Advanced Approaches for the Collection, Quality Control, and Bias Correction of Smartphone Pressure Observations and Their Application in Numerical Weather Prediction.

Callie McNicholas

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Clifford Mass

Gregory Hakim

Dale Durran

Program Authorized to Offer Degree

Atmospheric Sciences

University of Washington

Abstract

Advanced Approaches for the Collection, Quality Control, and Bias Correction of Smartphone Pressure Observations and Their Application in Numerical Weather Prediction.

Callie McNicholas

Chair of the Supervisory Committee:
Dr. Clifford Mass
Atmospheric Sciences

Distributed worldwide, over 300 million smartphones are now capable of measuring atmospheric pressure, providing a potential surface observing network of unprecedented density and coverage. To examine the utility of such a network and test potential approaches for collection, quality control, and bias correction of smartphone pressures, a novel smartphone pressure app was developed. Within this app, observational uncertainty was quantified, quality control was performed, and sources of error were minimized. New machine learning techniques were developed to proactively correct observation bias. To test the performance of the app and evaluate the utility of smartphone pressure observations, WRF-EnKF experiments were performed for two case studies. In both case studies, smartphone pressures were able to constrain forecasts and analyses of observed and unobserved variables. In the second case study, full-cycling experiments assimilating smartphone pressures were able to successfully predict the forecast track and intensity of a major wind storm. Partial cycling experiments revealed that by improving initial conditions smartphone pressure assimilation can enhance forecasts at short lead times.

Table of Contents

Chapter 1

THE UTILITY OF SURFACE PRESSURE AND THE POTENTIAL OF CROWDSOURCING

1.1) Review of crowdsourcing surface pressure networks.

Over the past quarter-century, the advent of the Internet and advances in telecommunication and sensor technology have resulted in a substantial increase in the availability of real-time surface pressure observations from local MESONETs and privately operated citizen weather observer networks. Three of the most prominent citizen weather networks are the Citizen Weather Observer Program (CWOP) which has grown to over 6000 members, the WeatherBug network which has deployed over 8000 weather stations at neighborhood schools and public facilities across the US, and the Wunderground network which consists of over 30,000 private weather stations worldwide. Combined, these networks provide tens of thousands of near-real-time surface pressure observations across the United States. While these networks have rapidly increased the number of available pressure observations across the US, sizeable gaps in data availability remain, especially in rural areas, as most private weather stations are located in urban areas. Furthermore, quality control issues and limited temporal resolution remain a challenge for existing privately operated weather networks. Fortunately, a rapid increase in the temporal and spatial resolution of surface pressure observations may be achieved with the aid of smartphones.

Over the past three years, well over half a billion smartphones capable of measuring atmospheric pressure have been sold worldwide. Two crowd-sourcing applications that have taken advantage of the proliferation of smartphone barometers are PressureNet and OpenSignal. In 2015, the mobile collection of surface pressure from both of these apps exceeded **100,000**

observations an hour. In August 2016, The Weather Channel app, with over 50 million downloads, began collecting upwards of **3 million** observations an hour from a mere 3% of their user base. By mid-2017, **30 billion** pressures will be retrieved every month from the Weather Channel app. If infrastructure and data quality challenges can be met, these observations could revolutionize numerical weather prediction (NWP) by providing unprecedented observational coverage and density. Potentially even more smartphone pressures could be collected if other major apps were provided with pressure-collection codes.

1.2) Background and Literature Review

As of January 2017, over 170 smartphones are capable of measuring surface pressure. Surface pressure is arguably the most valuable observed surface variable due to its ability to provide information on all scales of motion ranging from convectively induced cold pools to mid-latitude cyclones. Surface pressure reflects atmospheric structure aloft above the planetary boundary layer (PBL), in contrast to surface temperature, moisture, and wind. Unlike other surface measurements, surface pressure is less affected by representation error as the distribution of surface pressure, corrected for elevation variation, is typically more spatially homogeneous than other surface parameters and has little variation between the interior and exterior of buildings. Measurements of pressure exhibit fewer exposure problems and are less sensitive to the placement of instrumentation than measurements of other surface variables. For these reasons, among others, several studies during the last decade have examined the impact of surface observations, and in particular surface pressure, on numerical simulations and forecasts across a variety of atmospheric scales.

At the planetary scale, Whitaker et al. (2004) showed that assimilating surface pressure at a relatively sparse collection of locations could produce realistic lower and middle tropospheric

analyses, with 500 hPa errors similar to present 2.5-day forecast errors. Compo et al. (2006) assimilated surface pressure observations using ensemble techniques and found surface pressure observations alone could reproduce the entire upper-tropospheric circulation, even during the first half of the 20th century. This remarkable conclusion led to the construction of the 20th Century Reanalysis using surface pressure observations alone (Compo et al. 2011).

Dirren et al. (2007), assimilating observations from an idealized regional network of surface pressure observations, was able to constrain lower-tropospheric fields and geopotential heights through the middle troposphere. A recent study examining the impact of assimilation frequency by Lei et al. (2014) came to a similar conclusion, noting that assimilation of surface pressure observations at all frequencies constrained observed state variables and reduced error and uncertainty through the depth of the troposphere. As assimilation frequency was increased, the analysis error through the depth of the troposphere decreased. This conclusion is in agreement with a previous idealized study by Anderson et al. (2005), which observed a monotonic decrease of prior error for all state variables as the density and frequency of assimilated surface pressure observations was increased.

While high frequency (hourly or less) assimilation of surface pressure has produced significant reductions in analysis error in idealized planetary and synoptic scale simulations, a major question asks whether the same result could be achieved in mesoscale experiments. Wheatley and Stensrud (2010) assimilated altimeter setting and 1-h surface pressure tendency along with traditional surface observations such as temperature, moisture, and winds from METAR stations. The hourly assimilation of altimeter setting, and to a limited extent surface pressure tendency, constrained errors in meso-high position and intensity, resulting in improved model depictions of cold pools associated with MCS events. Madaus et al. (2014) performed a

similar study utilizing a high-density network of routine METAR and non-traditional MESONET observations from private operators such as the Citizen Weather Observer Program (CWOP). In this study, quality control and bias correction techniques were implemented, using the Rapid Update Cycle Surface Assimilation System (RSAS) altimeter analysis as truth. Implementation of these techniques resulted in substantial improvements in 3-h forecasts of altimeter setting. In full-cycling experiments, Madaus et al. (2014) observed a significant reduction in domain averaged error when additional, high-density, pressure observations were assimilated. However, without bias correction, this error reduction was not observed. Since pressure tendency is independent of bias, Madaus et al. (2014) examined the effects of assimilating pressure tendency and found that increasing the frequency of pressure tendency assimilation from 3-h to 1-h produced a more realistic distribution of mesoscale precipitation during a Puget Sound Convergence Zone event.

Recently available smartphone pressures have characteristics that are similar to the MESONET observations employed in the aforementioned studies. Both smartphone and MESONET pressures suffer from quality control issues and representation error due to uncertainty in elevation. Like MESONET sites, smartphones are capable of submitting observations at high temporal frequency. But smartphone pressures observations (SPOs) outnumber MESONET pressure observations by orders of magnitude.

During the past two years, several smartphone case studies appeared in the literature. Mass and Madaus (2014) described the potential of crowdsourcing smartphone pressures for mesoscale weather prediction and provided an example of SPO assimilation during a convective event in eastern Washington State. Their convective case study showed that SPOs could modify and potentially enhance short-term mesoscale forecasts. Hanson and Greybush (2016) performed

idealized simulations with synthetic smartphone observations. They concluded that if observational uncertainty can be well represented, smartphones pressures could improve model forecasts of surface variables. Madaus and Mass (2017) assimilated quality-controlled smartphone pressure observations from networks like PressureNet and OpenSignal but failed to produce improvements in forecasts of surface variables. The forecasts skill for some surface variables, like 2m-temperature and 2m-humidity, decreased after the assimilation of quality-controlled SPOs. This result, suggests that significant advances in quality control and post-processing of smartphone observations need to occur before SPOs may be of value to NWP.

1.3) Motivation and problem statement.

The utility of surface pressure assimilation has been demonstrated in the literature. With the availability of millions, and potentially, billions of SPOs from providers like the Weather Company, a global pressure network of unprecedented density has become feasible. Lei et al. (2014) and Anderson et al. (2005) suggested a monotonic relationship between observation density and prior error. If even a fraction of smartphone observations can be utilized for NWP, their superior spatial coverage could potentially provide significant improvements in forecast skill. Unfortunately, crowdsourcing pressure apps have not been designed to provide SPOs suitable for NWP. Incoherency and a lack of consistency among SPOs currently undermine the use of the smartphone pressures in NWP. If smartphones are to contribute to NWP and become a viable atmospheric observing platform, sources of observational uncertainty must be constrained and quantified. New approaches to quality control and bias correction are needed, as are insights into the best design of pressure-collection protocols on mobile platforms. To address these challenges a novel crowdsourcing pressure app, *uWx* was developed.

Chapter 2

DEVELOPMENT AND PERFORMANCE OF AN INNOVATIVE PRESSURE COLLECTION APP, *uWx*

2.1) In-app quality control framework

2.1.1 Motivation

A quality control (QC) framework was developed within the app to reduce sources of observational error, quantify uncertainty and bias, and provide estimates of pressure change. Previous apps such as PressureNet and OpenSignal did not QC SPOs. Since these apps only provided pressure and location information, attempts at QC were limited in scope. In Madaus and Mass (2017), QC of SPOs involved validity, statistical, and spatial consistency checks that borrowed heavily from the QC framework used by the Meteorological Assimilation Data Ingest System (MADIS) (Miller et al., 2005). Due to the poor quality of SPOs, spatial consistency checks could not be performed internally and instead were performed with MESONET and METAR observations. Even with these QC procedures in place, a significant fraction of SPOs increased the ensemble mean analysis error in the no-cycling experiments performed by Madaus and Mass (2017). This suggests that traditional QC approaches for SPOs are insufficient. In this thesis, an effort is made to improve the quality of SPOs before they are retrieved in-app. Benefits of in-app QC include immediate access to a wealth of potentially useful environmental data from photodetectors, accelerometers, and Global Positioning System (GPS) hardware. In app-QC also speeds up post-processing of SPOs by eliminating the burden of performing QC for thousands of phones remotely.

2.1.2 Constraining Observation Error and Quantifying Uncertainty

The largest source of systematic error for SPOs is elevation error. The elevation of a smartphone is required to compute altimeter setting, a reduction of pressure to sea-level using the U.S. Standard Atmosphere. Since a smartphone's location is inherently linked to its elevation, errors in the position of a smartphone can lead to large errors in elevation. To reduce location errors, the in-app QC framework forces all location retrievals to utilize the onboard GPS in high accuracy - high power mode. Wi-Fi and the Cellular Network location detection are used in tandem with the GPS to enhance location retrieval. By mandating the use of the GPS, location accuracy can be constrained to tens of meters (Fig. 1). Since multiple GPS locks can improve location accuracy, up to two GPS locks are performed if a smartphone's position deviates more than 60 meters from its previous position. The ground elevation at the smartphone's location (latitude, longitude) is retrieved from a USGS Digital Elevation Model (DEM) with a resolution of 30 meters and an RMSE of 1.55 meters (Gesch and al., 2014). Elevation uncertainty is found by summing the RMSE of the DEM and the variance of the surrounding DEM grid points. Since the 90% confidence interval of GPS location typically does not exceed 90 meters, the local elevation grid used to estimate elevation uncertainty was limited to the nine nearest DEM grid

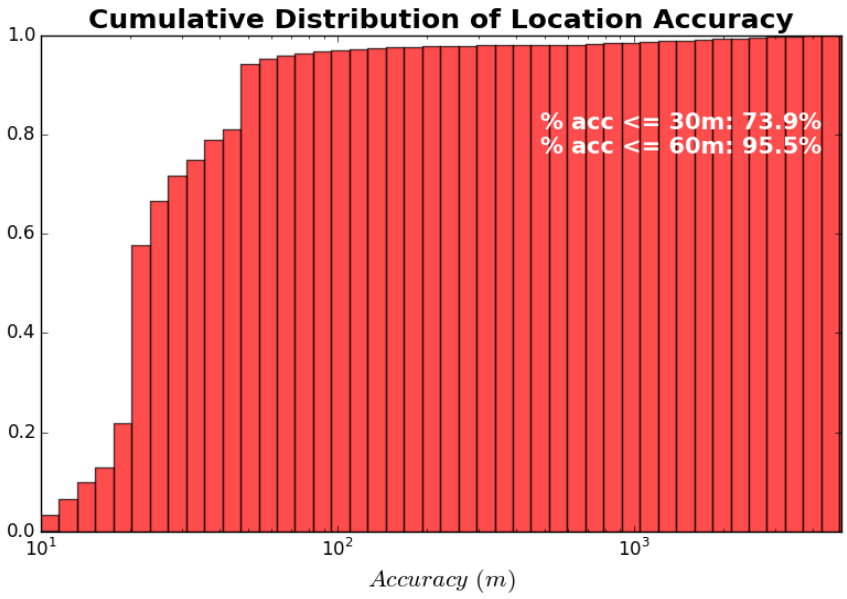


Figure 1: CDF of location accuracy of uWx SPOs averaged over a 1 month period.

points.

A secondary source of error in SPOs is dynamic lag. Dynamic lag is a consequence of the measurement process. Smartphone pressure sensors are microelectromechanical (MEMS) sensors that utilize a sensing diaphragm formed on a silicon substrate, which bends with applied pressure. This bending causes a deformation in the crystal lattice of the diaphragm. The deformation initiates a change in the band structure of piezo resistors placed on the diaphragm, leading to a change in the resistivity of the material. A typical MEMS sensor like the BMP280, used in a variety of high-end smartphones, has a relative accuracy of ~0.17 hPa and absolute accuracy of ~1 hPa (Bosch, 2016). Since MEMS sensors are extremely sensitive, oversampling and an internal Infinite Impulse Response (IIR) filter are often used to reduce signal bandwidth and help suppress high-frequency noise caused by wind, doors/windows shutting, etc. A sample IIR filter formula from the BMP280 is provided below:

$$x_f(t) = \frac{\left[x_f(t-1) * (k-1) + x(t)\right]}{k}$$

$$where\ k = \ filter\ coefficient\ ;\ x_f = \ filtered\ data\ ;\ x = unfiltered\ data$$

Filter coefficients for a typical MEMS sensor can range from two to sixteen. When a MEMS sensor transitions from sleep to measurement mode, the last measured pressure is used to initialize the IIR filter. The difference between this pressure and newly measured pressures can be substantial, as changes in atmospheric pressure and phone elevation can occur between measurements. Figure 2., shows the effect of different filter coefficients on the response of a MEMS sensor to a pressure change between measurements of 5 hPa.

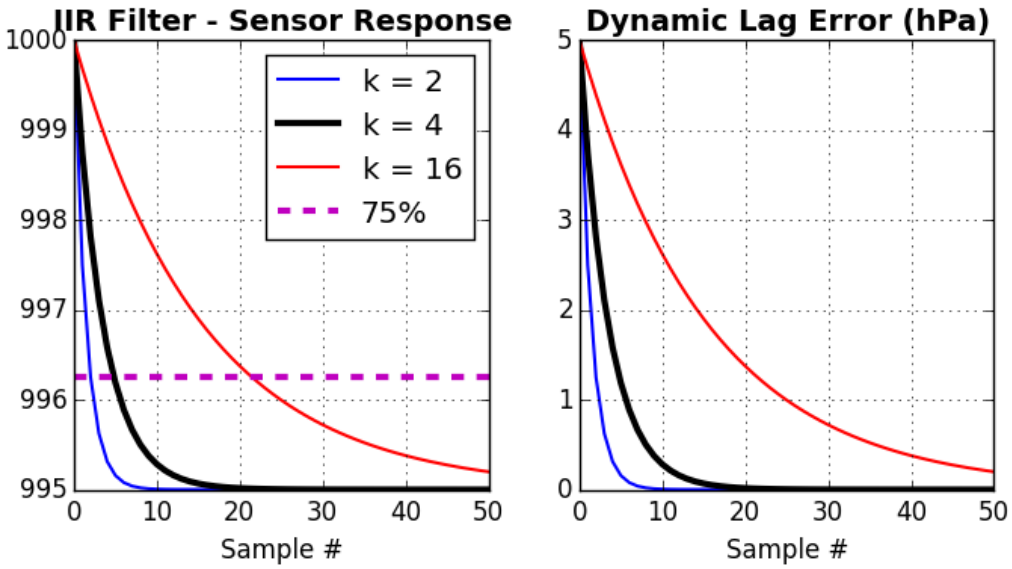


Figure 2: An example of the dynamic lag caused by an IIR Filter with various filter coefficients.

