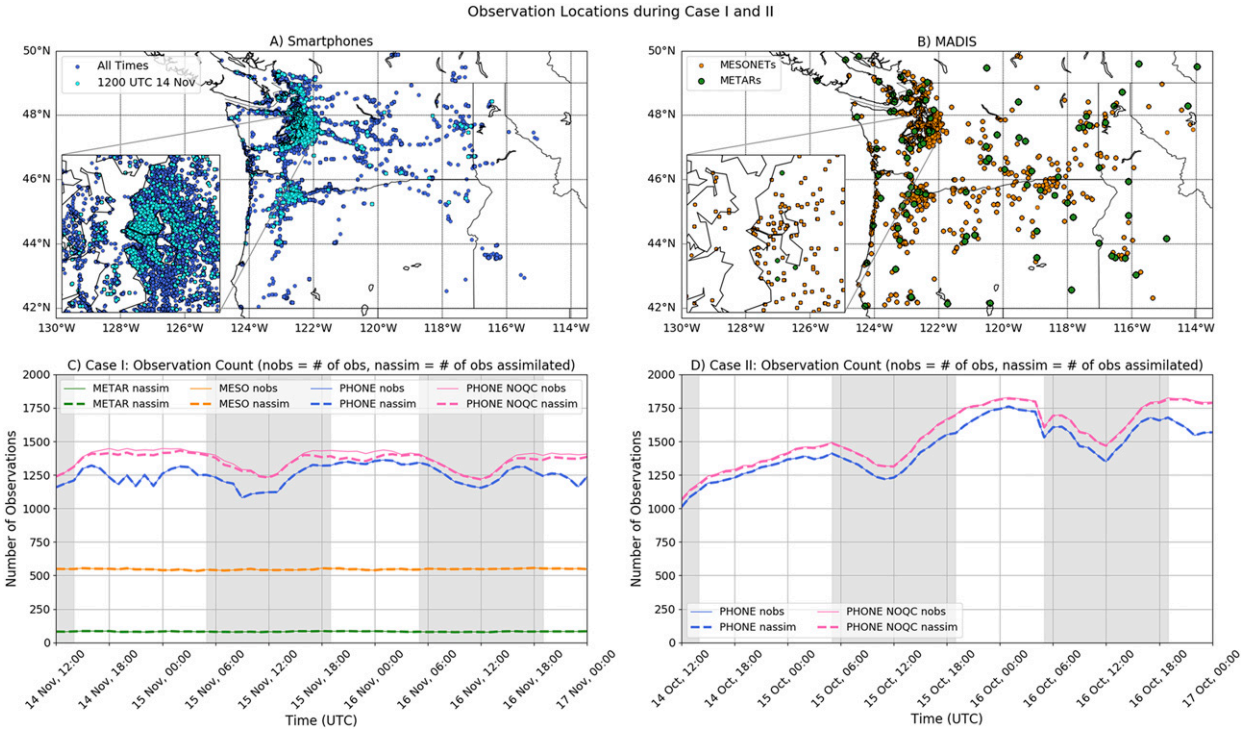

FIG. 5. (a) SPO locations during case I (1200 UTC 14 Nov–0000 UTC 17 Nov 2016) and during a single assimilation cycle (1200 UTC 14 Nov 2016). (b) MADIS mesonet and METAR altimeter observation locations. Inset plots depict observation locations in and around Seattle. The numbers of pressure observations retrieved (nobs) and assimilated (nassim), each hour, by observation type, in no-cycling DA experiments during (c) case I and (d) case II. Time series are displayed for both uncorrected (PHONE_NOQC) and corrected (PHONE) smartphone pressure observations.

Figures 6c–e display time series of 2-m temperature, 2-m dewpoint, and 10-m wind speed analysis error, relative to the prior (CNTRL) error. In both the PHONE and PHONE_NOQC experiments, SPOs generally provided no added benefit to the CNTRL analyses of 10-m wind speed, while a slight period-average improvement in 10-m wind analyses was observed when mesonet/METAR altimeter setting observations were assimilated. Assimilating corrected SPOs reduced the dewpoint and temperature analysis errors approximately 0.1 and 0.18 K, respectively. RMSE improvements from assimilating the mesonet altimeter setting were comparable to those achieved by assimilating corrected SPOs. There were improvements to temperature and dewpoint analyses when uncorrected SPOs were assimilated, and, assimilating uncorrected SPOs reduced temperature analysis errors to a greater degree than assimilating corrected SPOs.

In the CNTRL, the domain-average temperature bias was negative due to persistent 2–3-K (negative) temperature biases in the Columbia River basin, east of the Cascade Mountains (not shown). SPO assimilation in the PHONE experiment, and, to a greater degree, in the PHONE_NOQC experiment produced negative pressure increments. In the CNTRL, ensemble correlations between pressure and temperature were mostly negative. Consequently, negative pressure increments were associated with positive increments to the temperature field. Analysis increments in the PHONE_NOQC experiment were larger than in the PHONE experiment, as uncorrected SPOs deviated more from the CNTRL analysis and were more numerous/widespread than corrected SPOs. Accordingly, positive temperature increments in the PHONE_NOQC experiment helped offset negative temperature biases in CNTRL, to a greater degree than in the PHONE experiment. As a result, the 2-m temperature analysis RMSE was smaller in the PHONE_NOQC experiment than in the PHONE experiment.

b. Correlation length scale

Altimeter assimilation produced RMSE improvements of different magnitudes for each observed surface variable (Fig. 6). It was initially hypothesized that assimilating pressure should improve wind analyses, since pressure and wind are intimately related; however, this was not the case in the no-cycling DA experiments. To explain this lack of improvement in the wind statistics, the magnitude of the correlation coefficients between

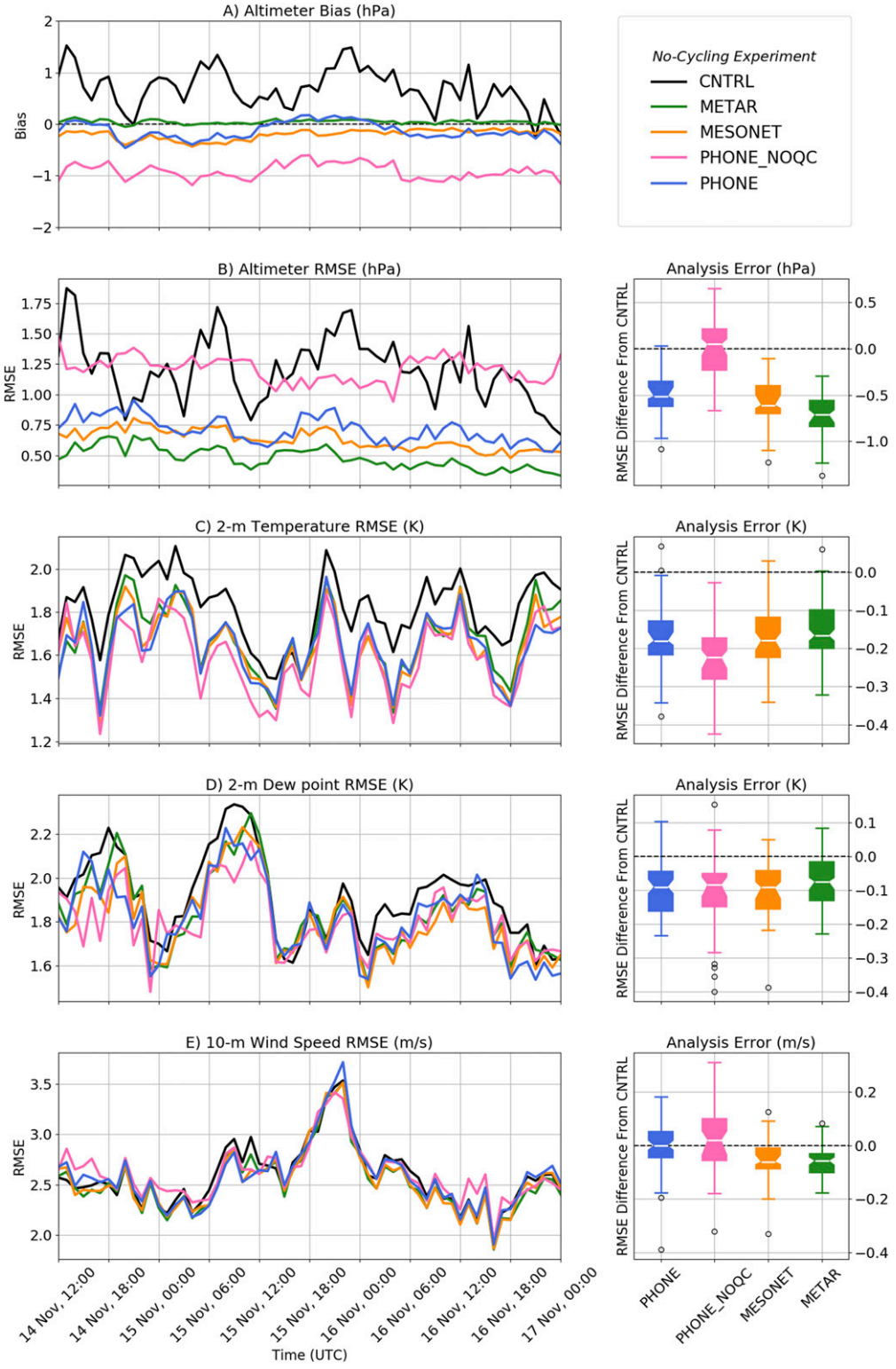

FIG. 6. (a) Altimeter setting bias and (b) RMSE of altimeter setting, (c) 2-m temperature, (d) 2-m dewpoint, and (e) 10-m wind speed in the no-cycling experiments. (left) Time series of RMSE/bias for the CNTRL (black), PHONE (blue), PHONE_NOQC (pink), MESONET (orange), and METAR (green) experiments. (right) Boxplots display the distribution of RMSE differences between the CNTRL 1-h forecast (prior) RMSE and the analysis (posterior) RMSE for each experiment. Notches in the boxplots represent the confidence interval (95%) for the median.