In the android operating system, the IIR filter coefficient is set to four in smartphones containing the BMP280. At this setting, five samples are required to capture 75% of the response. With a filter coefficient of sixteen, twenty-two samples are required to capture 75% of the response. The right panel of Figure 2, shows that capturing 75% of the response with five samples leads to a 1.25 hPa dynamic lag error.

The frequency of pressure collection is quasi-periodic in nature. Operating systems regulate background tasks by limiting their use when the device has been inactive for some time. Since phone activity and is dependent on the user, the frequency of pressure collection is a function of user behavior. Atmospheric pressure and phone elevation can change significantly over time. Consequently, the quasi-periodicity of pressure collection can result in a wide range of observed pressure changes between measurements. Since the dynamic lag error is proportional to the magnitude of the pressure change between measurements, the in-app QC framework employs a long spin-up period to allow the pressure signal to reach a state of quasi-equilibrium before pressure retrieval is performed. A typical spin-up period can last from 30-40 seconds. With a

sampling rate of 20 Hz, between 600-800 measurements are retrieved during this spin-up period, minimizing dynamic lag error. Since all MEMS sensors likely employ some form of IIR filtering a spin-up time of at least 15 seconds is recommended to mitigate dynamic lag error.

2.1.4 Bias Estimation

It is important to note that in many cases smartphones are not at ground level. Thus, even with a precise location, the ground elevation estimate can deviate substantially from the true elevation of a smartphone. Most dwellings are built on foundations that can be a foot to several feet above ground level. In urban and suburban areas phone elevations can deviate tens to hundreds of meters in multi-story buildings, high rises, and underground rail stations. These deviations are quantified by the in-app QC framework as an additional component of observational uncertainty. To quantify this added vertical uncertainty an estimate of the true altimeter setting is retrieved from surrounding, MESONET and METAR observations, retrieved from the MADIS network.

To interpolate altimeter observations to a smartphone location, an optimum interpolation (OI) scheme was developed. Nearby MADIS observations are placed into four quadrants spanning the four cardinal directions around the smartphone location. The nearest two observations, in each quadrant, are retained for OI. The OI is only performed if at least three quadrants contain observations within 300 km of the SPO. This is to ensure that the OI interpolant will be representative and not spatially biased. A piecewise cubic spline is used to temporally interpolate MADIS observations to the time of the SPO. An inverse distance weighting (IDW) technique is then employed to spatially interpolate nearby MADIS observations to the location of the SPO. Cross-validation is used to estimate an appropriate power factor for IDW and jackknifing is performed to estimate the uncertainty of the synthetic

MADIS observation at the location of the SPO. The difference between this synthetic observation and the SPO is defined as phone bias and archived for later post-processing.

2.1.5 Pressure Change QC

In previous studies, such as Madaus and Mass (2017), pressure change was computed during post-processing. Observations with the same unique identifier, separated in time by a given time interval, and within thirteen meters of each other, were used to compute pressure change. With these requirements in place, only a small fraction of phones contributed pressure change observations. No consideration to location accuracy, elevation variance, and phone movement was given since such information was not provided by previous SPO providers.

To improve the quality of smartphone pressure change observations (SPCOs) sensor data from auxiliary environmental and motion sensors such as the accelerometer and photodetector (light sensor) were retrieved. Illuminance, significant motion detection, pressure noise, location (latitude/longitude), location accuracy, and GPS speed were used to estimate the probability that a smartphone moved a distance exceeding a set threshold, defined by the terrain variance and location accuracy. In regions with minimal terrain variance, smartphones were allowed up to ninety meters of movement between observations. By retrieving additional sensor data in-app, the QC framework was able to retrieve frequent SPCOs from over 90% of smartphones (Fig. 3).

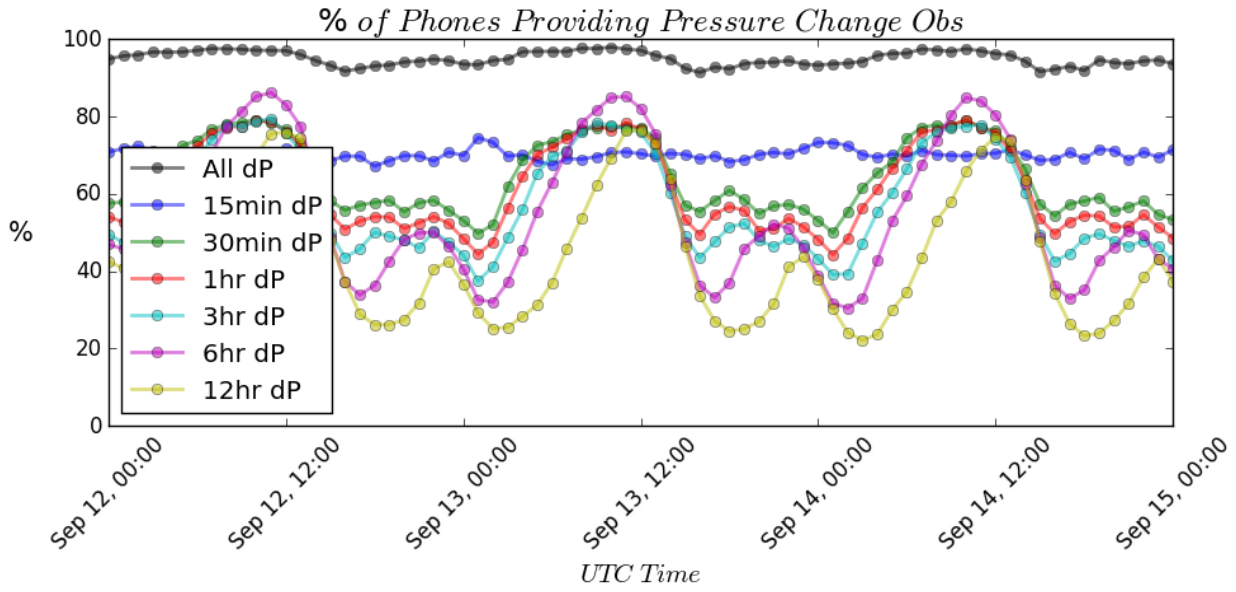


Figure 3: Time series of the percentage of *uWx* smartphones contributing pressure change observations.

2.2) Pressure Retrieval Procedure

One of the motivations behind developing the pressure collection app, *uWx*, was to define the best practices for mobile pressure collection. Extensive testing was performed with *uWx*, to achieve a balance between power consumption, observation frequency, and measurement accuracy. The following procedures are the final product of these tests.

1) An android background service is initiated on the smartphone triggering the collection and submission of pressure data. The user-adjustable frequency of this service varies from 5 - 60 minutes and is initially determined by the battery capacity of the smartphone.
2) The barometer, light sensor, significant-motion sensor (if available), and temperature sensor (if available) are activated and allowed to spin up for 10-15 seconds to avoid the influence of dynamic lag.
3) The GPS receiver is powered on and set to high-accuracy mode (if permissible). Location retrieval is aided by utilizing nearby Wi-Fi networks (if available) and the cellular

network, in conjunction with the GPS. A maximum of two location updates are retrieved to improve location accuracy. To limit power consumption, only one location update is retrieved if the phone's position has moved less than sixty meters from the location of the last observation.

4) Sensor measurements are processed and pressure along with a plethora of other statistics (diagnostics like GPS speed, accuracy, etc.) are submitted to the app server. The server returns the elevation and elevation uncertainty at the location of the phone. This information is used to compute phone altimeter setting.
5) 15 minute, 30 minute, 1 hour, 3 hour, 6 hour, and 12-hour pressure change observations are computed by examining prior observations to determine if they fall in one of the aforementioned time windows. If so, the current and prior observation are evaluated by the in-app QC framework to determine whether the current observation was taken at the location of the prior observation.
6) If location accuracy was sufficient (< 60 meters), observational data retrieved over an hour prior are uploaded for bias estimation. By this time, nearby MADIS observations are available allowing for the computation of a synthetic altimeter observation at the phone's prior location. The difference between the phone altimeter setting and the synthetic observation is defined as the phone bias.

2.3) App Summary and Current Status

As of January 2017, the pressure collection app, *uWx*, has been downloaded over 5000 times. 3800 users actively use the app. Approximately, 1900 users actively contribute pressure observations. The large discrepancy between the number of active users and the number of users contributing pressures is in part because the app can be downloaded on any smartphone,

regardless of whether it contains a barometer or not.

Since the background service that retrieves pressure is run at pseudorandom intervals, *uWx* collects pressures continuously. On average, ten observations are retrieved every eight seconds. Hourly retrieval rates vary from 3500-4000 observations per hour. During National Weather Service (NWS) issued severe weather watches/warnings *uWx* doubles the pressure collection frequency allowing for hourly retrieval rates to approach 8000 observations per hour.

A key reason for the success of *uWx* has been its limited power consumption and minimal economic cost. On a typical Android device, the power draw from *uWx* averages around 20 milliamps (mA) but can range from 5- 40 mA depending on the frequency of pressure collection and app use. The average android device has a battery capacity of ~ 2800mA. Thus, it would take *uWx* alone, approximately 140 hours (5 days, 20 hours) to drain a typical android battery. By comparison, a background music player app, like the VideoLan Client (VLC) app, can consume 150-200 mA on a typical android. The modest power footprint of *uWx* demonstrates the viability of sub-hourly retrieval of crowd-sourced smartphone pressures. This is an important achievement as it opens the door for large-scale crowdsourcing of smartphone pressures.

2.4) App comparison

Previous work, performed by Hanson and Greybush (2016) and Madaus and Mass (2017) utilized crowd-sourced SPOs from PressureNet and OpenSignal. As of August 2016, the Weather Company has provided SPOs from a small percentage of their user base. PressureNet, OpenSignal, and The Weather Channel did not apply QC techniques in-app. To evaluate the utility of in-app QC, a comparison between *uWx* and the three aforementioned SPO providers is provided in Table 1.

Provider	*uWx*	*PNet*	*OSig*	*TWC*
% dP	90-95%	10-15%	N/A	10-15%
Mean Loc Acc	60.3m	567m	471m	527m
Median Loc Acc	21.6m	36m	36m	42m
Mean Ob Freq	36 min	3h 31min	N/A	3h 24min
Median Ob Freq	18 min	3h 20 min	N/A	2h 18 min
% hourly	91.6%	11.6%	N/A	8.2%
Monthly Ob Count	**3.5M**	**100M**	**10M**	**2B (3%)**

Table 1: Comparison of crowd-sourcing pressure apps.

While *uWx* collects the fewest SPOs it outperforms PressureNet, OpenSignal, and The Weather Company app in all other categories. The mean location accuracy of SPOs collected by *uWx* is nearly an order of magnitude smaller than the mean location accuracy of SPOs from all other providers. *uWx* collects SPOs at nearly five times the frequency of all other SPO providers. As a result, over 90% of smartphones in the *uWx* network submit SPO's every hour. By using ancillary sensor data to aid pressure change collection, *uWx* can retrieve SPCOs from over 90% of phones in the network at any given time. By comparison, only 10-15% of phones, on average, contribute SPCOs from PressureNet and the Weather Company. The reason for this difference is the nature of the location retrieval procedure used by PressureNet and the Weather Company. Both apps likely use the default fused location provider which does not mandate the use of the GPS. The default fused location provider can switch between the Wi-Fi Network and Cellular Network to get a location. This behavior is demonstrated in Figure 4, which shows a sample time series of location accuracy from ten Weather Company phones.

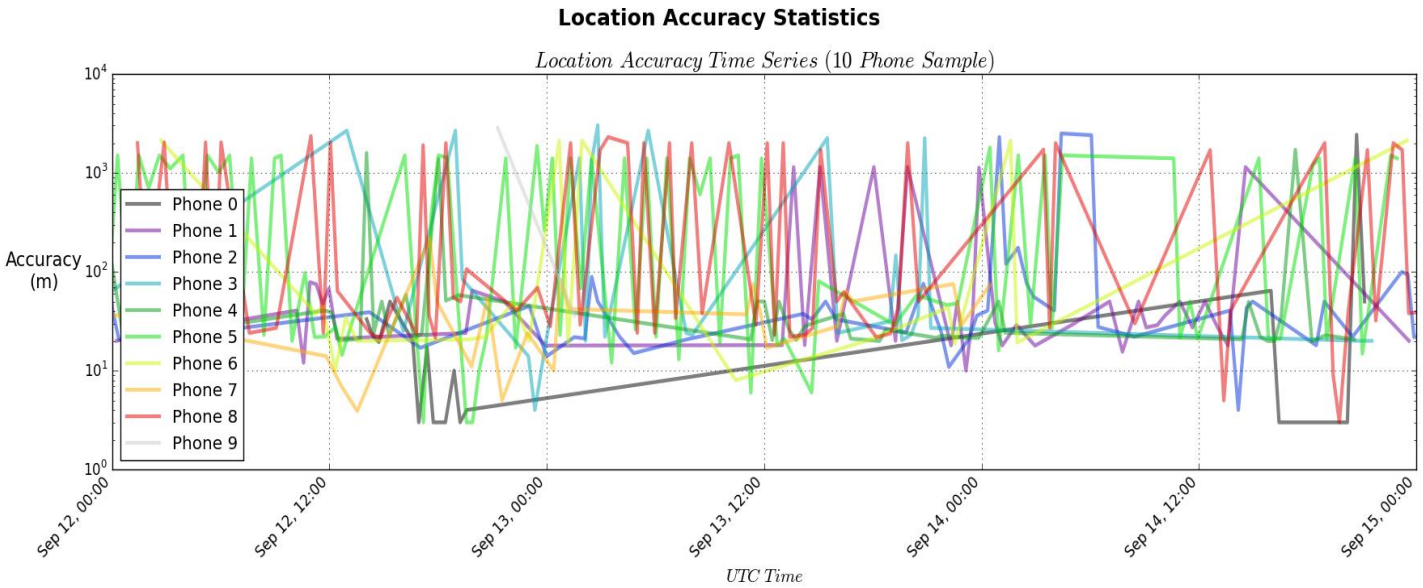


Figure 4: Time series of location accuracy from ten randomly sampled WC app users.

The frequent variability in location accuracy observed in Figure 4., is likely a result of the location retrieval switching between the Wi-Fi and Cellular networks. Since a location lock from a Wi-Fi Network often provides a different location estimate than a location lock from the Cellular Network, even when the phone is at the same location, this behavior can lead to spurious phone movement. This inhibits the collection of SPCOs, which are computed by comparing the spatial distance between two observations separated in time. *uWx* does not suffer from this problem as its in-app QC framework specifies stringent procedures for location retrieval substantially reducing the variability of location accuracy responsible for spurious phone movement (Fig. 5).

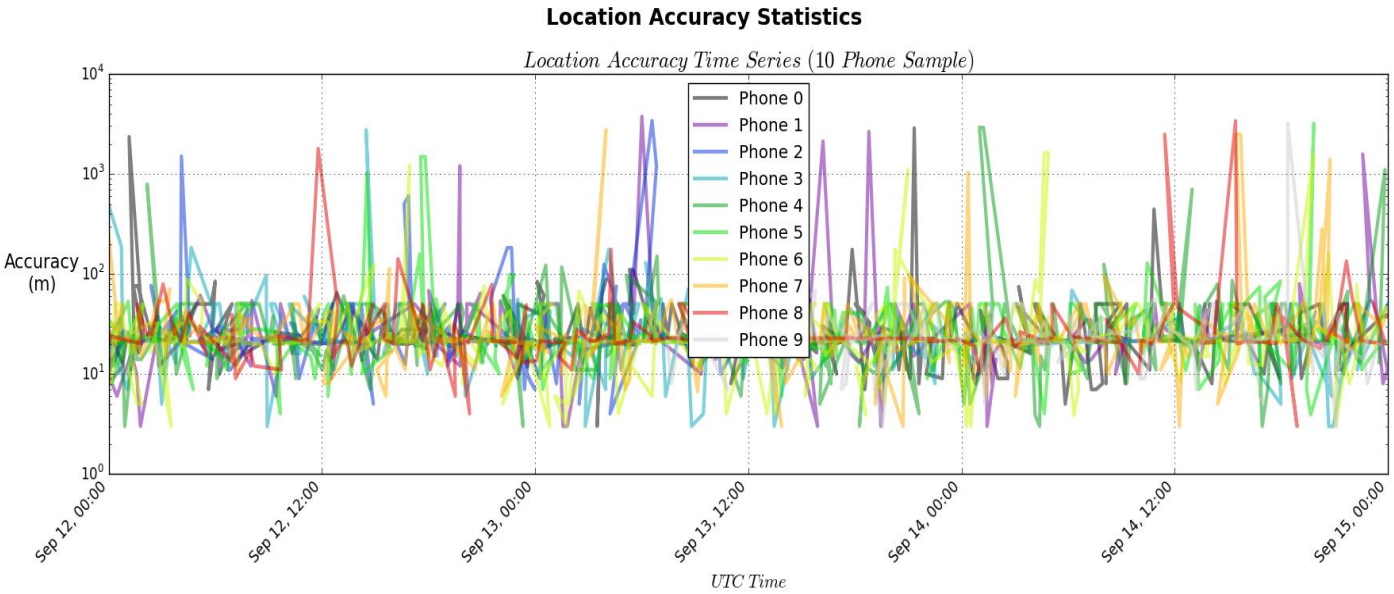


Figure 5: Time series of location accuracy from ten randomly sampled *uWx* users

Chapter 3

A MACHINE LEARNING APPROACH TO POST-PROCESSING SMARTPHONE PRESSURE OBSERVATIONS

3.1) Motivation

Analysis of observations from smartphones can reveal patterns associated with user behavior that can be useful for quality control applications. For example, phone statistics like location accuracy, elevation uncertainty, and phone bias often present some degree of similarity at frequented locations. For phones in motion, common modes of travel can be discerned from GPS data. For example, a GPS speed of 10 m/s indicates that the phone is likely traveling in a vehicle. With sufficient observations, relationships between statistics like GPS speed and elevation uncertainty can provide insights into dynamic lag and phone bias. Such relationships between phone statistics and phone bias are often subtle and nonlinear. This presents a formidable challenge for traditional multivariate regression techniques. In contrast, machine learning techniques excel at identifying hidden insights and quantifying patterns in data. A variety of machine learning techniques have been applied to both *uWx* and Weather Company (WC) data to determine whether phone bias can be predicted from data derived from environmental sensors and GPS hardware. Due to the sheer size of the WC dataset, only WC data from the Pacific Northwest (PNW) has been used in this study.

3.2) Methodology

To determine the optimal machine learning algorithms a variety of different machine learning algorithms were tested on *uWx* data collected between August 15 – October 15, 2016.

The algorithms included extremely randomized trees, random forests, adaptive boosted trees (AdaBoost), gradient boosted trees (Gradient Boost), and logistic regression. Since the behavior of every phone is unique, each algorithm was trained on time-series data from a single phone. While generating a unique algorithm for each phone is costly, it is expected that an optimum machine learning algorithm will eventually be incorporated into *uWx* relieving the burden of generating a distinct algorithm for every phone during post-processing.

To quantify the performance of each machine learning algorithm, four-fold cross-validation was performed. Data from each phone was randomly ordered and split into four quartiles. One quartile was used for verifying the machine learning algorithm, while the remaining three quartiles were used for training. This process was repeated three additional times so that all quartiles were used for both evaluation and training. Phones with fewer than fifty observations during the two months were discarded. The input matrix for each machine learning algorithm included the eleven variables collected in real-time by *uWx*: latitude/longitude (arbloc), luminosity (lux), elevation (elev), elevation variance (estd), pressure (pres), altimeter setting (alts), pressure noise (pstd), location accuracy (gps_acc), GPS Fix boolean (gps_fix), number of satellites in view (gps_sat), indoor/outdoor boolean (loc), significant motion boolean (sigmotion). This multivariate input was regressed to the phone bias. Once the algorithm was trained, data withheld for verification was provided to the machine learning algorithm which then returned an estimate of phone bias. This estimate of phone bias was compared to the observed phone bias. Recall from Chapter 2, that the observed phone bias was defined as the difference between a synthetic MADIS observation and the phone observation. The difference between the machine learning estimate of phone bias and the observed phone bias is hereafter referred to as the *bias prediction error*.

For each phone, the absolute value of bias prediction error was computed and then averaged to produce an estimate of mean absolute error (MAE). In Figure 6, the distribution of MAE for all phones is displayed for a variety of machine learning algorithms. The distribution of MAE from a simple averaging technique is provided for comparison. The simple averaging technique is akin to a persistence forecast, it computes the average phone bias during the training period and uses this value as the predicted phone bias for all verification data.

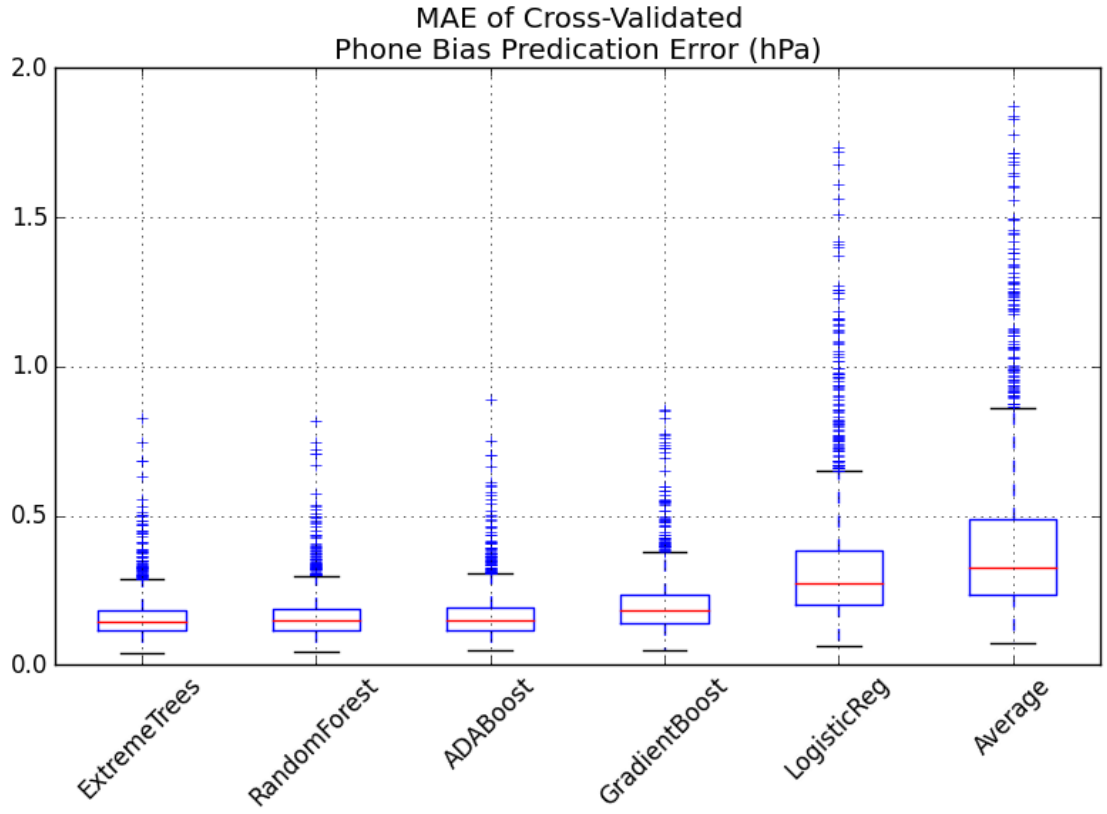


Figure 6: Comparison of machine learning techniques applied to *uWx* SPOs.

Compared to this averaging technique, all decision tree-based learning algorithms do a remarkable job of predicting the phone bias. The median, MAE of bias prediction error is markedly low because for any given phone the vast majority of observations are taken at a handful of frequented locations where, through repeated sampling, the distribution of phone bias has become more narrowly peaked. From the four decision tree-based learning algorithms

examined, the Random Forest algorithm (Breiman, 2009) was selected for SPO bias prediction due to its efficiency, simplicity, and diagnostic capabilities.

3.3) Bias Correction with Random Forests

3.3.1) Random Forests

Random forests are a form of ensemble decision tree learning. Decision tree learning works by sub-setting/splitting input data based on information gain (classification) or variance reduction (regression) until further splitting is not possible or provides no added value. Since decision trees mimic human decision-making, they are easy to interpret and comprehend. Individual decision trees can lack robustness and suffer from overfitting, due to excess complexity. One technique used to reduce this overfitting problem is ("bagging") or bootstrap aggregation (Breiman 1996). Bagging creates an ensemble of decision trees by selecting random samples with replacement from a training data set and fitting individual decision trees to these samples. This bootstrapping technique can improve performance by decreasing variance without increasing bias but is still prone to overfitting if a few independent variables (features) in the input training data are strong predictors of the predictand.

To improve the diversity of decision trees within a bagged ensemble, the random subspace method can be employed (Ho 1998). This method randomizes the selection of features during each split in the tree learning process. For regression problems, the number of features randomly sampled at each split is typically one-third of the total number of features. At each candidate split, the randomly sampled feature which minimizes the mean-squared error of the prediction is chosen as the feature to split. By randomly sampling features, strong predictors can be prevented from dominating bootstrapped trees, ensuring that trees in the ensemble do not become correlated. By employing bagging (randomizing observations) and the random

subspace method (randomizing features), random forests can produce an ensemble of uncorrelated trees that, when averaged, produce predictions with small mean squared error and low bias.

3.3.2) Feature Importance

The ensemble nature of random forests precludes a direct interpretation of the tree learning process. Nevertheless, random forests can be used to evaluate the importance of independent variables (features). Consider a feature *f*. For each node in a tree, which is split on *f*, the variance reduction of the node is weighted by the number of training observations that reached the node. This weighted variance reduction estimate is summed for all nodes in the tree which split on *f*, providing an estimate of the importance of *f* for a single tree. This process is repeated for all trees in the ensemble so that an average across the ensemble of trees can be computed, producing an estimate of the importance of *f* for the entire random forest. Feature importance is typically normalized for easier interpretability.

The random forest feature importance for *uWx* SPO's collected between August 15 – October 15, 2016, is displayed in Figure 7. Notably, no single feature dominates. The best predictor of phone bias, latitude/longitude location (arbloc), has a modest importance of ~0.2 on a scale from 0-1. Since all features exhibit relatively low importance it can be concluded that sensor and GPS data are individually weak

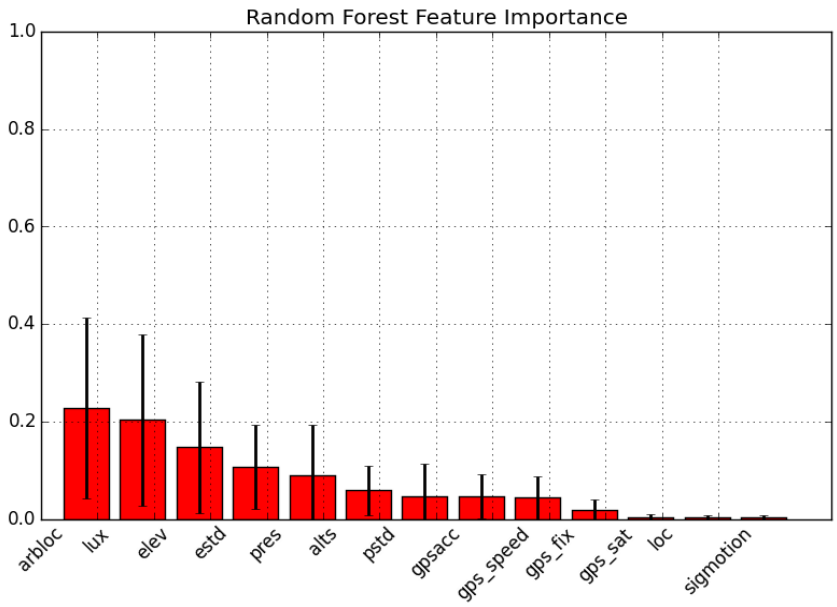


Figure 7: *uWx* random forest feature importance.

predictors of phone bias. Interestingly, the illuminance from the phone light sensor (unit: lux) is the second most important feature. This result was unanticipated but sensible. Illuminance is correlated with the environment. In residential homes illuminance averages between 160-320 lux, in the workplace illuminance is often greater (320-640 lux) as many workplaces are lighted with industrial fluorescent lights. Outside, the illuminance can vary from 1000 lux on an overcast day to over 10,000 lux in clear-sky conditions. Differences in illuminance aid phone bias prediction by serving as a secondary predictor of location.

3.3.3) Training Window

To evaluate how the extent of the training window affects bias prediction error, random forests were trained on *uWx* data over a one, two, and three-month period spanning August 15 to November 15, 2016 (Fig. 8). Since a continuous stream of users installed and/or uninstalled *uWx* during the training period, the extent of the training data set for every phone is unique. Around October 15th the number of users doubled due to additional app advertising during a local windstorm. This jump in users, mainly in the Seattle metro area, resulted in a slight increase in the number of outliers and the size of the interquartile range (IQR) for the three-month analysis. Interestingly, it also produced a notable change in feature importance. During the 3-month window, the importance of location markedly increased, while the importance of nearly all other featured decreased. It is unclear why this occurred as little change in feature importance was observed between the one and two-month cases. For this reason, it is hypothesized that changes in feature importance are more likely to be a consequence of the doubling of users than any increase in the length of the training window.

Both the MAE and MSE of bias prediction error did not change significantly when the training period was expanded from one to three months (Fig 8).

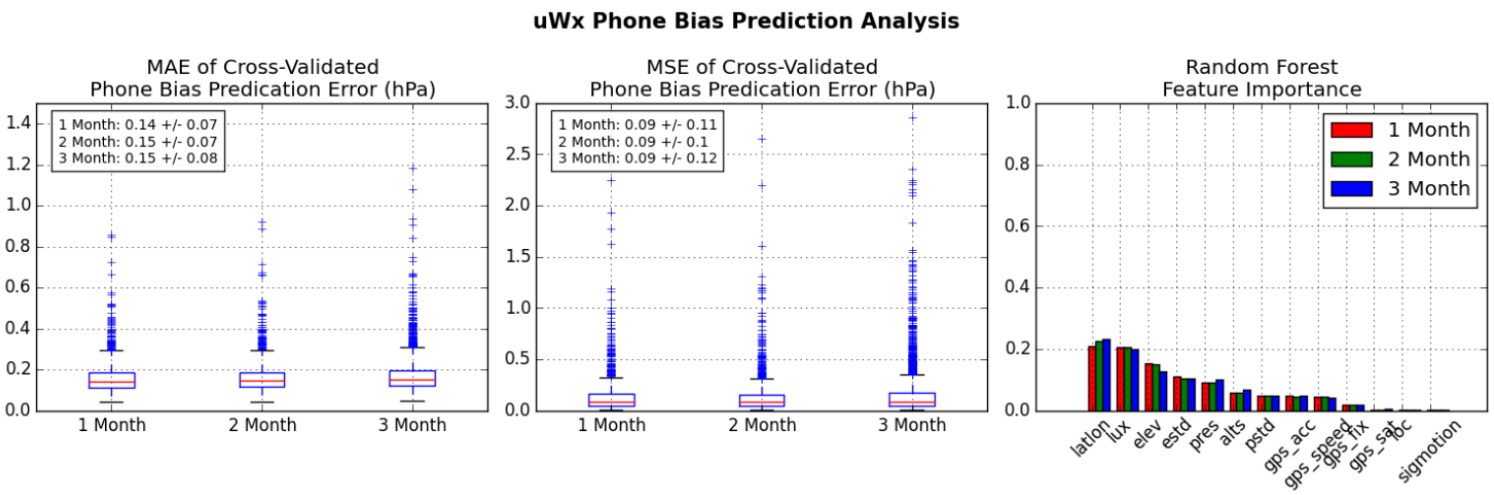


Figure 8: Training period comparison (*uWx*)

This is likely since the MAE of bias prediction error had saturated after one month. For all three cases, the median MAE of bias prediction error is approximately equal to the relative accuracy of a typical MEMS barometer (~0.12 hPa). Since *uWx* collects pressures at an hourly frequency, hundreds to thousands of observations may be taken from a phone over a month, providing sufficient training data to accurately predict phone bias.