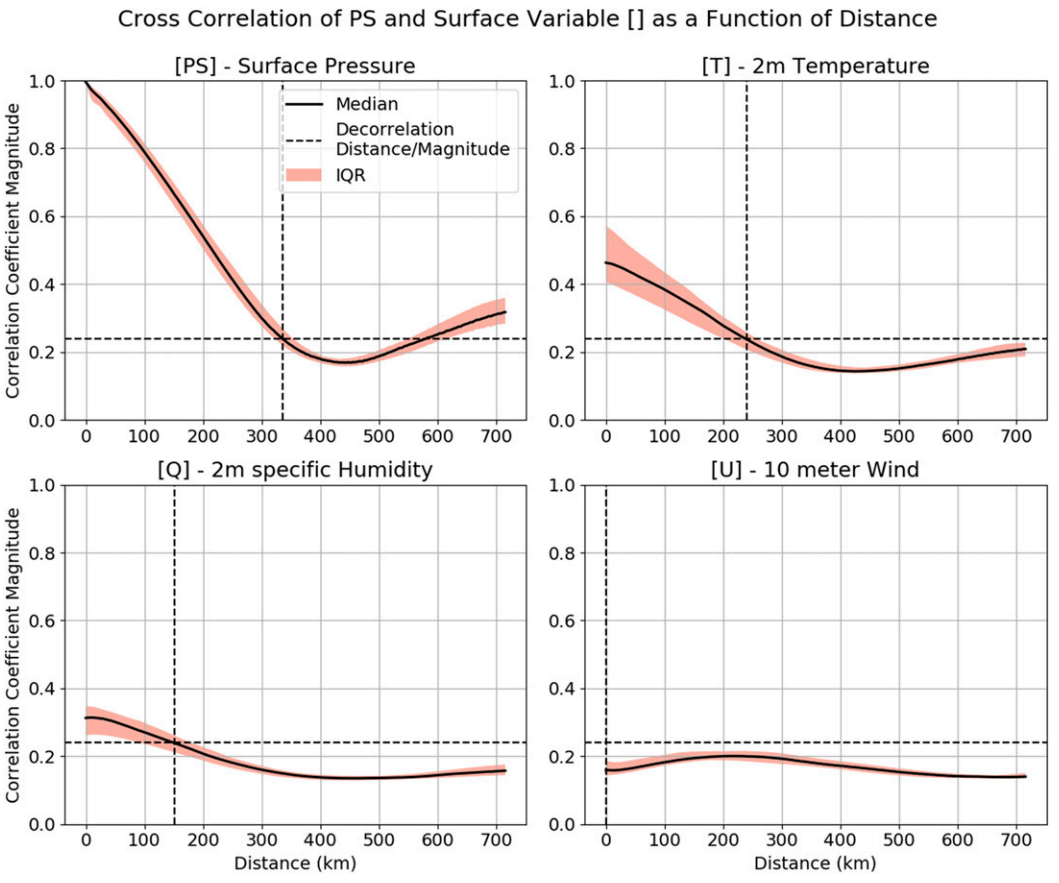

FIG. 7. Correlation coefficient magnitude of surface pressure (PS) with itself, 2-m temperature T, 2-m specific humidity Q, and the zonal 10-m wind U as a function of distance. These plots were generated from the CNTRL ensemble. Ensemble estimates of surface pressure at a single grid point X were correlated with ensemble estimates of each surface variable, at all grid points. Time averages were taken of the correlation coefficient magnitude at each grid point and the correlation coefficients were binned as a function of distance from grid point X. This process was repeated for all grid points, providing a distribution of cross-correlation magnitude, as a function of distance. The above plots display the interquartile range (IQR; shaded) and median (bold line) of this distribution. Dashed lines indicate the decorrelation magnitude, defined as the correlation coefficient magnitude below which correlations are, on average, not statistically significant. The distance at which correlations between PS and each surface variable are no longer significant is defined as the correlation length scale (decorrelation distance).

surface pressure and itself, 2-m temperature, 2-m specific humidity, and the zonal component of the 10-m wind was computed as a function of distance for each grid point in the CNTRL ensemble (Fig. 7). Figure 7 reveals that surface pressure is correlated with itself at distances of up to 320 km. This distance, defined as the correlation length scale for pressure, is in good agreement with the effective localization radius for SPOs noted in section 2. The second and third most closely correlated variables with surface pressure were 2-m temperature and 2-m specific humidity, respectively. The smaller the correlation magnitude, the smaller the covariance, and the smaller the analysis increments. Analysis error reductions were greater for 2-m temperature than 2-m dewpoint due to temperature's longer correlation length scale and larger correlation with surface pressure. Little to no improvement was observed for 10-m wind analyses in the no-cycling DA experiments since correlations between ensemble estimates of pressure and wind were minimal.

c. *Sensitivity experiments*

In previous research, a connection was found between the surface pressure observation density and analysis error (Anderson et al. 2005; Lei and Anderson 2014; Madaus et al. 2014). This relationship is tested here for corrected/uncorrected SPOs by assimilating varying sample sizes of SPOs over the duration of case I. At each assimilation step, specified numbers of SPOs were selected by random sampling without replacement, which

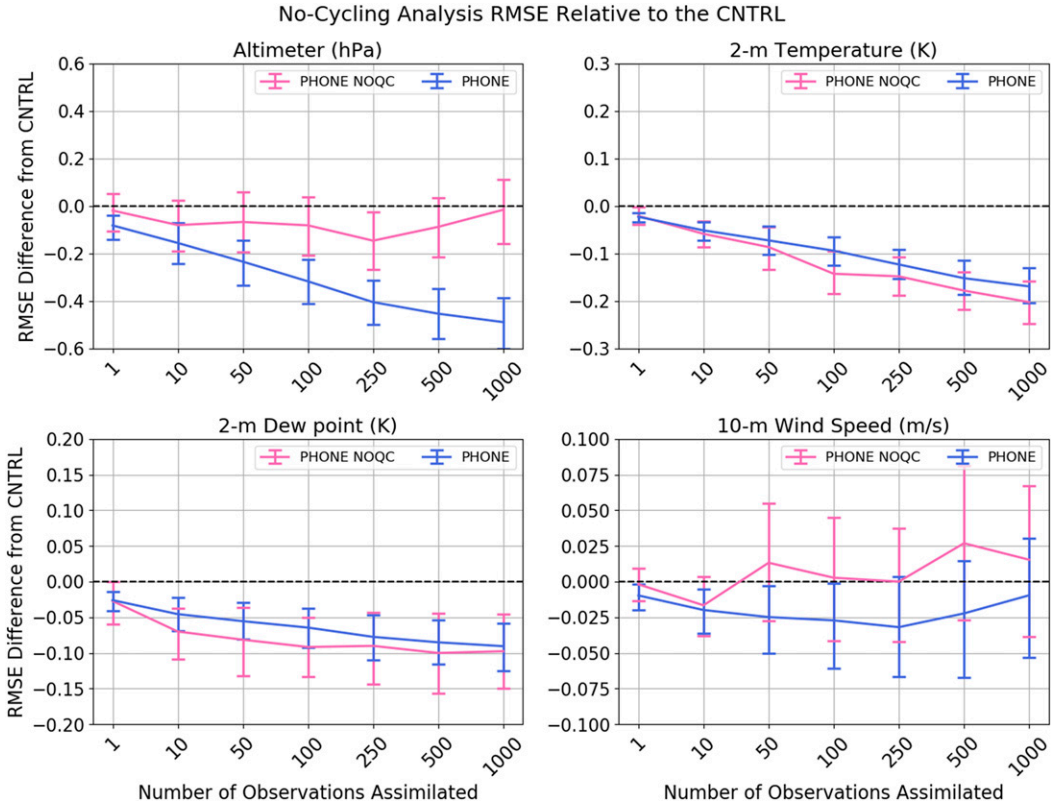

FIG. 8. Sensitivity of no-cycling experiments to the number of observations assimilated. The RMSEs of the altimeter setting, 2-m temperature, 2-m dewpoint, and 10-m wind speed are displayed for the PHONE and PHONE_NOQC experiments. These plots were generated by computing the mean domain-averaged RMSE for all 60 assimilation cycles. Error bars represent bootstrapped 95% confidence intervals for the mean RMSE difference between each experiment and CNTRL. In each sensitivity experiment SPOs of sample size N were selected, prior to each assimilation cycle, by randomly sampling without replacement.

ensured that SPOs from the same smartphone were not necessarily assimilated every hour.

Figure 8 displays the results of this sensitivity experiment. Assimilating corrected SPOs resulted in a monotonic decrease of the analysis altimeter, 2-m temperature, and 2-m dewpoint RMSEs relative to the prior (CNTRL) as the number of corrected SPOs assimilated was increased. A similar reduction in 2-m temperature and 2-m dewpoint RMSE was observed when the number of uncorrected SPOs assimilated was increased. Decreases in analysis RMSE of each variable in the PHONE experiment were consistent with the correlation length scale between the variable and surface pressure (Fig. 7). Surface variables more correlated with surface pressure and with longer correlation length scales exhibited larger reductions in analysis RMSE. In the PHONE experiment the largest reductions were observed for altimeter setting, followed by 2-m temperature and 2-m dewpoint. In both the PHONE and PHONE_NOQC experiments, wind analysis RMSE was independent of the number of observations assimilated since the sample covariance between the wind and pressure was, on average, minimal.

d. *Cycling experiments*

To evaluate the cumulative impact of SPO assimilation, four cycling DA experiments were performed with corrected SPOs (PHONE), uncorrected SPOs (PHONE_NOQC), mesonet altimeter observations (MESONET), and METAR altimeter observations. For all cycling experiments, 1-h forecast errors were computed by subtracting METAR observations from the prior ensemble-mean 1-h forecast at the location of each METAR observation. A domain-averaged 1-h forecast RMSE was computed at each assimilation step for the ensemble mean altimeter setting, 2-m temperature, 2-m dewpoint, and 10-m wind speed.

Time series of the domain-averaged 1-h altimeter forecast RMSE for each cycling experiment are displayed in Fig. 9a, with period-averaged differences in RMSE between the CNTRL 1-h forecast and the four DA experiments displayed in the right panel. Assimilating corrected

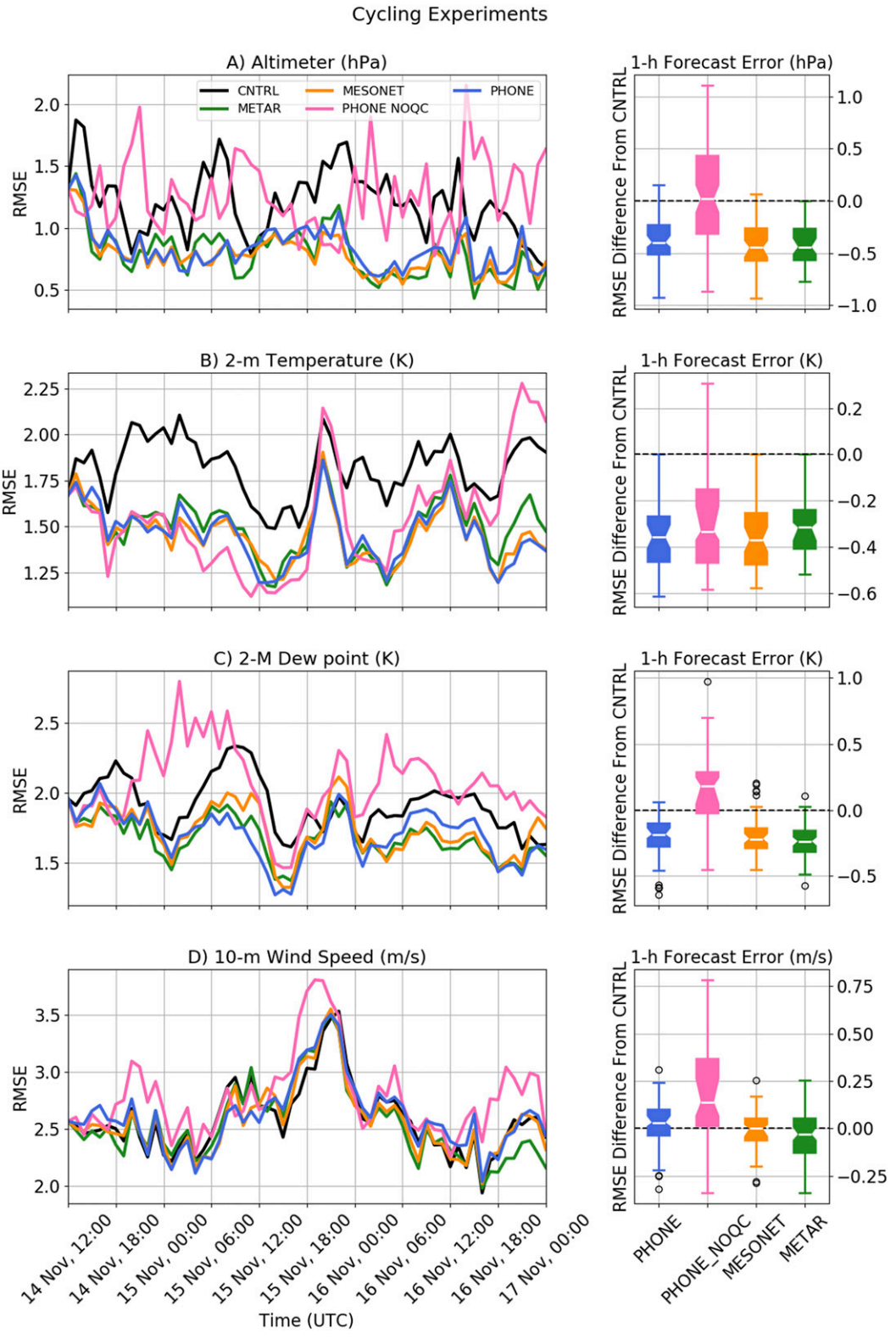

FIG. 9. RMSEs of (a) altimeter setting, (b) 2-m temperature, (c) 2-m dewpoint, and (d) 10-m wind speed in the cycling experiments. (left) As in Fig. 6, time series of RMSEs for the CNTRL (black), PHONE (blue), PHONE_NOQC (pink), MESONET (orange), and METAR (green) experiments. (right) Boxplots display the distribution of RMSE differences between the 1-h forecast (prior) RMSEs and 1-h forecasts from each experiment.

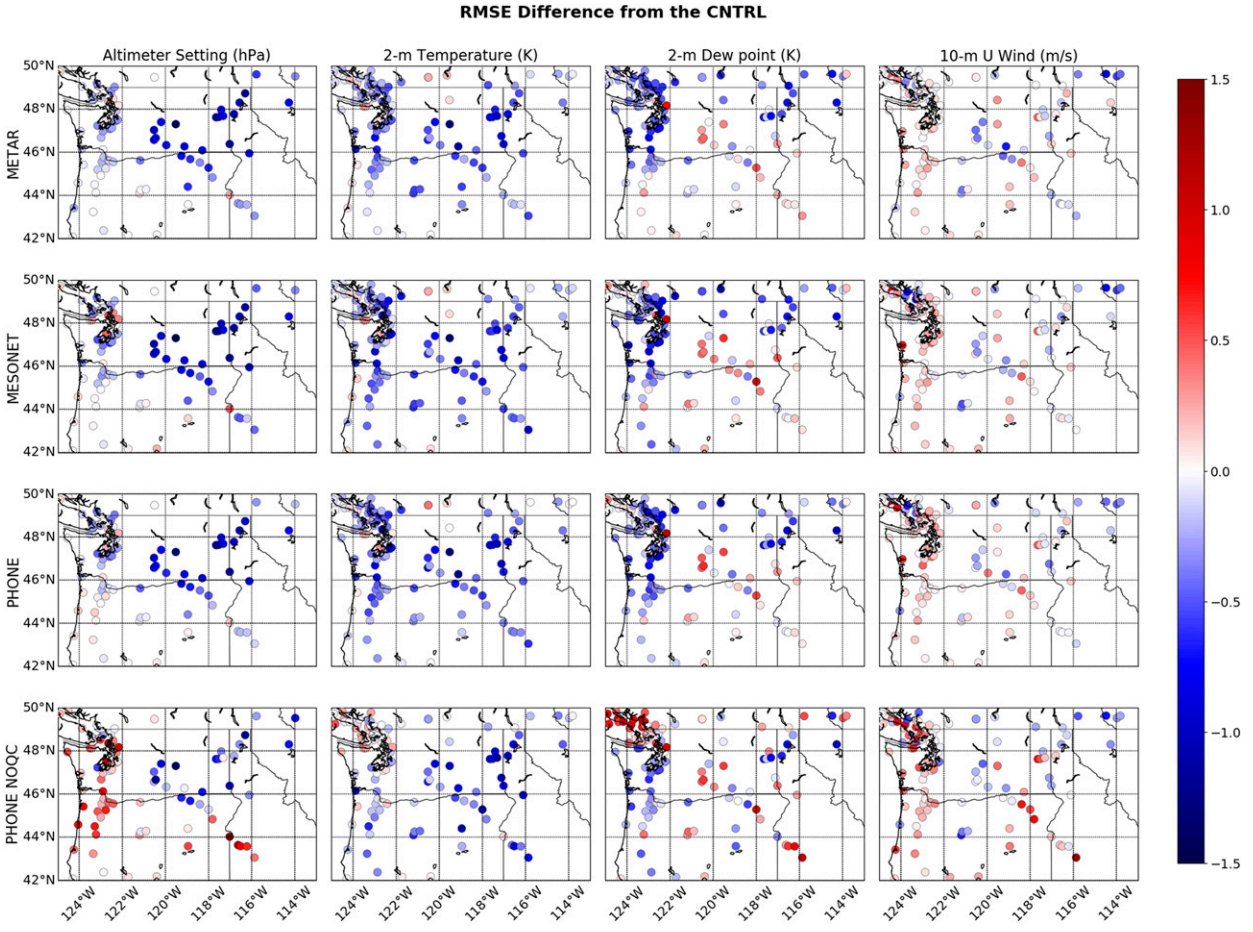

FIG. 10. Time-averaged spatial distribution of 1-h forecast RMSEs, relative to CNTRL, for the METAR, MESO, PHONE, and PHONE_NOQC cycling experiments. RMSE differences for each cycling experiment are displayed in rows, with each column showing the RMSE difference for a given surface variable. At each verification site, the 1-h forecast RMSE is computed for the altimeter setting, 2-m temperature, 2-m dewpoint, and 10-m u wind over the duration of case I. This calculation is performed for all DA experiments and the CNTRL experiments.