The analysis described above for *uWx* data was repeated for WC observations (Fig. 9).

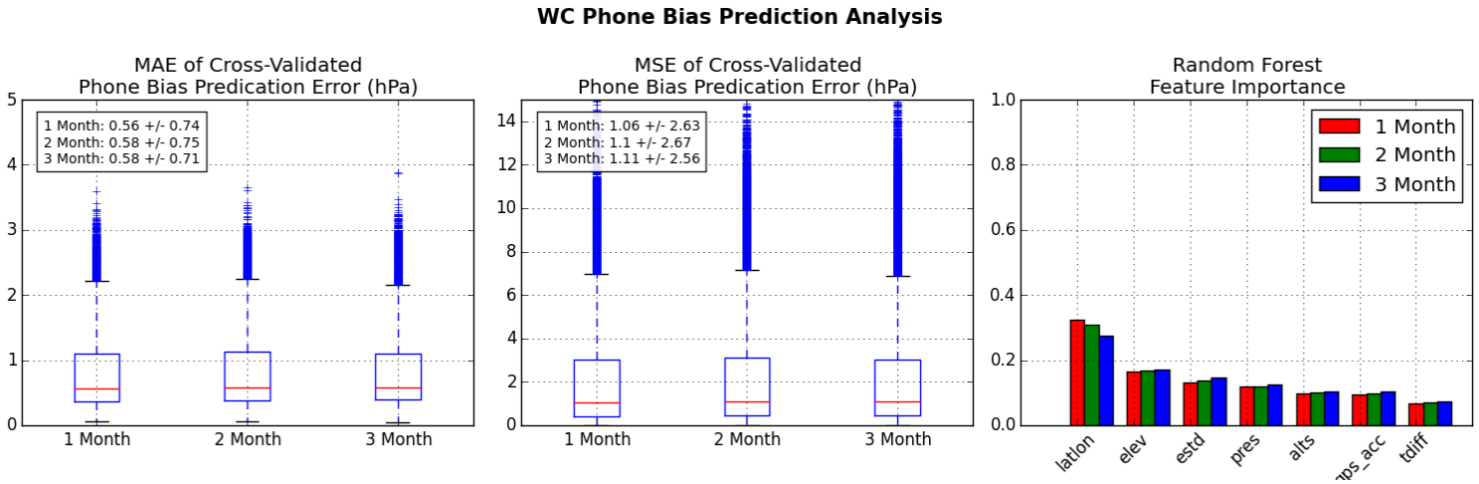


Figure 9: Training period comparison (Weather Company)

Like the *uWx* analysis, the MAE of bias prediction error was consistent for all training datasets,

irrespective of duration. The most notable difference between the *uWx* and WC analysis is the magnitude and skewness of the errors. Unlike the *uWx* analysis, the WC analysis is right-skewed toward outliers. The spread in the MAE of bias prediction error is also notably larger with the IQR for all cases spanning ~ 0.8 hPa. This result could be caused by a failure to account for dynamic lag, fewer features, and/or poor location retrieval. The latter issue is known to affect WC observations (Table 1). While WC data lacks some of the features provided by *uWx,* such as illuminance and GPS speed, the order of WC feature importance is in agreement with that found in the *uWx* analysis (Fig. 9). It is unclear how the absence of certain features, like illuminance, affects random forest performance.

For a better comparison between *uWx* and WC data, *uWx* feature selection was limited to those features available in the WC dataset. Three months of uWx and WC data were then used to cross-validate random forests with the same input features. One exception to this is the *tlag* feature. This variable is defined as the time delay between the observation time and the retrieval time of the observation. Since the majority of *uWx* data is retrieved in real-time (within thirty seconds of the observation time) the *tlag* feature was not derived from *uWx* data. Even with four features removed, random forests still performed well on the *uWx* data (Fig. 10).

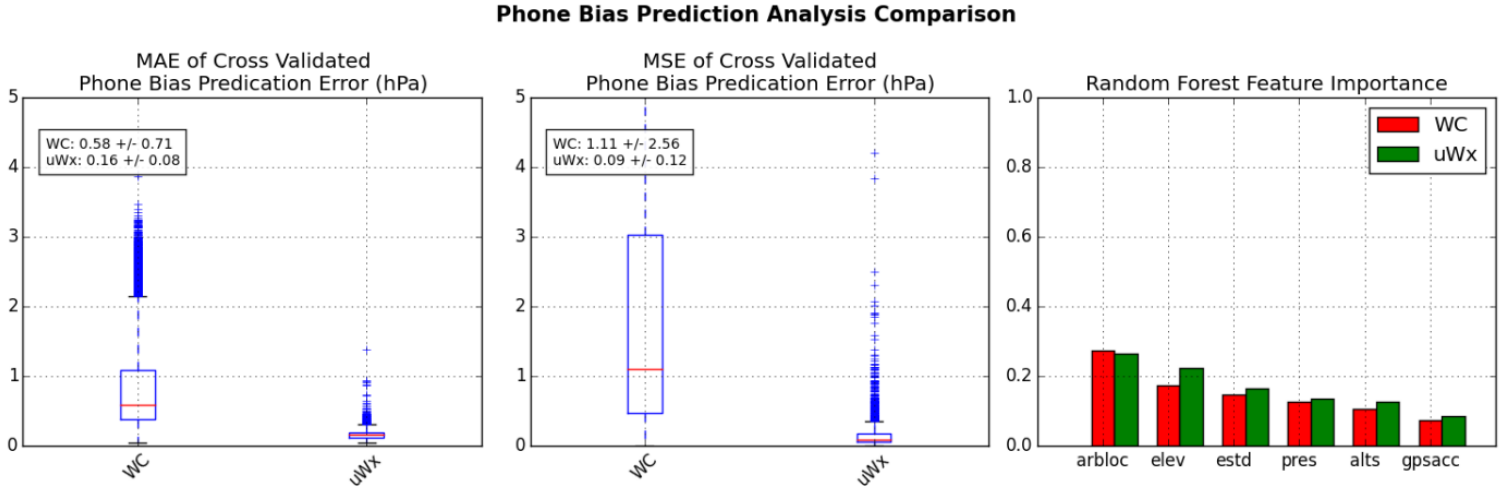


Figure 10: Comparison between *uWx* and WC random forest bias prediction.

This suggests, that the difference in the magnitude and skewness of the *uWx* and WC error distributions is more closely related to location accuracy. Improved location accuracy refines estimates of altimeter setting (i.e., the synthetic observations) used as the ground truth to train the random forest algorithm. Better location accuracy also provides better elevation and elevation uncertainty estimates, which in turn produce more accurate phone altimeter observations. The variance of location accuracy also plays an important role. If location accuracy frequently varies at a given location (Fig. 4) due to changes in location provider (like switching from GPS to Wi-Fi/cell network), this will lead to a wider distribution of phone bias and poorer random forest performance.

3.3.4) Clustering Analysis

When MESONET and METAR observations are bias-corrected, the average bias is simply taken to be the mean of the observation error over some time. This calculation is straightforward since the position of the observing instrument is fixed. Such analysis is not practical for smartphones that are capable of travel in three dimensions and often in motion. Fortunately, many phones spend considerable amounts of time at a select few locations. These common locations, such as homes and workplaces, can serve as de-facto observation sites.

To identify these common locations a data mining clustering technique, density-based spatial clustering of applications with noise (DBSCAN), was employed. DBSCAN is an ideal technique for clustering smartphone data as it is capable of identifying arbitrarily shaped clusters and is robust to outliers. An example of DBSCAN clustering analysis, performed on a month of *uWx* data from the author's smartphone, is displayed in Figure 11. In this analysis, a cluster radius of sixty meters was applied. The DBSCAN analysis reveals statistically significant differences in phone bias and illuminance between clusters.

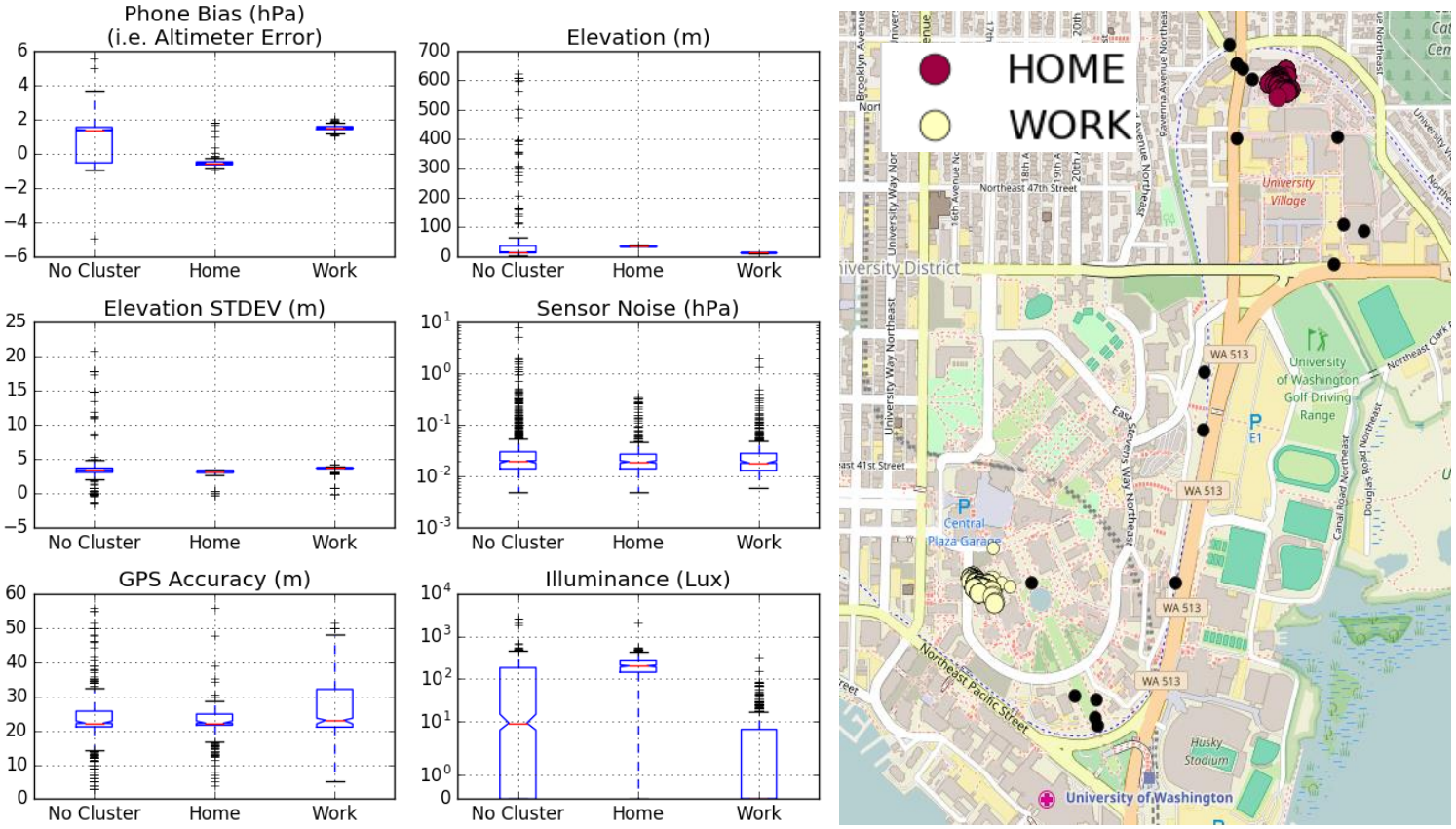


Figure 11: DBSCAN analysis of one month of *uWx* data from the author's smartphone

For both clusters, the IQR of phone bias is reduced by an order of magnitude, from ~2 hPa to ~0.2 hPa. This suggests that sensor bias is fairly conservative, as the IQR of phone bias at each cluster is roughly proportional to the relative accuracy of a typical MEMS pressure sensor. The median sensor noise and GPS accuracy are roughly the same for each cluster; however, at the work cluster, the variance of GPS accuracy is notably larger. Decomposing phone statistics with DBSCAN helps visualize the subtle relationships and patterns in the data that random forests utilize to predict phone bias.

Random forests are more successful in constraining bias prediction error with *uWx* data than with WC data. To better understand why this is the case, DBSCAN analysis for a randomly selected WC phone was performed (Fig. 12).

Like *uWx* data from the author's phone, the bias of this WC phone is notably different at the two identified cluster locations. Significant differences in the location accuracy between the two clusters are also observed. DBSCAN analysis of WC and *uWx* data suggests that clustering could be used to help predict phone bias.

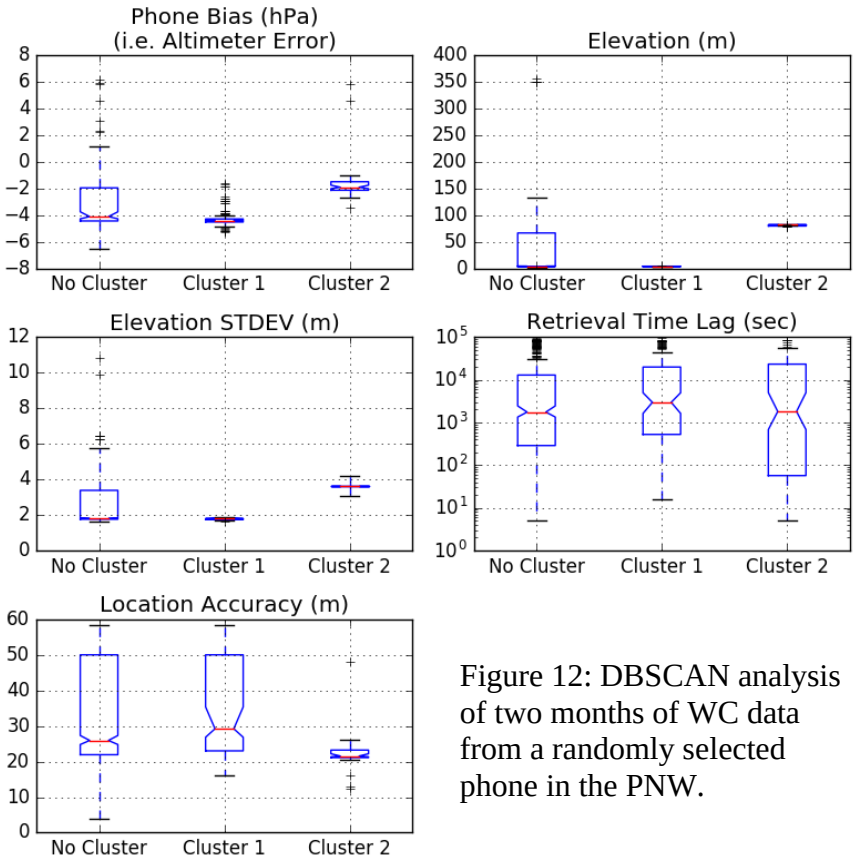


Figure 12: DBSCAN analysis of two months of WC data from a randomly selected phone in the PNW.

3.3.5) *Clustered Forests*

To examine whether clustering could improve bias prediction, three DBSCAN analyses were performed on 60 million WC observations, retrieved between August 15th and October 15th, 2016. Distance thresholds of 15, 30, and 60 meters were used to define cluster points. Different cluster radii were set to evaluate how clustering and location accuracy impact bias prediction error. For each cluster radius, only observations with a location accuracy less than or equal to the radius were used. Thus, for DBSCAN analysis with a cluster radius of fifteen meters only observations with a location accuracy less than or equal to fifteen meters were used. For a cluster to be defined, at least twenty-five observations had to fall within the cluster radius. Observations not falling within a cluster were labeled with a zero, while observations falling within a cluster were labeled with a positive real number. This cluster identifier was supplied to the random forest algorithm in place of the latitude/longitude (arbloc) feature. The performance of cluster-

based random forests is displayed in Figure 13. Previously, it was hypothesized that poor and variable location accuracy may explain why random forests are less effective at constraining bias prediction error for WC phones. Although bias prediction error decreases as the threshold on location accuracy and cluster radius decreases, reducing the threshold on location accuracy to a mere fifteen meters does not produce error statistics comparable to those derived from *uWx* data. This suggests that location accuracy alone cannot explain the gap between *uWx* and WC random forest performance.

As the cluster radius and constraint on location accuracy is reduced, the shape of the distribution of MAE does not change (Fig. 13).

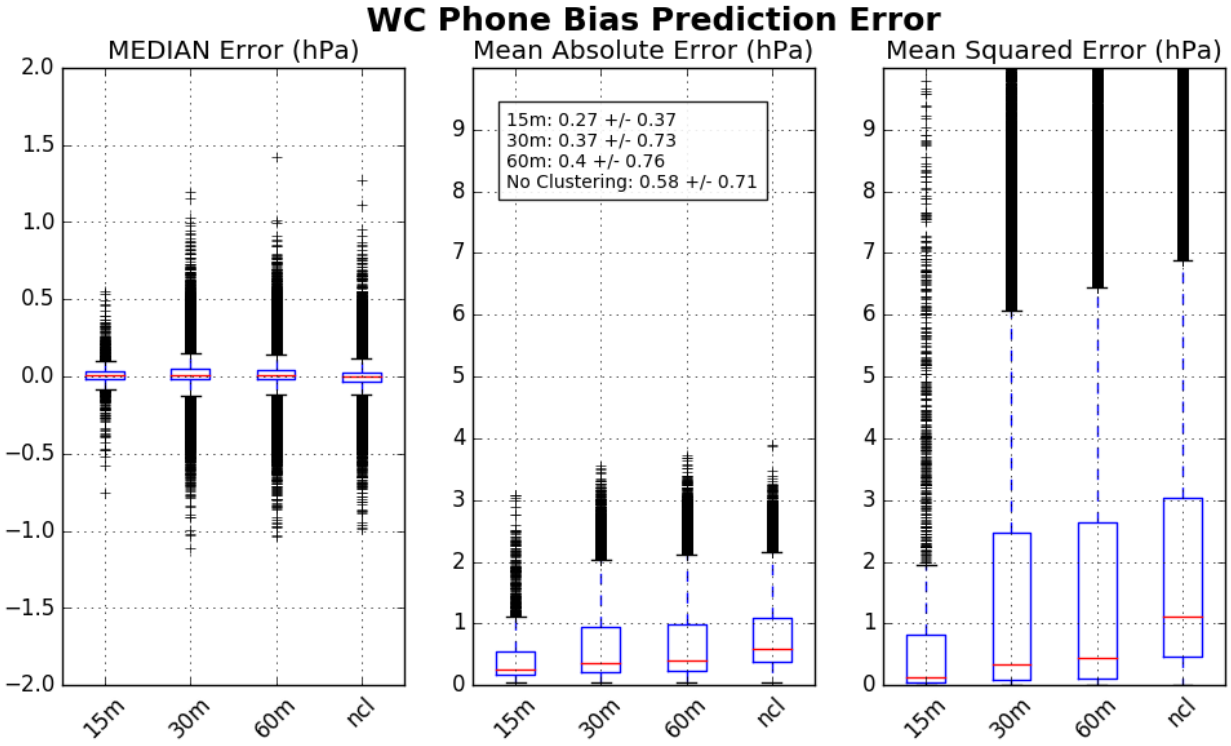


Figure 13: Performance of random forest bias correction with clustering for WC observations. NCL refers to the control case, for which clustering was not performed.

Regardless of location accuracy, the distribution of the MAE is right-skewed. It is hypothesized

that this skewness could be explained by a combination of infrequent pressure collection and a failure to account for dynamic lag. If dynamic lag was not accounted for, the variable frequency of pressure collection employed by the WC app would result in a wide range of dynamic lag errors. Figure 14, shows the distribution of the mean observation frequency for WC phones.

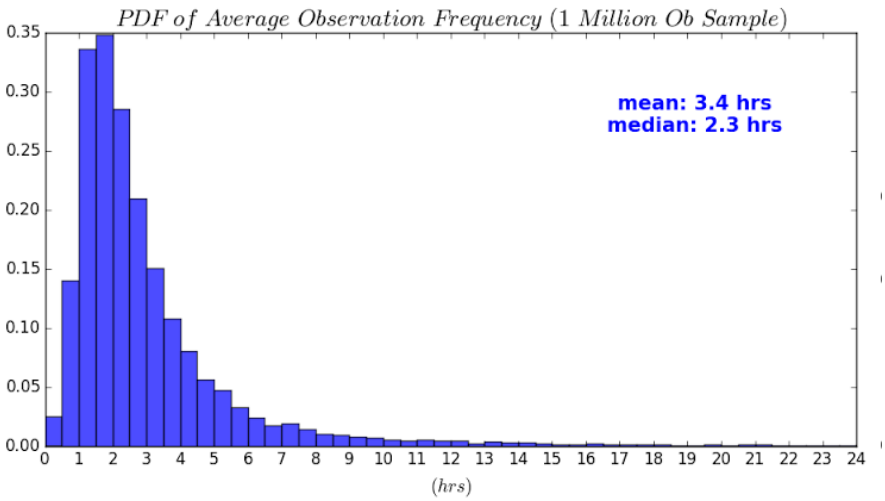


Figure 14: Histogram of WC observation frequency (one million observations sampled).

The distribution of observation frequency, like the distribution of MAE, is right-skewed. Dynamic lag error is linked to observation frequency as it scales with the magnitude of pressure change (Fig. 2). Synoptic-scale changes in atmospheric pressure often increase in magnitude as time progresses. For this reason, WC phones falling in the right tail of the observation frequency PDF may be expected to exhibit larger dynamic lag errors. This mechanism explains why the right skew of MAE is observed in Figure 13. It may also clarify why error statistics comparable to those derived from *uWx* are not achieved even when location accuracy is not an issue. A failure to fully account for dynamic lag adds a pseudo-random source of uncertainty to phone bias since the magnitude of dynamic lag error is a function of changes in elevation and atmospheric pressure. This added uncertainty limits the predictability of phone bias and the performance of random forests.

3.4) Quality Control with Radial Basis Functions

3.4.1) *Motivation*

Machine learning with random forests allows for phone bias to be proactively predicted from sensor and GPS data collected in real-time. One consequence of using random forests is that phone bias predictions will not exceed the range of phone bias observed during training. This can occasionally lead to poor predictions of phone bias which appear as outliers in bias-corrected altimeter analysis. Outliers can also arise when a phone ventures to an unfamiliar location where observational uncertainty is large, like the interior of a moving subway car. Since random forests cannot adequately predict phone bias 100% of the time, novel QC techniques have been developed to intelligently remove outliers from bias-corrected observations.

3.4.1) *uWx Quality Control*

Before performing QC, data from *uWx* phones are bias-corrected by random forests trained on past data. The first stage of QC employs simple validity checks to remove prominent outliers (i.e., altimeter setting < 890 hPa or > 1100 hPa). The second stage of QC involves a statistical check which removes statistical outliers exceeding four standard deviations from the mean of the observational dataset. Statistical outlier thresholds are modified to adjust for skewness using the split-histogram technique outlined in McNicholas and Turner (2014). The third and final stage of QC is a spatial consistency check which utilizes a radial basis function (RBF) in the form of a thin plate spline (TPS). RBFs are ideal for fitting pressure observations as they produce smooth surfaces and lack free parameters in need of tuning. Once a TPS is fit to the observations, outliers are determined by comparing the variance of the surface around the observation to the error of the interpolation (i.e. the difference between the observation and surface).

Figure 15 provides an example of the RBF spatial consistency check (RBF check). Observations displayed in this figure were bias-corrected with random forests trained during the month prior.

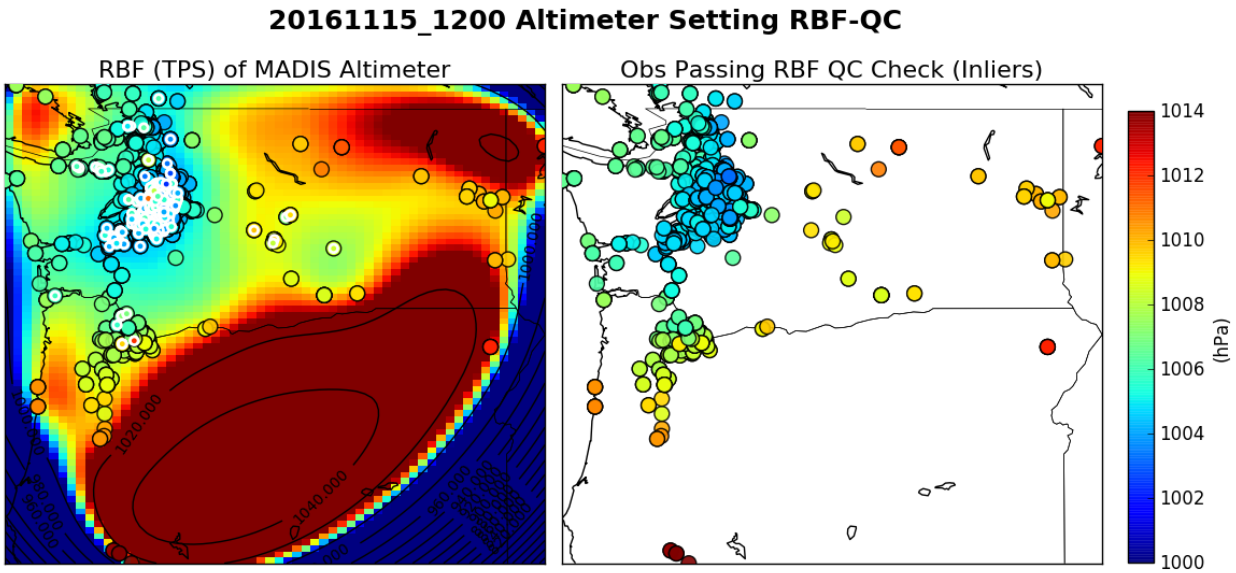


Figure 15: Example of *uWx* TPS analysis used during the spatial RBF check stage of QC. In the left plot, outliers are outlined in white.

In rural regions, where smartphone observations are sparser, the quality of the TPS is reduced and the surface is less smooth. In urban areas, the plethora of available smartphone observations produces a very smooth surface with low variance. Consequently, outlier removal is more liberal in urban areas and more conservative in rural areas. In regions where data does not exist, the TPS analysis is unphysical. On the peripheries of data-sparse regions, the TPS analysis exhibits large gradients increasing the outlier detection threshold. In urban areas around Seattle and Portland, the TPS analysis is fairly smooth decreasing the outlier detection threshold. By taking advantage of the properties of RBFs, intelligent, density-dependent outlier removal can be achieved allowing for crucial rural observations to be retained and for the removal of unnecessary noisy observations in urban areas.

3.4.2) *WC Quality Control*

The same set of QC techniques applied to *uWx* observations are applied to WC data, with one exception. The RBF-check for WC observations utilized a multiscale (hierarchical) RBF model with a Gaussian basis. A different form of RBF was used for QC as a large number of observations demanded a more efficient RBF algorithm (Bochkanov, 1999). To further reduce the computational burden of QC, a spatially filtered sample was extracted. WC observations were binned into 1x1 degree latitude/longitude boxes and then sorted by total uncertainty. Total uncertainty was defined as the sum of the random forest uncertainty (mean absolute bias prediction error), pressure uncertainty (elevation variance/sensor noise), and bias uncertainty (uncertainty of MADIS interpolation used to estimate phone bias). If the number of phones within a bin exceeded ten, then the ten phones with the lowest total uncertainty were selected. For bins with fewer than ten observations, all observations were added to the sample distribution.

An example of the spatially filtered RBF check described above is provided in Figure 16. WC observations displayed in this figure passed the first two stages of QC and were bias-corrected with random forests trained during the month prior. Although the sampled observations were not entirely self-consistent, especially in rural areas, the RBF fit to the data is still able to

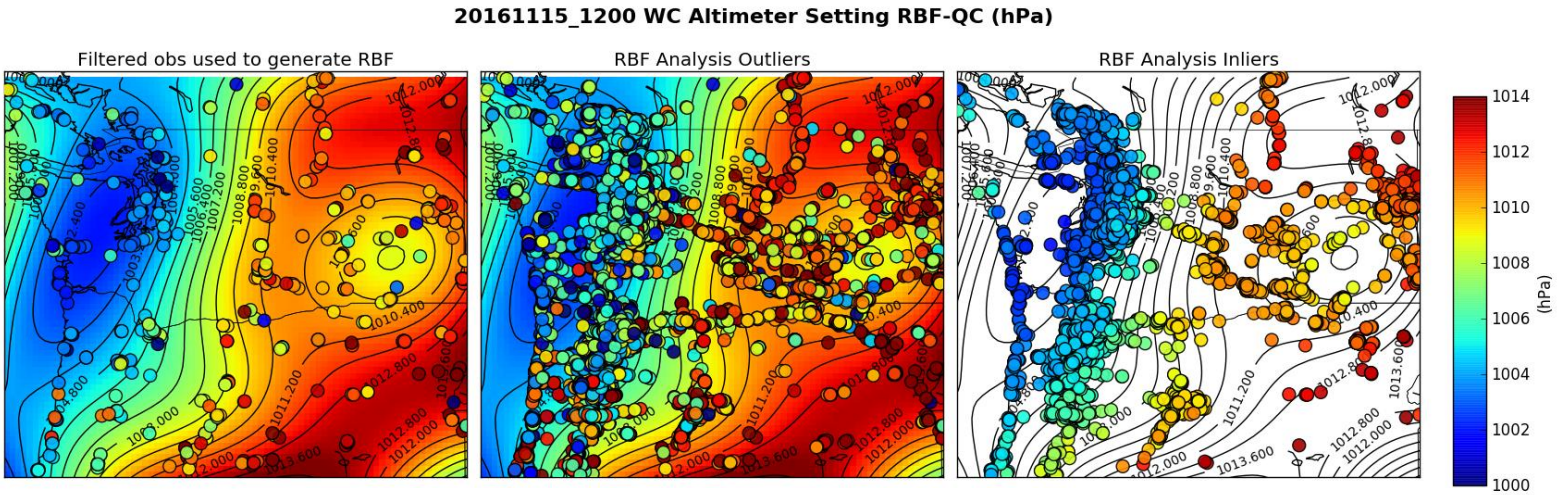


Figure 16: Example of the spatial filtered RBF check applied to WC observations.

produce a realistic altimeter analysis from the selection of spatially filtered observations. This is due to the limited number of outliers and the intrinsic smoothness of the RBF.