SPOs consistently reduced 1-h forecast altimeter RMSEs, with a median RMSE reduction of 0.4 hPa. This reduction in RMSE was not significantly different than the reduction of 1-h forecast altimeter RMSEs observed in the MESONET and METAR experiments. Assimilating uncorrected SPOs provided no benefit to 1-h forecasts of altimeter setting.

Figures 9b–d display the domain-averaged 1-h forecast RMSEs for 2-m temperature, 2-m dewpoint, and 10-m wind, as well as the CNTRL ensemble-mean 1-h forecast RMSE. Assimilating corrected SPOs consistently improved the 1-h temperature forecasts and, to a lesser degree, the 1-h dewpoint forecasts. This result is expected since pressure is more strongly correlated with 2-m temperature than dewpoint. On average, corrected SPOs slightly degraded the performance of 1-h wind forecasts. When uncorrected SPOs were assimilated, reductions in 2-m temperature were observed but were not sustained. Large increases in 2-m temperature RMSE were observed in the PHONE_NOQC experiment toward the end of the period. In contrast to the no-cycling experiments, the assimilation of uncorrected SPOs degraded the 1-h forecasts of 2-m dewpoint and 10-m wind speed.

To examine the time-averaged spatial distribution of the forecast error, the 1-h forecast RMSE was computed at each METAR verification site over case I, for all cycling DA experiments and the CNTRL experiment. The results are displayed in Fig. 10, which shows the 1-h forecast RMSE difference between each assimilation experiment and the CNTRL experiment, at all verification sites, for surface variables altimeter setting, 2-m temperature, 2-m dewpoint, and the 10-m zonal u-wind component. Figure 10 reveals that reductions in the 1-h forecast RMSE for altimeter setting were widespread in all experiments except for the PHONE_NOQC

experiment, in which 1-h forecasts of altimeter setting were degraded in western Washington and southeastern Oregon. Improvements to 1-h forecasts of 2-m temperature were observed throughout the domain in all cycling experiments; however, in the PHONE_NOQC experiment the 2-m temperature forecasts were only marginally improved in western Washington. In the PHONE_NOQC experiment the 1-h forecast RMSE for 2-m dewpoint was increased, relative to the CNTRL, over northwest Washington and Vancouver Island, Canada. In this region the assimilation of uncorrected SPOs produced large negative (positive) pressure (temperature) increments resulting in anomalously warm and dry conditions at the surface. In the METAR, MESONET, and PHONE experiments the 1-h dewpoint forecast RMSE was markedly reduced, relative to the CNTRL, throughout western Washington, where most assimilated observations were located. A slight increase in the *u*-wind 1-h forecast RMSE, relative to CNTRL, was observed across western Washington in all DA experiments. Since pressure and wind were poorly correlated in this ensemble, wind analysis increments were prone to spuriousness. This was particularly true in western Washington, where most observations were assimilated, and the analysis increments were largest.

e. Precipitation skill

During case I, the surface low passage was associated with both frontal and postfrontal precipitation. To evaluate the impacts of SPO assimilation on precipitation forecasts, fractions skill scores (FSSs) were computed for 1-h ensemble precipitation forecasts for $\geq$1 mm (see the appendix for details). Gridded observations from NCEP Stage IV 1-h precipitation accumulation analyses were used to compute the FSS at a variety of spatial scales (Fig. 11a). On average, the FSS remained below the ''useful'' skill threshold of 0.5 suggested by Roberts and Lean (2008). Nevertheless, assimilation of corrected SPOs and, to a lesser degree, uncorrected SPOs improved the time-averaged FSS relative to CNTRL (Fig. 11b). There were several times when the FSS of the 1-h precipitation forecasts in the PHONE and PHONE_NOQC exceeded the useful skill threshold when CNTRL did not (Fig. 11c). A notable example of this is at 1600 UTC 16 November when the FSS in the PHONE experiment peaked during a postfrontal period characterized by a decline in FSS in CNTRL. This peak in the time series of FSS for the PHONE experiment was observed for all neighborhoods (not shown).

Figure 12 compares the fractional coverage of gridded precipitation from the Stage IV precipitation analyses and the fractional coverage of the ensemble members that met/exceeded the forecast precipitation threshold of 1 mm within a 68-km neighborhood at 1600 UTC 16 November. Figure 12 reveals that CNTRL failed to capture postfrontal precipitation while the PHONE_NOQC experiment overforecast precipitation. This is not surprising as the assimilation of uncorrected SPOs introduced a systematic low pressure bias that promoted precipitation. In contrast, the assimilation of corrected SPOs in the PHONE experiment resulted in a more skillful mesoscale 1-h precipitation forecast. In the PHONE experiment, SPO assimilation reduced the pressure just offshore of the Oregon coast. This reduction in pressure, relative to CNTRL, encouraged the development of shallow convection along the Oregon coast that produced a more realistic distribution of precipitation in the PHONE experiment.

f. Free forecasts

To examine the impact of SPOs on forecasts at longer lead times, 11 free-forecast runs were performed during case I. The 0–6-h free forecasts were initialized with analyses from the cycled PHONE, PHONE_NOQC, and CNTRL ensemble every 6 h from 1200 UTC 14 November to 0000 UTC 17 November 2016. The RMSE of ensemble mean forecasts from all 11 runs was computed for mean sea level pressure, 2-m temperature, 2-m dewpoint, and 10-m wind speed as a function of forecast lead time (Fig. 13). The assimilation of corrected SPOs improved the 2-m temperature and 2-m dewpoint RMSEs at forecast lead times up to 6 h, while MSLP forecasts were improved at 3–5-h lead times. When uncorrected SPOs were assimilated, MSLP forecasts were degraded relative to CNTRL. The assimilation of uncorrected SPOs degraded the 2-m dewpoint forecasts at short lead times and reduced the 2-m temperature RMSE at all forecast lead times. In both the PHONE and PHONE_NOQC free-forecast experiments, significant improvements to 10-m wind speed forecasts were not observed.

6. Case II: Data assimilation and forecast results

The second case represents a very different synoptic/mesoscale evolution from case I, with an intense, compact midlatitude cyclone moving northward just offshore of the Pacific coast, with substantial errors in track and intensity in the operational forecasts.

a. Cycling experiments

Figure 14a displays the domain-average altimeter bias and RMSE for cycling experiments assimilating corrected

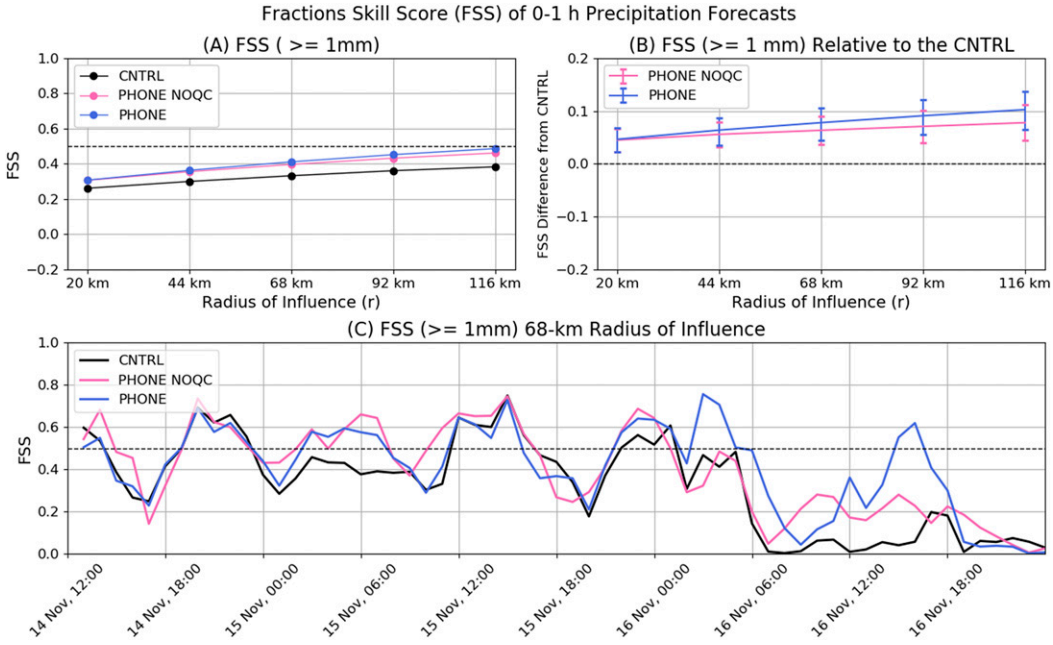

FIG. 11. (a) FSSs for 0–1-h ensemble precipitation forecasts from the CNTRL, PHONE, and PHONE_NOQC cycling experiments. Here, the FSS is determined by the fraction of ensemble members that forecast ≥ 1 mm of precipitation at each grid point or within a radius of influence r of that grid point. (b) FSSs for the PHONE and PHONE_NOQC experiments relative to FSSs from CNTRL. Error bars represent bootstrapped 95% confidence intervals for the difference between the FSSs of each experiment and CNTRL. (c) Time plot of FSS, at a spatial scale of 68 km, for the CNTRL, PHONE, and PHONE_NOQC cycling experiments.

(PHONE) and uncorrected (PHONE_NOQC) SPOs. Period-averaged differences in RMSE between CNTRL and the two DA experiments are displayed in the right panel. For the period average, assimilating uncorrected (corrected) SPOs degraded (improved) the 1-h forecasts of altimeter setting. The time plot of altimeter RMSE reveals that between 0000 and 0600 UTC 16 October the 1-h forecast altimeter RMSE was substantially

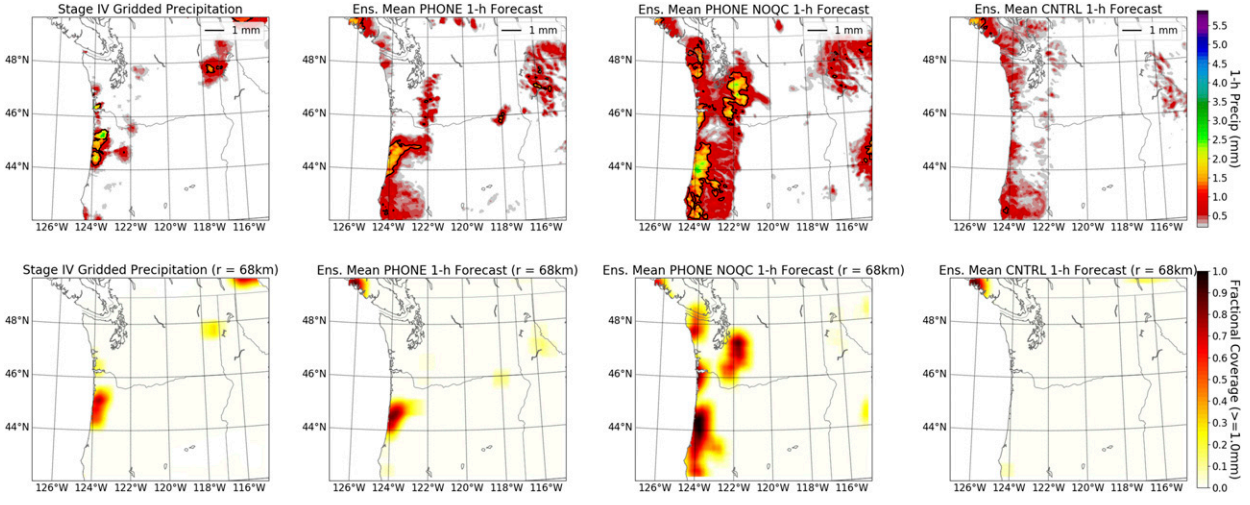

FIG. 12. (top) Stage IV gridded precipitation analysis and 0–1-h accumulated precipitation forecasts from the PHONE, PHONE_NOQC, and CNTRL experiments, valid at 1600 UTC 16 Nov 2016. (bottom) Fractional coverage fields for precipitation analysis and 1-h forecasts. For the precipitation analysis the fractional coverage represents the fraction of analysis grids within 68 km of a grid point that meet the ≥1 mm precipitation threshold. For the 0–1-h ensemble precipitation forecasts, fractional coverage represents the fraction of ensemble members that exceed the precipitation threshold within 68 km of a grid point.

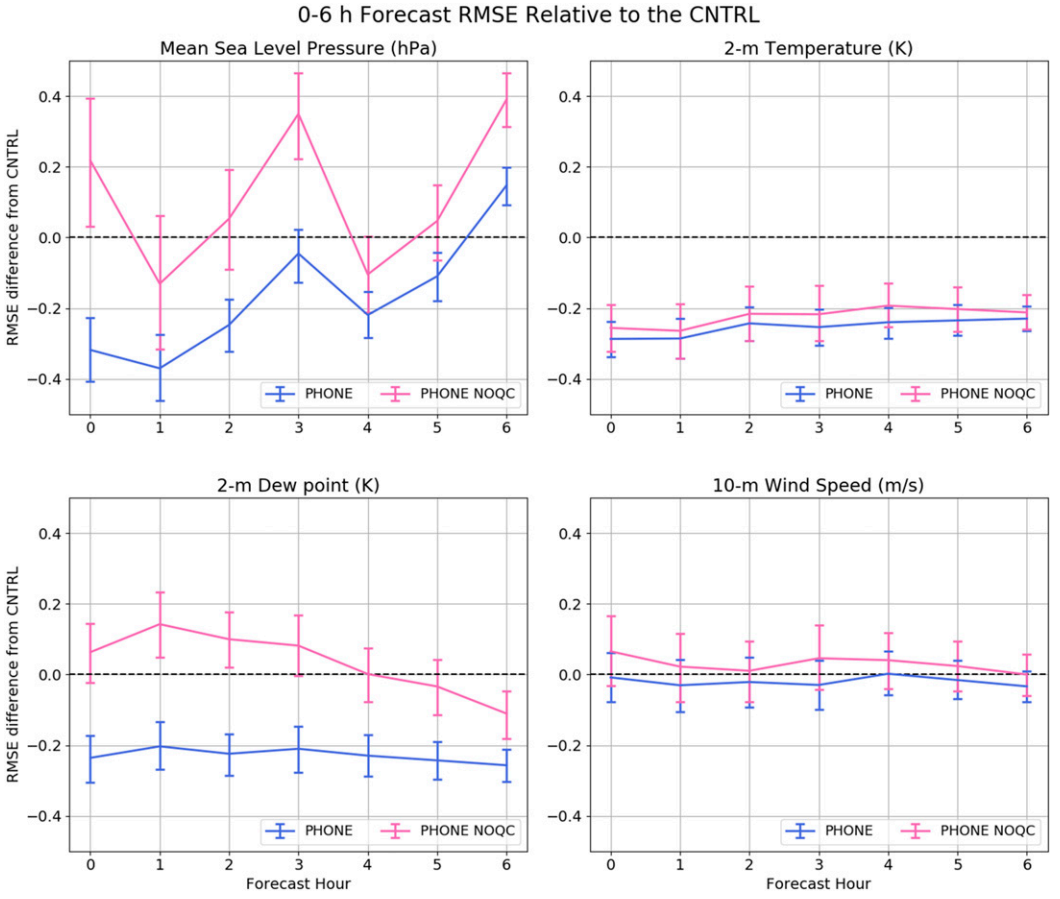

FIG. 13. RMSE as a function of forecast hour for 0–6-h free forecasts of MSLP, 2-m temperature, 2-m dewpoint, and 10-m wind speed. These plots were generated from 11 free forecasts initialized every 6 h starting at 1200 UTC 14 Nov and ending at 0000 UTC 17 Nov 2016. At each forecast hour a domain-averaged RMSE was computed from all 11 runs. The error bars in each plot depict bootstrapped 95% confidence intervals for the RMSE difference between each free forecast initialized from the PHONE and PHONE_NOQC cycling experiments and free forecasts initialized from CNTRL.

reduced, relative to CNTRL, in both the PHONE and PHONE_NOQC experiments. During this period, the surface low approached and made landfall on Vancouver Island. Uncorrected SPO errors were less than those in CNTRL. As in case I, uncorrected SPOs contributed to a low bias in 1-h altimeter forecasts. Domain-averaged 1-h altimeter forecasts were low biased in both the CNTRL and PHONE_NOQC experiments during the period when the low made landfall. Assimilation of corrected SPOs slightly overcorrected the domain-average low pressure bias during this period.

Figure 14b displays the domain-average 10-m wind speed bias and RMSE for the CNTRL, PHONE, and PHONE_NOQC cycling experiments. In this case, wind forecast errors were dominated by errors in the track of the surface low. In CNTRL, 10-m wind speeds were overforecast during the time when the low made landfall. This positive bias was mostly corrected in the PHONE_NOQC and PHONE experiments. In the period average, assimilating SPOs provided no added benefit to domain-averaged wind forecasts.