Since spatially filtering observations based on uncertainty may not always produce an optimum sample free of outliers, an alternative approach to spatial QC was developed. This approach utilized clustered observations to generate the RBF surface used to detect and remove outliers. In this approach, the random forest algorithm is applied to bias-corrected observations whose location fell within a previously defined cluster. These clustered observations are then used to generate an RBF analysis for outlier removal (Fig. 17)

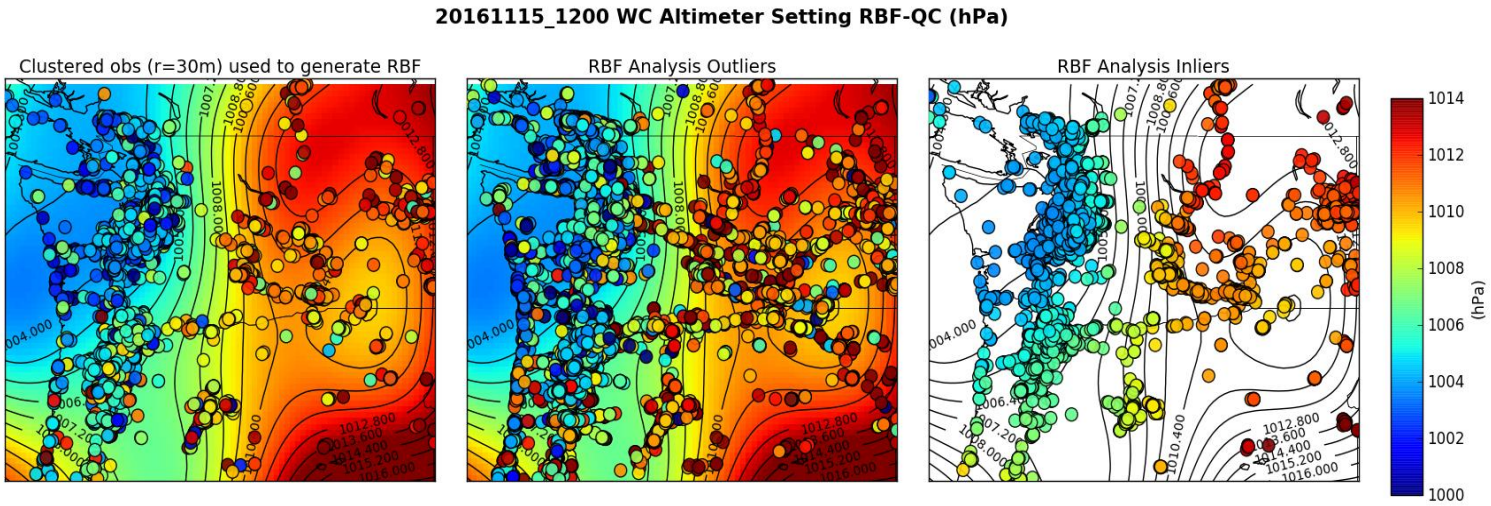


Figure 17: Example of the clustered RBF check applied to WC observations.

The observations used to generate the RBF surface were bias-corrected with random forests trained on observations clustered with DBSCAN. DBSCAN clustering was performed with a cluster radius of 30 meters. Like the spatial filtering technique, the cluster approach to spatial QC can produce a realistic altimeter analysis. One downside of the cluster approach is that fewer rural observations contribute to the RBF fit. Consequently, not as many observations are retained in rural regions like northeastern Oregon. Since Figures 15-17 display bias-corrected QC observations from the same date/time, a comparison can be made. Compared to the *uWx* TPS

analysis, both the spatially filtered RBF analysis and the clustered RBF analysis underestimate the north-south pressure gradient, west of the Cascades. The clustered RBF analysis is more consistent with the *uWx* analysis in northwestern Washington, while the spatially filtered RBF analysis overestimates the east-west pressure gradient across the cascades. In this case, the clustered RBF analysis appears slightly better than the spatially filtered RBF, due to its closer resemblance to the higher quality *uWx* analysis.

3.4.3) *Post Processing Performance*

uWx altimeter analyses, at each stage of post-processing, are displayed in Figure 18. The post-processing procedure begins with bias correction, simple statistical checks, and density-dependent spatial consistency checks. Random forests are only trained on phones that have submitted at least fifty observations. Out of all the phones in the *uWx* network, less than 15% fail to meet this requirement for bias correction. Overall, the random forests do a remarkable job at predicting and correcting *uWx* phone bias (Fig. 18).

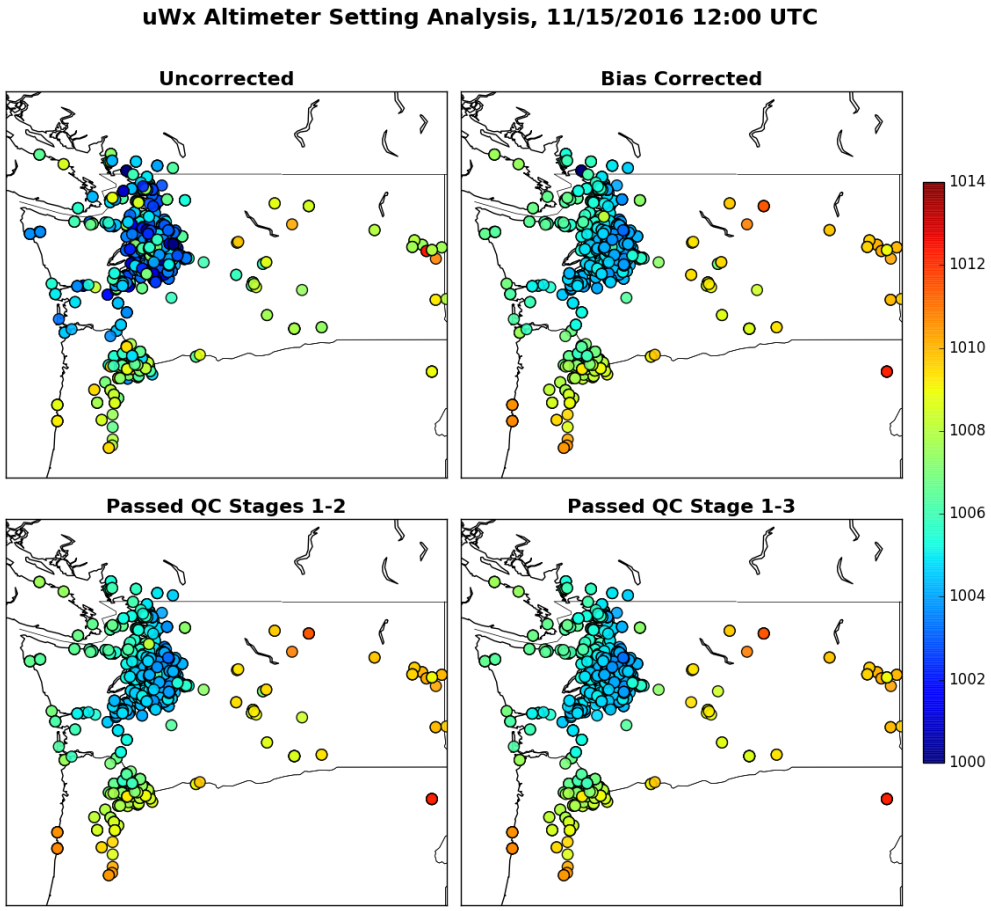


Figure 18: *uWx* altimeter setting at different stages of post-processing.

Even before QC checks are applied, bias-corrected *uWx* observations already appear self-consistent and coherent. Table 2, provides a summary of post-processing statistics, for the case displayed in figure 18.

uWx	% of all observations	% of bias-corrected observations
Ineligible for Bias Correction (insufficient training data)	14.08	N/A
Failed Validity Check	0	0
Failed Statistical Check	0.39	0.45
Failed RBF Check	13.8	16.06
Completed Post-Processing and passed all QC checks	71.73	83.48

Table 2: Summary of *uWx* post processing statistics for observations from 11-12 UTC Nov 15, 2016

When statistical and validity checks were applied, fewer than 1% of observations were removed. When the TPS spatial check was applied approximately 16% of bias-corrected observations were rejected as outliers. The vast majority of observations rejected by the QC framework were located in urban centers where sufficient observational density already exists. This is why the spatial extent and density of *uWx* observations do not appear to significantly change before and after post-processing (Fig. 15). In total, over 80% of bias-corrected observations passed all three stages of QC.

WC altimeter analyses, at each stage of post-processing, are displayed in Figure 19. dataset. In contrast to the *uWx* dataset, the random forest bias correction does not perform well

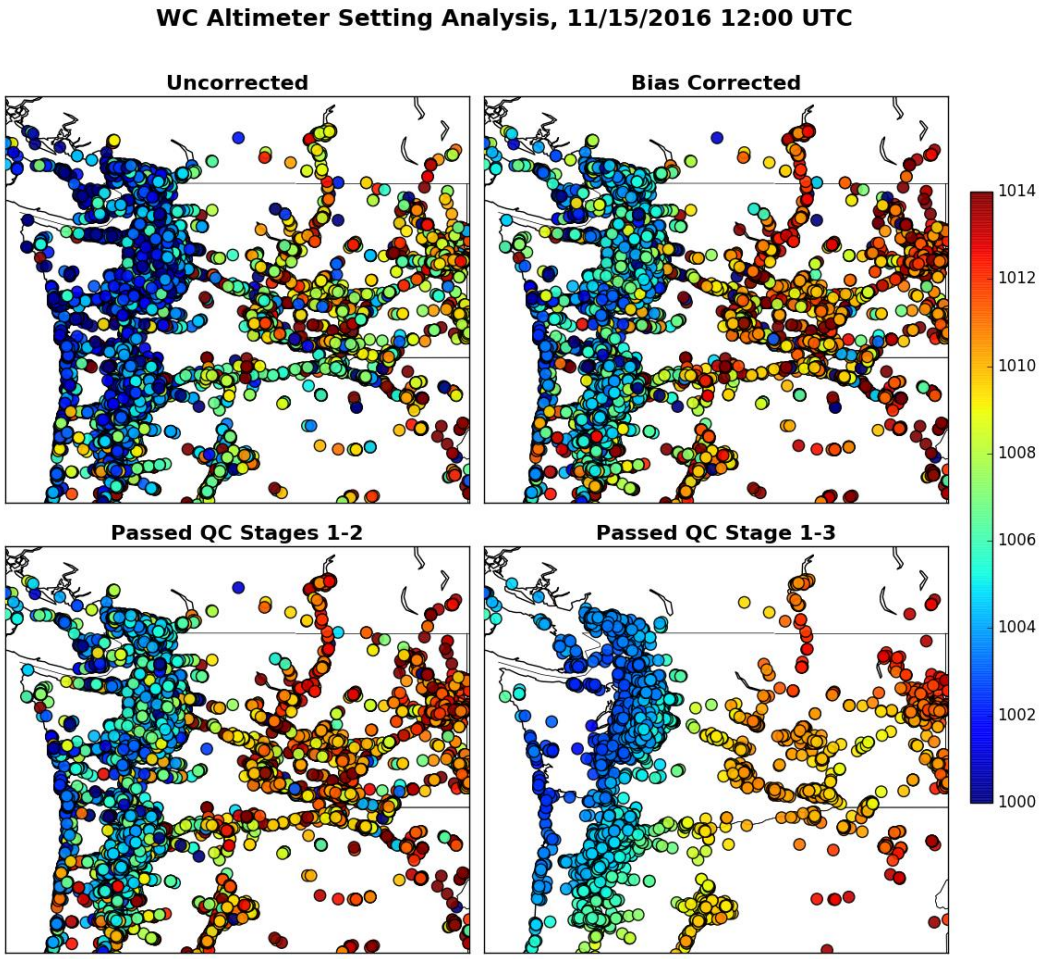


Figure 19: WC altimeter setting at different stages of post-processing.

on WC data. A lack of consistency between observations remains after bias correction. Further statistical and RBF-QC checks do a satisfactory job of eliminating discontinuities in the WC altimeter setting analysis. However, this comes at a significant price. Table 3, provides a summary of post-processing statistics, for the case displayed in figure 19.

WC	% of all observations	% of bias-corrected observations
Ineligible for Bias Correction (insufficient training data)	22.69	N/A
Failed Validity Check	0.09	0.11
Failed Statistical Check	0.47	0.6
Failed RBF Check	49.58	64.12
Completed Post-Processing and passed all QC checks	27.18	35.16

Table 3: Summary of WC post processing statistics for observations from 11-12 UTC Nov 15, 2016

Over 60% of bias-corrected observations are removed by the spatially filtered RBF check, with less than 1% removed by statistical and validity checks. The majority of the removed observations are in urban areas, however, a notable amount are also removed from roadways and more rural areas. The limited frequency of observation retrieval also contributes to the reduction of viable observations. Recall that bias correction was only performed on phones that had collected at least fifty observations. Over 20 % of WC phones failed to meet the criterion for bias correction. In total, only 35% of bias-corrected WC observations are included in the final altimeter setting analysis.

While post-processing bias-corrected observations can produce realistic altimeter analyses, it remains unclear whether these analyses are accurate. To evaluate the accuracy of the post-processed altimeter analysis displayed in Fig 18-19, a comparison was made between conventional in-situ observations and phone altimeter observations (Fig. 20).

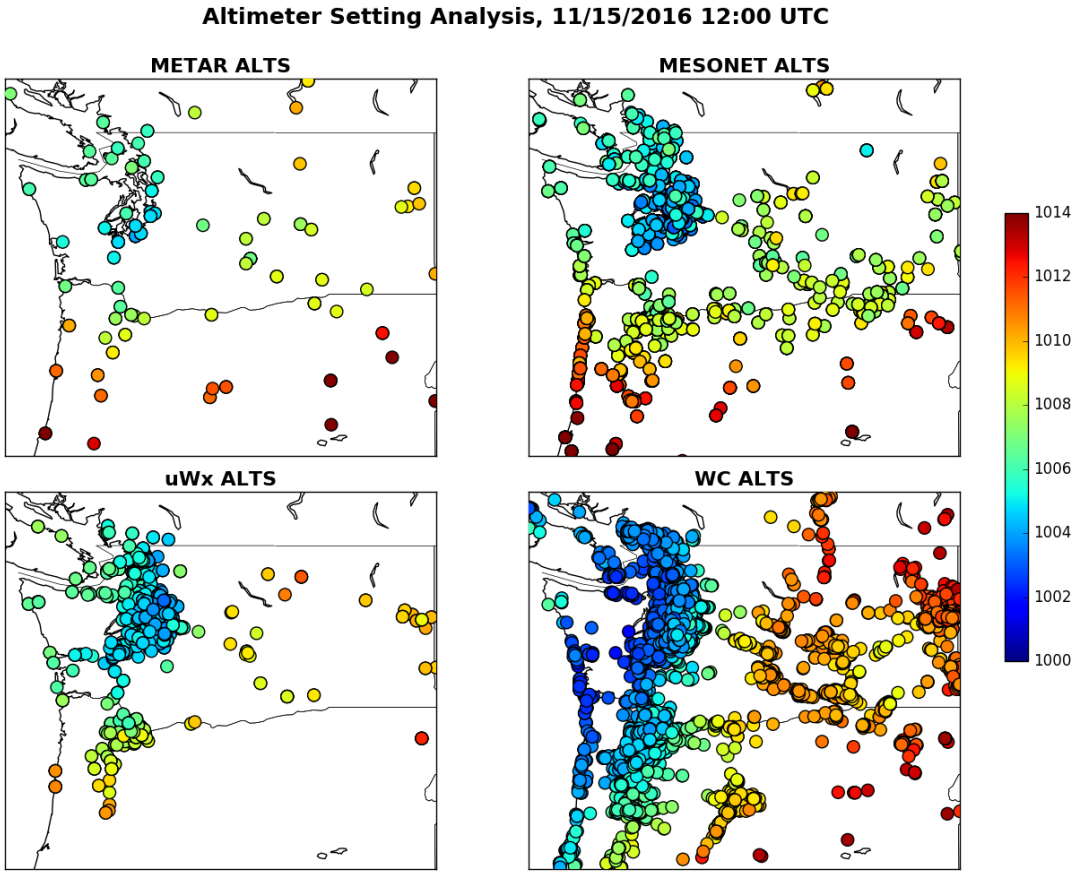


Figure 20: Comparison between phone altimeter and conventional altimeter observations

The *uWx* analysis compares well with both the METAR and MESONET analysis. The magnitudes of both the cross and along-barrier pressure gradients are roughly equivalent in the *uWx*, METAR, and MESONET analyses. The *uWx* analysis, like the METAR analysis, is smoother than the MESONET analysis. A less homogeneous *uWx* analysis could be generated by relaxing the thresholds used to detect outliers in the RBF check. The only notable difference between the *uWx* analysis and the analyses from conventional observations is a slight high bias in observations to the east of the Cascades in central Washington. Compared to the *uWx* analysis the WC analysis performs poorly. The along-barrier pressure gradient is underestimated while the cross-barrier pressure gradient is overestimated in the WC analysis. Altimeter observations are low biased west of the Cascades and high biased east of the Cascades. The low in

northwestern Washington is also too deep in the WC analysis. The poor quality of the WC analysis can be traced back to the inability of random forests to constrain large bias prediction errors. In RBF spatial checks the majority of WC observations are rejected. For this reason, bias-corrected WC observations cannot produce meaningful altimeter analysis. Spatial consistency checks are only effective when the majority of observations are accurate. This is not the case for WC observations. More work must be done to reduce bias prediction error and improve consistency among observations so that spatial checks can be effective.