In this case, errors in the forecast track contributed to poor wind forecasts. Figures 15a and 15b display the forecast intensity and track of the surface low in analyses from the CNTRL, PHONE, and PHONE_NOQC cycling experiments. The analyzed intensity and track from the NOAA HRRR system is also plotted as an estimate of truth. Later in the period, when the surface low entered the MADIS maritime (buoy) and METAR observing networks, the minimum observed MSLP was plotted. Since the surface low did not pass directly over observing sites, this estimate can be considered a lower limit on the storm intensity.

Early in the period, prior to 2200 UTC, SPO assimilation had little impact on the analyzed track and intensity of the surface low as the surface low remained far

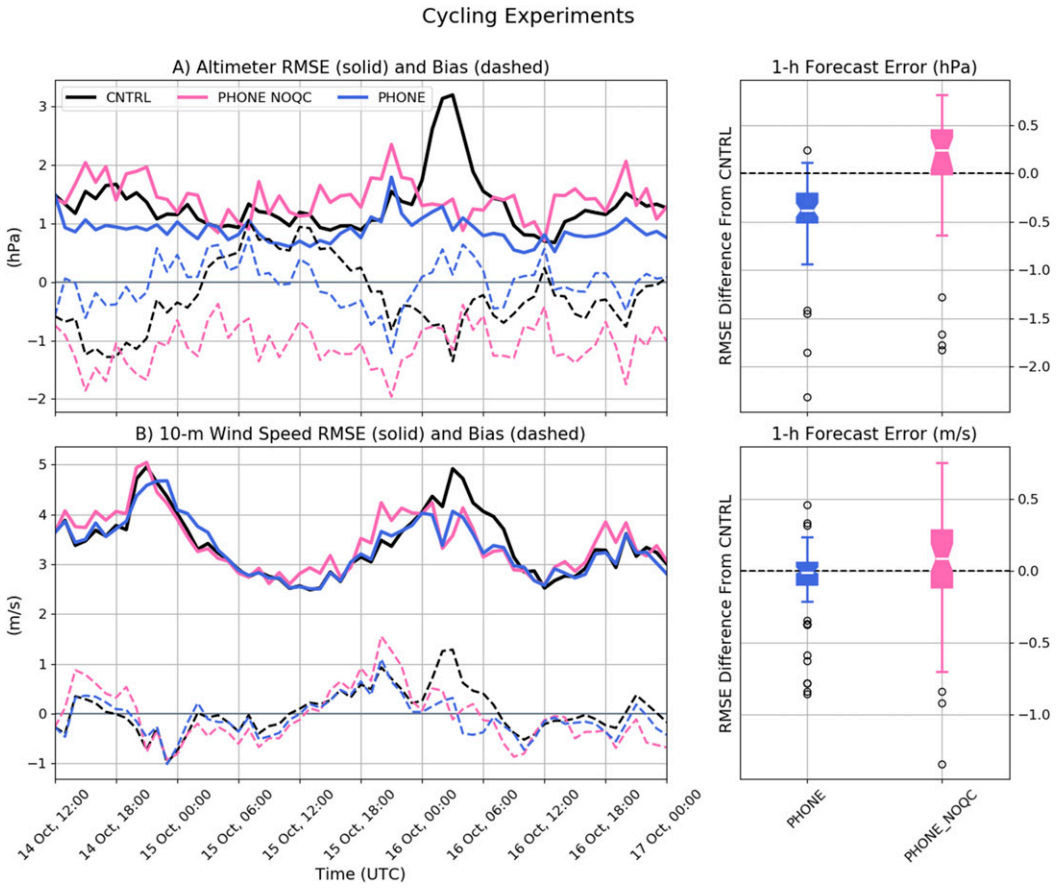

FIG. 14. RMSE and bias of (a) altimeter setting and (b) 10-m wind speed during case II cycling experiments. (left) Time series of RMSE and bias for the CNTRL (black), PHONE (blue), and PHONE_NOQC (pink) experiments. (right) Boxplots display the distribution of RMSE differences between the 1-h forecast (prior) RMSE and 1-h forecasts from each experiment.

offshore. At 2200 UTC, the analyzed forecast track in the PHONE_NOQC experiment shifted to the northwest of the CNTRL track, in better agreement with the HRRR analysis. A similar northwestward shift in the analyzed surface low position was observed an hour later (2300 UTC) in the PHONE experiment. At this time the magnitude of the analysis increments near the surface low increased substantially (not shown), since prior to this time the distance between the surface low and Seattle, where the majority of the SPOs were located, was less than the effective localization radius for SPOs.

In the PHONE_NOQC experiment the northwestward shift in the forecast track was observed an hour earlier because more SPOs were assimilated in this experiment than in the PHONE experiment. In this special case, model errors exceeded the average magnitude of the uncorrected SPO error. For this reason, the quality of observations was of less importance than the quantity, especially along the sparsely observed coastline. The cumulative impact of assimilating coastal SPOs at locations unobserved in the PHONE experiment facilitated an earlier shift in the storm track in PHONE_NOQC experiments by extending the analysis increments farther offshore than in the PHONE experiment. While the timing of the track shift differed in each SPO experiment, the location of landfall was the same in both experiments and in better agreement with the HRRR analysis than CNTRL. Likewise, surface low intensity analyses in the PHONE and PHONE_NOQC experiments were closer to the HRRR analysis and minimum observed MSLP than CNTRL. While not shown in Fig. 15, similar improvements to analyses were retained in 1-h cycled forecasts of the surface low intensity and position.

b. *Free forecasts*

To evaluate how SPO assimilation impacted forecasts of the surface low track and intensity at longer lead times, free forecasts were initialized at 2300 UTC 15 October from the cycled CNTRL, PHONE_NOQC,

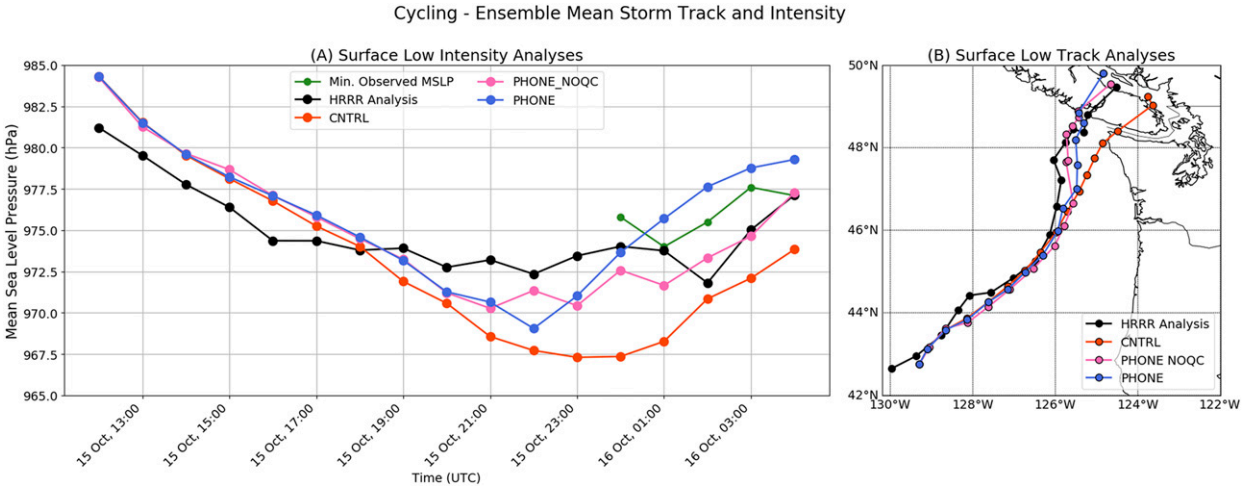

FIG. 15. (a) Surface low intensity and (b) track ensemble mean analyses in the HRRR, CNTRL, and cycling experiments PHONE and PHONE_NOQC. The minimum observed MSLP is plotted at the end of the period when the surface low was within the MADIS maritime (buoy) and METAR observing networks.

and PHONE ensembles. This initialization time was chosen because at this time the low positions in both SPO experiments had deviated from the position of the low in CNTRL. In case I, free-forecast experiments showed MSLP forecast improvements at lead times up to 5 h. For this reason, 0–5 h free forecasts were evaluated in case II.

Figures 16a and 16b display 0–5-h forecasts, initialized at 2300 UTC, of the surface low intensity and position for the HRRR, CNTRL, PHONE_NOQC, and PHONE experiments. In the PHONE and PHONE_NOQC experiments, the surface low intensity was closer to the HRRR analysis and minimum observed MSLP than in CNTRL. In both SPO experiments, initial improvements to the surface low intensity were retained at all forecast lead times. Similarly, improvements in the track of the surface low were observed at all forecast lead times in the PHONE and PHONE_NOQC experiments (Fig. 16b). At forecast lead times of 2–5 h, the surface low track in the PHONE_NOQC experiment overlapped with the surface HRRR analyzed track. In the PHONE experiment, the surface low tracked parallel to the HRRR analysis as the low approached land, making landfall approximately 25 km east of the analyzed HRRR track. In CNTRL the surface low tracked approximately 100 km east of the HRRR-analyzed track, making landfall on the Olympic Peninsula before crossing the Strait of Juan de Fuca.

c. *Wind forecast analysis*

In case II, the intensity and position of the surface low impacted the distribution and strength of near-surface winds along the western Washington coast and the interior. To evaluate the performance of ensemble near-surface wind and wind gust forecasts, the Brier skill score (BSS) was employed (see the appendix for details). Using the CNTRL ensemble as a reference forecast, the BSS was calculated for SPO cycling and free-forecast experiments. MADIS maritime and METAR near-surface wind observations were used for verification. Time plots of BSS for probabilistic forecasts of 10-m wind speed exceeding $10\,m\,s^{-1}$ and surface wind gusts exceeding gale force ($17.2\,m\,s^{-1}$) are shown in Fig. 17. The assimilation of uncorrected/corrected SPOs resulted in more skillful 10-m wind and surface gust forecasts from 0100 to 0400 UTC 16 October 2016. By this time, the surface low intensity had decreased, and the low position had shifted northwest relative to CNTRL. Free forecasts initialized with analyses from the PHONE and PHONE_NOQC cycling experiments produced more skillful 10-m wind speed and surface gust forecasts than CNTRL at 2–5-h forecast lead times. Improvements in near-surface wind forecast skill were greatest in the PHONE_NOQC free-forecast experiment, as in this experiment the surface low track was farthest from the CNTRL track and closest to the analyzed HRRR track. In both SPO forecast experiments, improvements to the initial surface low intensity and position were retained at forecast lead times up to 5 h, facilitating more skillful wind forecasts at equivalent forecast lead times.

7. Conclusions

This paper examines the impact of smartphone pressure observations (SPOs) for two events: the first involving the passage of a trough and associated cold front

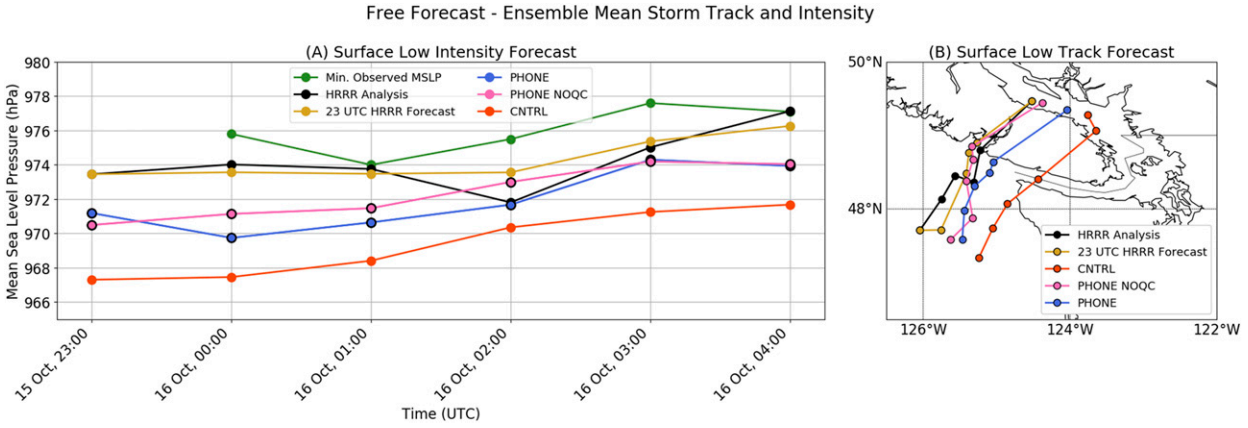

FIG. 16. (a) Surface low intensity and (b) track forecasts for ensemble mean free forecasts initialized at 2300 UTC 15 Oct 2016. HRRR intensity and track analyses are displayed for reference. The minimum observed MSLP is plotted at the end of the period when the surface low was within the MADIS maritime (buoy) and METAR observing networks. Two-tailed t tests were applied to test the difference between the ensemble mean surface low intensity in the PHONE_NOQC and PHONE experiments and CNTRL. In the intensity plot, bold markers indicate analyses/forecasts from these experiments whose p value was <0.05. The forecast position of the surface low is plotted in (b).

and the second associated with the landfall of an intense low pressure center. For each case, there is an evaluation of the impacts of advanced quality control strategies and machine learning for the bias correction of smartphone pressure observations. In addition, the impact of improved quality control/bias correction of smartphone pressure observations on forecast skill is examined, building on previous work (McNicholas and Mass 2018).

In case I, a surface low/trough traversed the Puget Sound region, where SPO density was greatest. During this

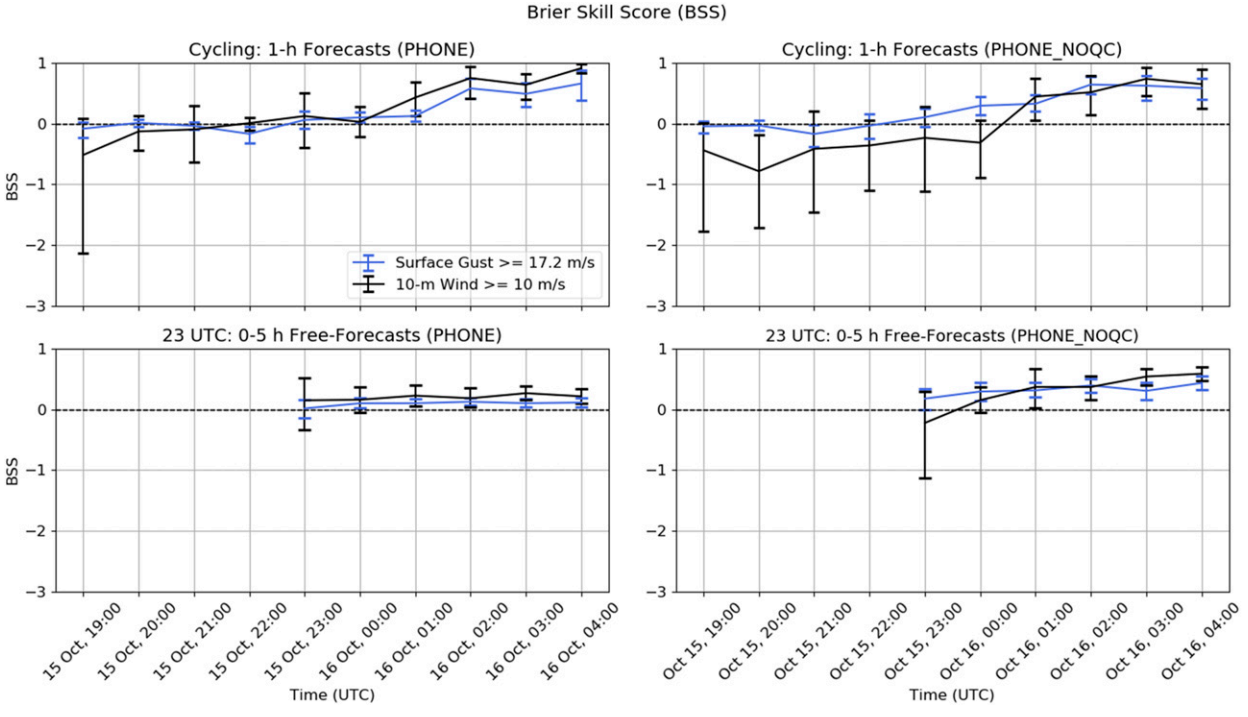

FIG. 17. BSS for surface wind gusts exceeding gale force (17.2 m s^{-1}) and 10-m wind speed exceeding (10 m s^{-1}). (top) BSSs for cycling experiments PHONE and PHONE_NOQC. (bottom) BSSs for 0–5-h free forecasts initialized from the PHONE and PHONE_NOQC cycled ensembles. BSSs were computed using the CNTRL ensemble as a reference forecast. Positive BSSs indicate improvements in skill, relative to CNTRL, and vice versa.

event, corrected SPOs consistently reduced the analysis error of the altimeter setting, 2-m temperature, and 2-m dewpoint. Reductions in 1-h forecast errors for these surface variables were achieved when corrected SPOs were assimilated in the cycling mode. Such reductions in RMSE were consistent in time and space. Compared to experiments that assimilated pressures from traditional mesonets, the assimilation of corrected SPOs resulted in nearly equivalent reductions in domain-averaged altimeter and temperature forecast/analysis errors. Likewise, the spatial distribution of forecast improvements was markedly similar in experiments assimilating corrected SPOs and pressures from traditional mesonets and METARs.

The assimilation of uncorrected SPOs did not improve the altimeter analyses and 1-h forecasts in case I; however, uncorrected SPOs were able to improve analysis/forecasts of 2-m temperature. This unintuitive result was the consequence of a cancellation of biases, wherein negatively biased smartphone pressures induced positive temperature increments that reversed a systematic negative temperature bias in the control experiment. In no-cycling/cycling DA experiments, SPOs did not improve wind analyses/forecasts, a result reflecting the lack of correlation between ensemble estimates of pressure and wind. The magnitude of the analysis and forecast error reductions, achieved by assimilating corrected SPOs, was directly proportional to the number of observations assimilated and the magnitude of the correlation of surface pressure with the surface variable evaluated.