3.4.4) *QC of Moving Phones*

One of the more novel aspects of random forests is their ability to accurately predict phone bias even for phones in motion. Figure 21, displays post-processed *uWx* observations from smartphones moving at speeds exceeding five meters per second (derived from GPS data).

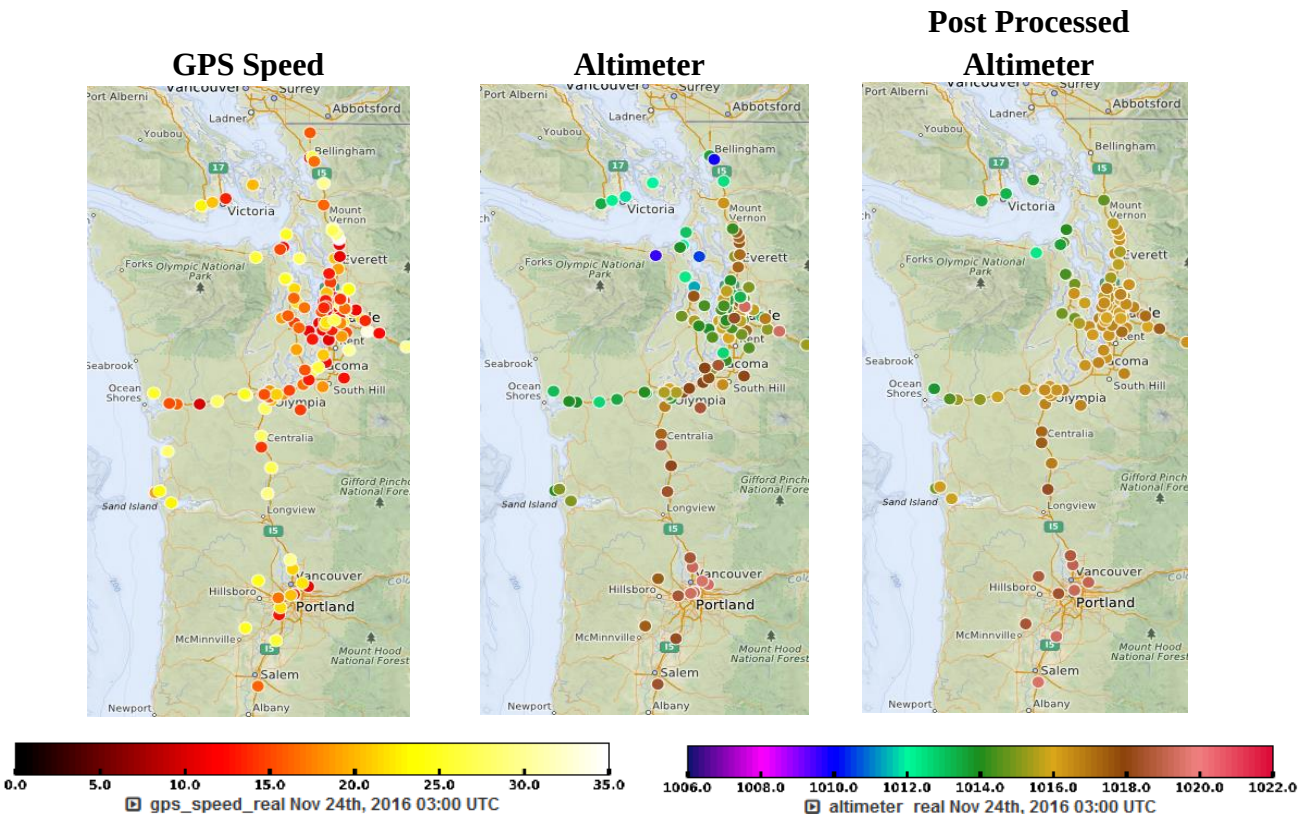


Figure 21: *uWx* altimeter setting analysis for phones moving at speeds exceeding 5 m/s.

Even for phones in motion, the random forest performs remarkably well. This may be a consequence of the random forest algorithm identifying weak relationships between phone bias, GPS speed, and elevation variance. Bias correction of moving phones is also likely aided by an algorithm built into *uWx* which modifies pressure retrieval when the GPS speed of the phone exceeds walking/running speeds. When the phone is moving at a significant pace, the pressure sensor is run for a few seconds after the GPS location has been retrieved. The delay in pressure collection is proportional to the speed of the phone. Pressure collection is delayed for phones traveling at vehicular speeds to ensure that the retrieved pressure is representative of the actual pressure at the location recorded by the GPS.

Chapter 4

DART-WRF ENKF EXPERIMENTS WITH BIAS-CORRECTED SMARTPHONE PRESSURE OBSERVATIONS

4.1) Background

To examine the potential of smartphone pressures for NWP, potential case studies were evaluated during the Fall of 2016. During this period, the second Wind Forecast Improvement Project (WFIP2) was ongoing. WFIP2 is a Department of Energy project to improve short-term forecasts of wind energy generation in regions of complex terrain (Marquis et al, 2015). The WFIP2 field campaign, centered just east of the Columbia River Gorge (CRG), included a network of sonic detection and ranging (SODAR) instruments in addition to radar wind profilers. WFIP2 weather discussions and forecast verification tools were used to select an optimal case study for SPO experiments. Since the number of phones in the *uWx* network doubled in mid-October, an initial case study in mid-November was selected.

In this event, a surface low developed off the coast of Oregon before tracking northeastward from the Columbia River delta into north-central Washington. Accompanying the surface cyclone was a warm front, which passed through the CRG region between 10-12 UTC on November 15th. A subsequent cold front passed through the same region five hours later, between 1500-1700 UTC. Following the passage of the cold front, wind energy generation by the Bonneville Power Administration (BPA) climbed from near zero to 4000 Megawatts (MW) as westerly winds picked up behind the front (Fig. 22). West of the Cascades, widespread light to moderate precipitation persisted until around 1300 UTC when heavier banded precipitation

began. The surface low achieved its maximum strength just after making landfall around 1200UTC.

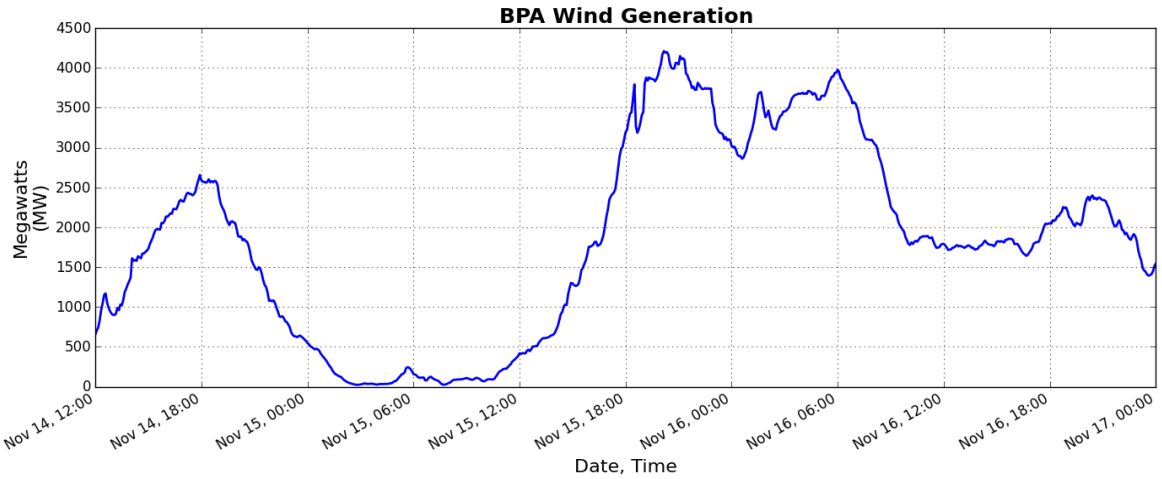


Figure 22: Time series of wind energy production from the Bonneville Power administration during the first case study (1200 UTC Nov, 14 – 0000 UTC Nov, 17).

This case was selected as the operational NOAA/NWS High-Resolution Rapid Refresh (HRRR) failed to capture the timing and strength of the wind-ramp event in the WFIP2 study region. Figure 23, shows a time-height cross-section from a SODAR near Rufus, Oregon. The HRRR brings strong winds down to the surface at 15 UTC, while the SODAR reveals that stronger winds did not occur in the rotor layer (80-150m) until 18 UTC. A similar two-hour timing error was also observed by a radar wind profiler at nearby Boardman, Oregon (not shown). Aside from the frontal timing issue, the HRRR, in general,

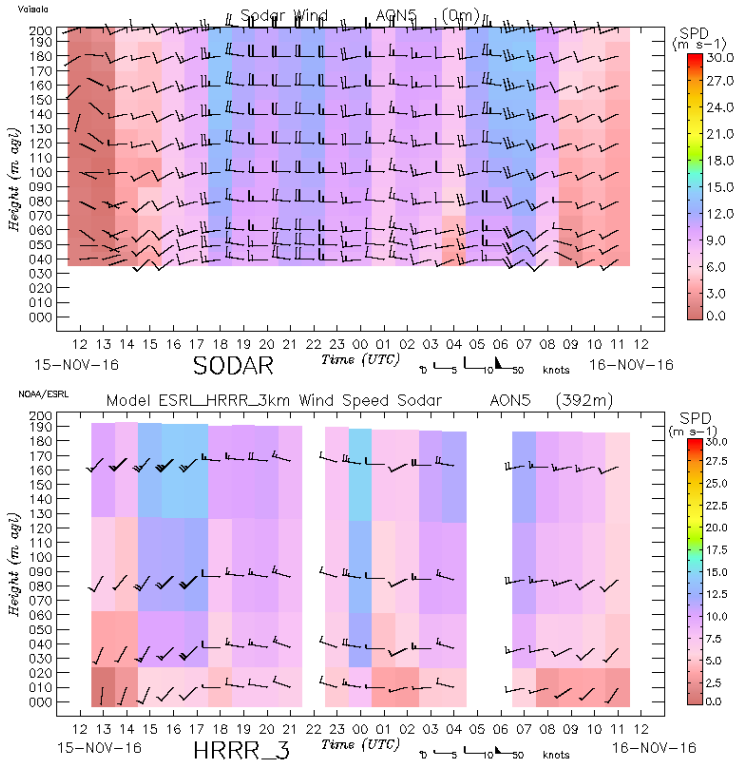


Figure 23: Time x Height cross-section of wind speed from a radar wind profiler (top) and from the 12Z HRRR (bottom) at Wasco, Oregon.

overestimates the strength of low-level winds. This is likely due to overmixing in the PBL. Failing to correctly time the onset of strong low-level winds poses a challenge to wind energy providers, who seek to anticipate when an up-ramp in wind energy production will occur. One goal of this study will be to determine whether SPOs can constrain near-surface wind forecasts during an up-ramp event of interest to wind-energy stakeholders.

A second case study unrelated to WFIP2 was examined in mid-October. As mentioned previously, it was around this time that the number of *uWx* users doubled from approximately 1000 to over 2000 active users. On October 12th, extended forecasts from the University of Washington's real-time Weather Research and Forecasting (UW-WRF) model suggested that the remnants of tropical storm Songda would intensify into a deep surface low over the next five days, potentially threatening the Pacific Northwest with a major wind storm. Initial forecasts from 0000 UTC on the 12th, placed a 955 hPa surface low offshore, a few hundred kilometers west of the Strait of Juan de Fuca. By the 13th, 3-day forecasts from the UW-WRF suggested that a more compact storm would make landfall near the center of the Olympic Peninsula. Since the 3-day track forecast placed the storm in an ideal location to generate strong winds in the densely-populated Puget Sound lowlands, media coverage of the storm increased.

With public awareness heightened by the threat of a major wind storm, a substantial effort was made to advertise *uWx* in an attempt to expand pressure collection ahead of the coming storm. On the morning of the 15th, the 4km 1200 UTC UW-WRF run forecast showed the surface low making landfall near Quillayute, Washington on the northwest coast of the Olympic Peninsula (Fig. 24). In this run, maximum wind gusts in the Seattle metro area were expected to exceed 50 mph. Due to the compact nature of the storm, the radius of strong winds was limited in extent. This complicated forecasts for cities like Seattle, which fell on the edge of

this radius. When the windstorm arrived between 00-0300 UTC on October 16, the surface low tracked offshore approximately sixty miles west of its forecast track. This tracking error resulted in substantial errors in wind forecasts in the Seattle area. Since the surface low took a more westward track, the strongest winds remained along the coast and the potentially devastating gale force gusts and damages to infrastructure never materialized.

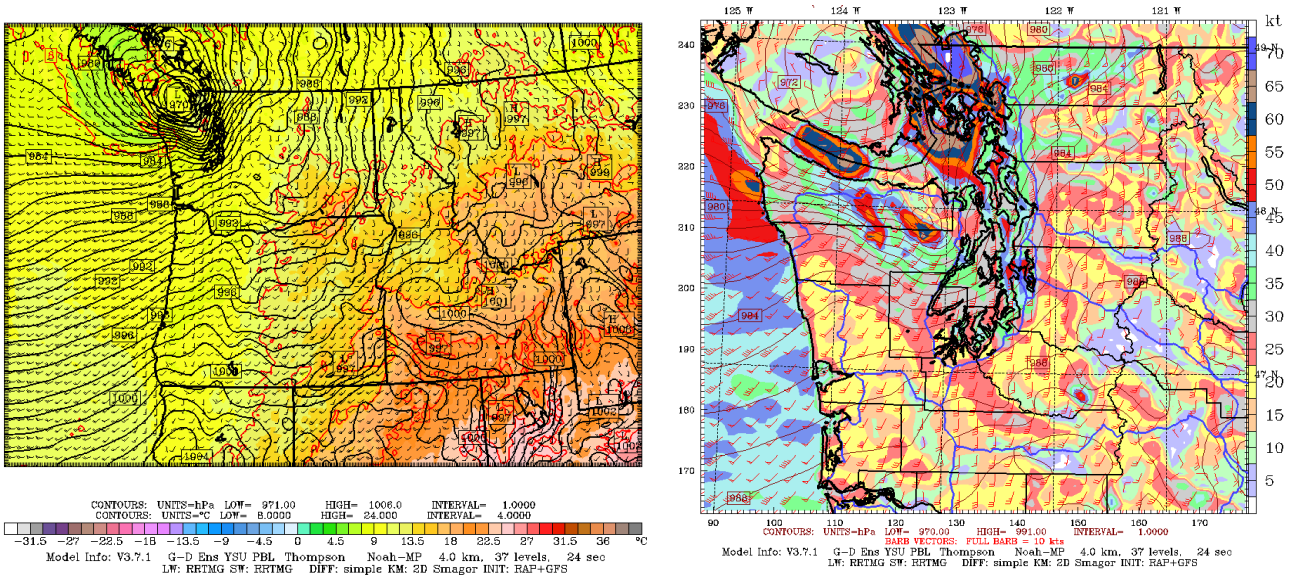


Figure 24: 15-h forecasts from the UW-WRF initialized at 1200 UTC on October 15th, 2016. MSLP, wind and temperature (displayed on the left), MLSP and 10 m Wind Gust (displayed on the right).

In anticipation of downed trees and widespread power outages, many Seattle residents opted to remain at home, resulting in lost revenue for stores, restaurants, and event holders who would have typically enjoyed good business on a Saturday night. If the forecast track of the windstorm could have been predicted with greater accuracy, earlier in the day, the economic impacts of this forecast bust may have been averted. To determine whether the large number of new smartphones added to the *uWx* network in the lead-up to this event, could have improved outcomes by better constraining the track and intensity of this windstorm, this event was selected as the second case study for review.