In case I, both corrected and, to a lesser degree, uncorrected SPOs improved 1-h forecast precipitation skill relative to the control simulation without smartphone observations. Improvements in the fractions skill score were most notable during the postfrontal period when the assimilation of corrected SPOs improved mesoscale forecasts of postfrontal convective precipitation along the Oregon coast. Free-forecast experiments showed that assimilating corrected SPOs resulted in a significant reduction in forecast RMSEs for altimeter setting, 2-m temperature, and 2-m dewpoint at forecast lead times of 3–6 h.

Case II considered a storm poorly forecast by operational systems. The assimilation of both corrected and uncorrected SPOs significantly improved altimeter and 10-m wind forecasts during the period of storm landfall. In this case, SPO quality had little impact on forecast performance since errors in uncorrected SPOs were dwarfed by the magnitude of the pressure errors in the control ensemble. In cycled SPO assimilation experiments, errors in the analyzed track and intensity of the windstorm were markedly reduced as the storm approached landfall. Free-forecast experiments demonstrated that such reductions in model analysis errors were associated with improvements in the forecast track and intensity of the windstorm at short lead times. In both cycling and free-forecast SPO experiments, improvements in the forecast storm track resulted in commensurate improvements to probabilistic near-surface wind forecasts.

In the region used in these experiments (the Pacific Northwest), there are likely over a million smartphones capable of retrieving pressure. This would imply that in the experiments discussed above less than 0.1% of potential SPOs were assimilated. In this study, sensitivity experiments revealed that domain-averaged analysis error, relative to the control, decreased monotonically as the number of assimilated smartphone observations was increased. Since just over a thousand hourly SPOs performed similarly to existing mesoscale pressure networks in constraining forecasts of pressure, temperature, and dewpoint, it is plausible that greater reductions in analysis/forecast error are possible if a considerably denser network was available. MM2018 showed that such a network is feasible by demonstrating that smartphone pressures can be efficiently collected and bias corrected at subhourly intervals. This study confirms the methodology of MM2018 and suggests that crowdsourced smartphone pressures can enhance operational numerical weather prediction.

Acknowledgments. The authors would like to acknowledge uWx users, whose cooperation made this research possible, the Weather Company (IBM) for their generous financial support of this research, and Microsoft for providing cloud-computing time and resources. Support was also provided by a grant from the NOAA CSTAR program through Grant NA10OAR4320148AM63.

APPENDIX

Verification Methods

In this study, ensemble forecasts were evaluated using the National Center for Atmospheric Research (NCAR) Model Evaluation Toolkit (MET; Fowler et al. 2018). MET was used to calculate several verification metrics for ensemble mean forecasts. The first metric, mean error (bias), was computed as the domain-average difference between the ensemble mean forecast f_i and verifying observation o_i at each observation location:

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i). \tag{A1}$$

The second metric, RMSE, was computed as an average over the model domain:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2}. \quad \text{(A2)}$$

In all DA experiments, RMSE and bias was evaluated with high quality METAR observations.

Ensemble probabilistic forecast skill was evaluated using the Brier score (BS; Brier 1950). The BS is analogous to the mean squared error for probabilistic forecasts:

$$\mathrm{BS} = \frac{1}{n}\sum_{t=0}^{n}(F_t - o_t)^2. \quad \text{(A3)}$$

In the Brier score, F_t represents the fraction of ensemble members that forecast an event to occur at time t, while o_t defines whether an event was observed to occur at time t. In this study, the BS is used to calculate the Brier skill score as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{FCST}}}{\mathrm{BS}_{\mathrm{CNTRL}}}, \quad \text{(A4)}$$

and the CNTRL ensemble is used as the reference forecast. When the BSS is negative (positive), the ensemble FCST is less (more) skillful than CNTRL.

To evaluate the performance of ensemble forecasts across spatial scales, the fractions Brier score (FBS; Roberts 2005) is used. The FBS is defined as

$$\mathrm{FBS} = \frac{1}{N}\sum_{i=1}^{N}(\langle P_f\rangle_i - \langle P_o\rangle_i)^2, \quad \text{(A5)}$$

where N is the number of neighborhoods. Neighborhoods N are defined using a radius of influence r. At each grid point i, a neighborhood is defined as a square grid of all grid points within r kilometers of i. In Eq. (A5), $\langle P_f\rangle_i$ represents the fraction of grid points (i.e., fractional coverage) of a binary metric (e.g., precipitation accumulation ≥ 1 mm) within a forecast neighborhood, at each grid point i. Likewise, $\langle P_o\rangle_i$ is the fractional coverage of a binary metric within an observed neighborhood, at each grid point i. In this study, the FBS is used within the context of the fractions skill score (FSS; Roberts and Lean 2008). The FSS is calculated as

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\sum_{i=1}^{N}\langle P_f\rangle_i^2 + \sum_{i=1}^{N}\langle P_o\rangle_i^2}, \quad \text{(A6)}$$

where the denominator represents the worst possible FBS (i.e., observed and forecast events have no spatial overlap). For the purposes of this study, the FSS is evaluated using the neighborhood ensemble probability approach outlined in Schwartz et al. (2010). In this approach, $\langle P_f\rangle_i$ represents the fraction of ensemble members that exceed a given threshold within neighborhood N, at each grid point i.

REFERENCES

Anderson, J., 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, https://doi.org/10.1175/MWR-D-11-00013.1.

——, and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463, https://doi.org/10.1175/JTECH2049.1.

——, B. Wyman, S. Zhang, and T. Hoar, 2005: Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.*, **62**, 2925–2938, https://doi.org/10.1175/JAS3510.1.

——, 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, https://doi.org/10.1175/2009BAMS2618.1.

Benjamin, S., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Berner, J., S.-Y. 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, https://doi.org/10.1175/2010MWR3595.1.

Blaylock, B., J. Horel, and S. Liston, 2017: Cloud archiving and data mining of High Resolution Rapid Refresh model output. *Comput. Geosci.*, **109**, 43–50, https://doi.org/10.1016/j.cageo.2017.08.005.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brier, G., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Compo, G. P., J. S. Whitaker, and P. D. Sardeshmukh, 2006: Feasibility of a 100-year reanalysis using only surface pressure data. *Bull. Amer. Meteor. Soc.*, **87**, 175–190, https://doi.org/10.1175/BAMS-87-2-175.

Dirren, S., R. Torn, and G. Hakim, 2007: A data assimilation case study using a limited-area ensemble filter. *Mon. Wea. Rev.*, **135**, 1455–1473, https://doi.org/10.1175/MWR3358.1.

Fowler, T., J. H. Gotway, K. Newman, T. Jensen, B. Brown, and R. Bullock, 2018: Model Evaluation Tools version 7.0 (METv7.0): User's guide 7.0. Developmental Testbed Center, 408 pp., https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v7.0.pdf.

Gaspari, G., and S. Cohn, 2006: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

Hanson, G. S., 2016: Impact of assimilating surface pressure observations from smartphones on regional, convective-allowing ensemble forecasts: Observing system simulation experiments. M.S. thesis, Dept. of Meteorology and Atmospheric Science, The Pennsylvania State University, 47 pp., https://etda.libraries.psu.edu/files/final_submissions/11853.

Lei, L., and J. Anderson, 2014: Impacts of frequent assimilation of surface pressure observations on atmospheric analysis. *Mon. Wea. Rev.*, **142**, 4477–4483, https://doi.org/10.1175/MWR-D-14-00097.1.

Madaus, L., and C. Mass, 2017: Evaluating smartphone pressure observations for mesoscale analyses and forecasts. *Wea. Forecasting*, **32**, 511–531, https://doi.org/10.1175/WAF-D-16-0135.1.

——, G. Hakim, and C. Mass, 2014: Utility of dense pressure observations for improving mesoscale analyses and forecasts. *Mon. Wea. Rev.*, **142**, 2398–2413, https://doi.org/10.1175/MWR-D-13-00269.1.

Mass, C., and L. Madaus, 2014: Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343–1349, https://doi.org/10.1175/BAMS-D-13-00188.1.

McNicholas, C., and C. Mass, 2018: Smartphone pressure collection and bias correction using machine learning. *J. Atmos. Oceanic Technol.*, **35**, 523–540, https://doi.org/10.1175/JTECH-D-17-0096.1.

Miller, P. A., M. F. Barth, and L. A. Benjamin, 2005: An update on MADIS observation ingest, integration, quality control and distribution capabilities. *21st Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology/14th Symp. on Education*, San Diego, CA, Amer. Meteor. Soc., J7.12, https://ams.confex.com/ams/Annual2005/techprogram/paper_86703.htm.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Roberts, N. M., 2005: An investigation of the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall. Met Office Tech. Rep. 455, 80 pp.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

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, https://doi.org/10.1175/2009WAF2222267.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., http://dx.doi.org/10.5065/D68S4MVH.

Wheatley, D., and D. Stensrud, 2010: The impact of assimilating surface pressure observations on severe weather events in a WRF mesoscale system. *Mon. Wea. Rev.*, **138**, 1673–1694, https://doi.org/10.1175/2009MWR3042.1.

Whitaker, J. S., G. P. Compo, X. Wei, and T. M. Hamill, 2004: Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.*, **132**, 1190–1200, https://doi.org/10.1175/1520-0493(2004)132<1190:RWRUED>2.0.CO;2.