4.2) Methods

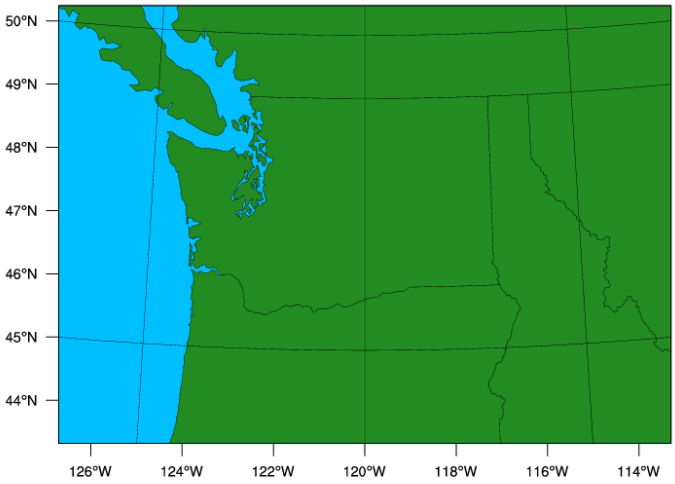


Figure 25: WRF-ENKF domain for both case

For both cases, assimilation experiments were executed with the Data Assimilation Research Testbed's (DART; Anderson et al., 2009) ensemble square root adjustment filter, and advanced with the Weather Research and Forecast (WRF) model (Skamarock, et al., 2008). Initial conditions were provided by the Rapid Refresh (RAP) model analysis, while hourly boundary conditions were generated with one-hour RAP forecasts. Physics parameterizations were set to mimic the operational HRRR model. WRF was run with a total of thirty-eight vertical levels and a horizontal resolution of four kilometers. The model domain consisted of a 1200 km x 900 km grid encompassing most of the PNW (Fig. 25). A total of 48 ensemble members were produced by allowing the Stochastic Kinetic-Energy Backscatter Scheme (SKEBS) to spin up perturbations in both WRF initial and boundary conditions (Berner et al., 2011). SKEBS parameters such as the total backscattered dissipation rate for stream function and potential temperature were set to $5x10^{-5}$ and $1x10^{-4}$, respectively. The decorrelation time scale for both stream function and the potential temperature was set to one hour. SKEBS was allowed to spin-up perturbations for twelve hours before the start of each case.

The first case study examined a frontal passage and subsequent wind ramp event on November 15th, 2016. For this case, both full cycling and no-cycling experiments were performed over sixty hours beginning at 1200 UTC on Nov. 14th and ending at 0000 UTC on Nov. 17^{th,} 2016. An initial control experiment was performed in which observations were used

for verification but not assimilated. In this run, the WRF ensemble was advanced one hour, every hour, with boundary conditions perturbed by SKEBS and provided by one-hour RAP forecasts. No-cycling experiments utilized previously generated 1-hr forecasts from the control ensemble to assimilate pressure observations. Since the model is not run after each assimilation step in the no-cycling runs, these experiments lack any memory of the assimilation and are less expensive to run. In full cycling experiments, the model is run after each assimilation step allowing the model to retain some memory of the assimilation. No-cycling and full-cycling experiments examined the impact of assimilating SPOs and conventional pressure observations from the MADIS network. In no-cycling experiments, the impact of assimilating pressure change was also examined. For this case, *uWx* SPOs were bias-corrected with random forests trained on data retrieved between August 15 – November 9, 2016. The root mean square bias prediction error was combined with estimates of location uncertainty and sensor noise to provide a unique estimate of observational uncertainty for every SPO.

The second case study examined a windstorm that evolved from the remnants of tropical storm Songda. For this case, both full cycling and partial-cycling experiments were performed over sixty hours beginning at 1200 UTC on Oct. 14th and ending at 0000 UTC on Oct. 17th, 2016. As in the first case, an initial control experiment was performed in which observations were evaluated but not assimilated. Two full cycling experiments examined the impact of assimilating SPOs and a combination of SPOs and SPCOs. In partial cycling experiments, the WRF ensemble was cycled until the start of a forecast run, during which the model was allowed to evolve freely. Two partial cycling experiments examined how the assimilation of SPOs impacted short-term forecasts of the extant and track of the windstorm. Partial cycling forecast runs were initialized at 1900 and 2100 UTC on Oct. 15th. In each forecast run, the WRF ensemble utilized SKEBS

perturbed RAP forecast boundary conditions. Both the 1900 and 2100 UTC runs produced ensemble forecasts out to 04 UTC on Oct. 16th. Since many SPOs collected during this case were retrieved from smartphones that had just joined the *uWx* network, bias correcting SPOs on past data was not feasible. As a result, SPOs assimilated in this case study were bias-corrected with random forests trained on data retrieved during the month after the event, between October 19 - November 23, 2016. Each SPO was assigned a unique observational uncertainty derived from the root mean square bias prediction error, location uncertainty, and sensor noise.

All of the aforementioned DART-WRF experiments were performed on the Microsoft Azure Cloud. Cloud computing resources were provided by Microsoft through a $20,000 Azure Research Grant. A 256-core, high-performance computing (HPC) cluster was built from sixteen H-Series virtual machine instances. Each virtual machine consisted of an intel processor with 16 cores, 7 GB of RAM per core, and a backend 32 GB/s InfiniBand network capable of Remote Data Memory Access (RDMA) ideal for low-latency high throughput communication. To the best of the author's knowledge, this study is the first to run DART on the cloud.

4.3) Results

4.3.1) Case I: No-Cycling Experiments

In previous work by Madaus and Mass (2017) a key finding was that no-cycling assimilation of SPOs degraded the analysis at a significant fraction of SPO locations. For comparison, a similar analysis was performed with *uWx* SPOs assimilated during no-cycling experiments. The difference between the ensemble mean posterior (analysis) error and the ensemble mean prior (background) error for all sixty assimilation steps was evaluated at the location of each assimilated observation to determine whether the assimilation of SPOs, SPCOs, and METAR altimeter setting improved the model analysis of the assimilated variable. Ideally,

assimilated observations should decrease the analysis error locally, ensuring that the ensemble mean state is nudged closer to the observations, producing an ensemble mean analysis with less error. In previous work, Madaus and Mass (2017) found that SPOs were unable to achieve this result over the entire domain. In their experiments, the analysis error was greater than the background error for ~45% of SPOs. In this study, a mere 13% of *uWx* SPOs increase the analysis error relative to the background error. Figure 26, highlights the change in analysis error relative to the background error in METAR and smartphone no-cycling assimilation

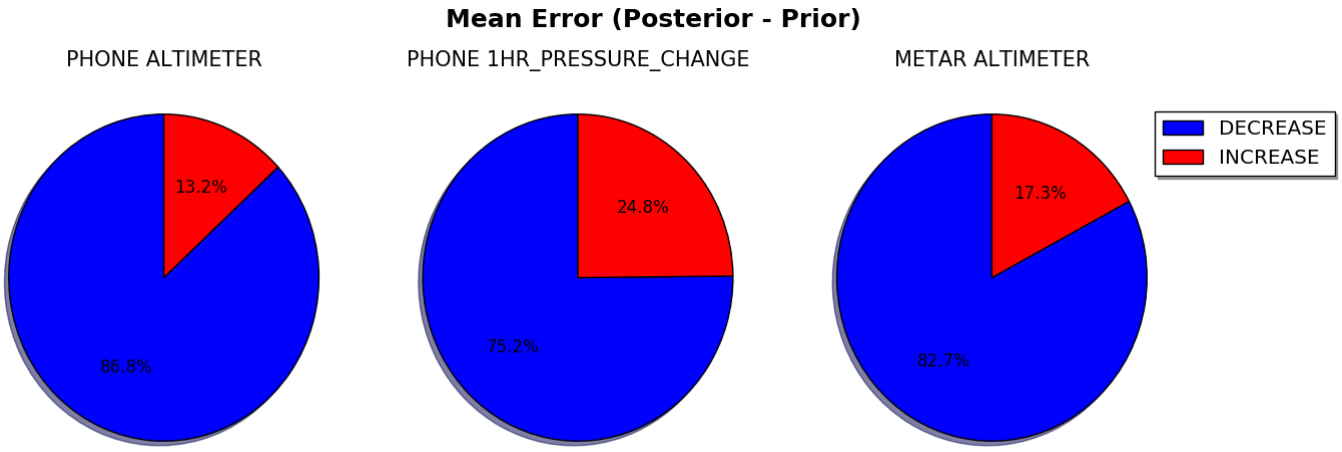


Figure 26: Mean Error difference between Posterior (analysis) and Prior (1h-forecast). If Posterior - Prior is negative (blue) assimilating X decreased the analysis error of X and vice versa (red).

experiments. The analysis error was less than the background error for ~82% of METAR pressures, ~87% of SPCOs, and ~75% of SPOs. This result is not entirely surprising given that qualitatively, smartphone pressure analyses appear self-consistent and coherent after bias correction with random forests and RBF-QC (Fig. 18).

While Figure 26 revealed that assimilating SPOs improved the analysis error at the plurality of SPO locations, it is unclear whether the same result would be achieved if the analysis was verified with high-quality METAR observations instead of SPOs. Figure 27, displays the

domain averaged time-series of altimeter analysis error, verified against METAR altimeter observations, for four unique no-cycling assimilation experiments. Relative to the background error (CNTRL) assimilating SPOs (PHONE_ALT) consistently reduces the analysis error at METAR locations by nearly 50%.

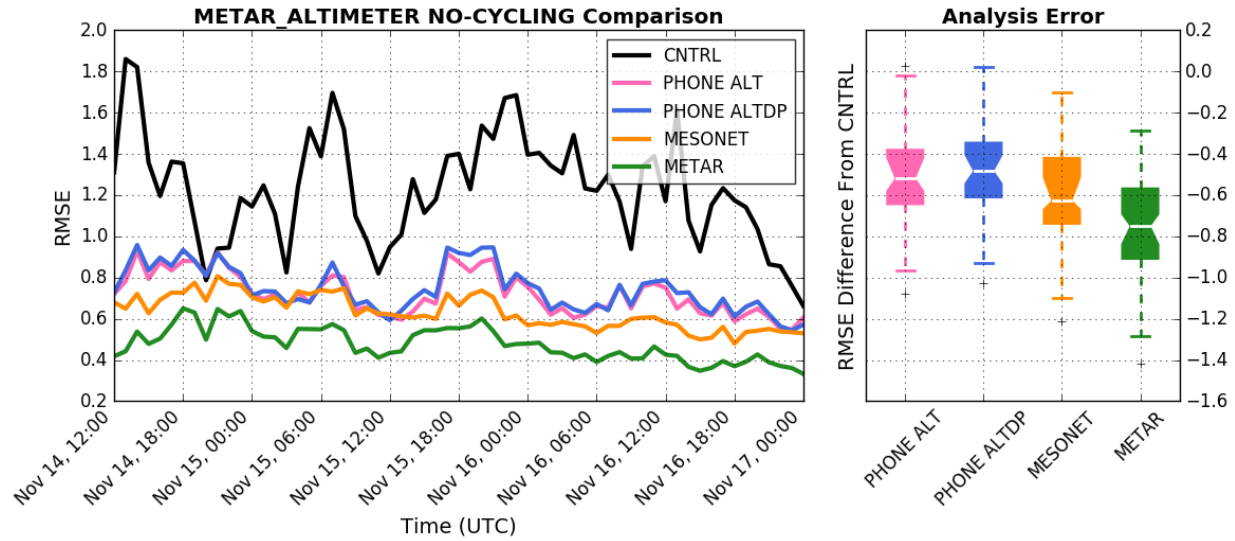
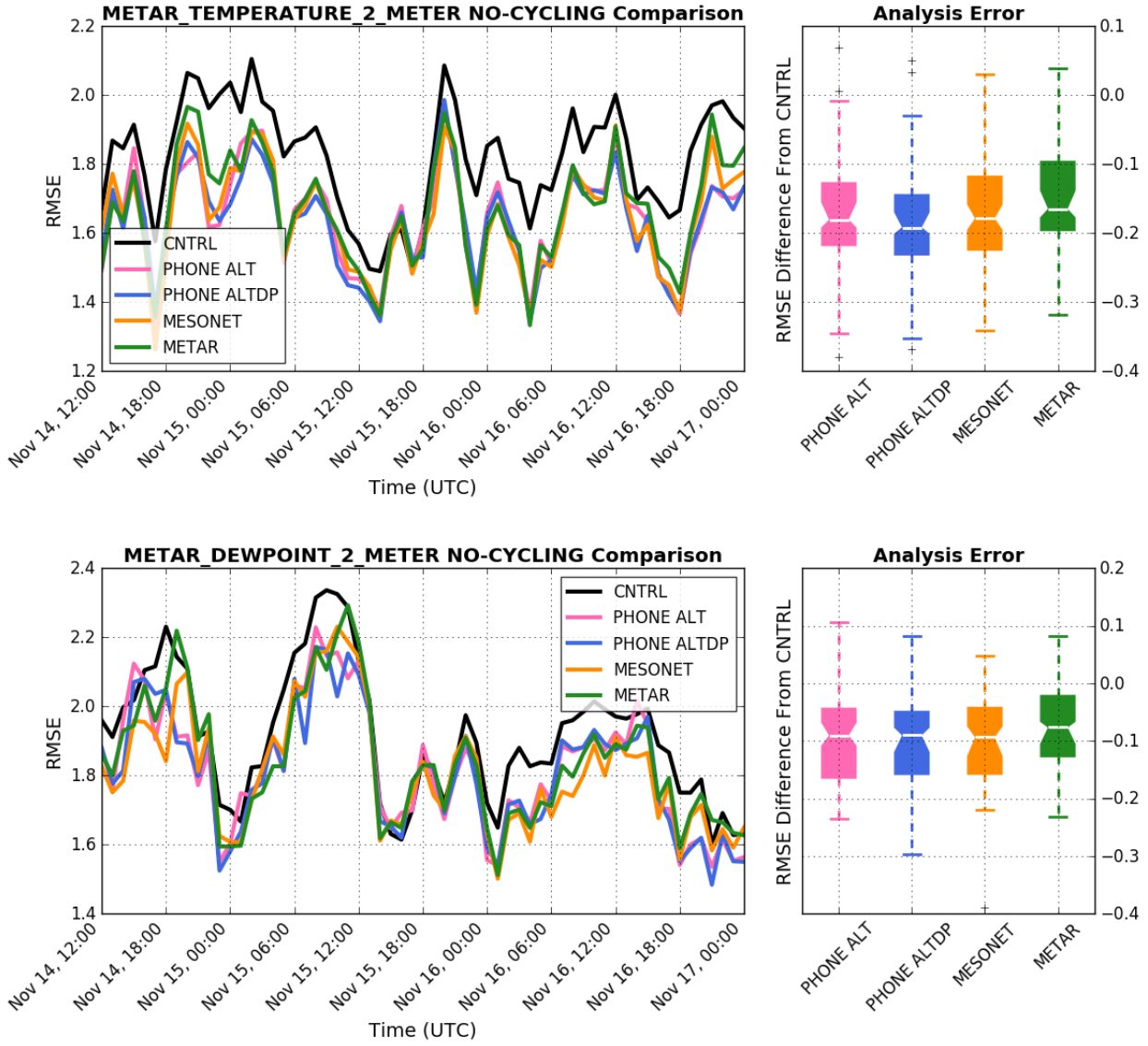


Figure 27: Time series of domain averaged altimeter analysis RMSE relative to the background altimeter RMSE from the control ensemble (no assimilation). METAR altimeter observations used for verification.

The median reduction in altimeter analysis error over the sixty hours was approximately 0.5 hPa. Assimilating 1hr SPOCs in addition to SPOs (PHONE_ALTDP) did not provide any added value, as reductions in analysis error were virtually equivalent to the PHONE_ALT assimilation experiment. Assimilating MESONET pressures provided a slight improvement over assimilating SPOs. This is not surprising as the spatial extent of MESONET observations is substantially greater than the spatial extent of SPOs in the eastern half of the model domain, east of the Cascade. When METAR pressures are assimilated, the greatest reductions in altimeter analysis error are achieved. Since verification is performed with METAR observations, it is expected that METAR pressures should have the biggest impact on the analysis at METAR locations.

To determine the effects of smartphone pressure assimilation on unobserved variables, the analysis error for 2-m temperature, 2-m dew point, and 10-m wind were examined. Figure 28, displays the time-series of temperature, dew point, and wind analysis error relative to the background (CNTRL) error.



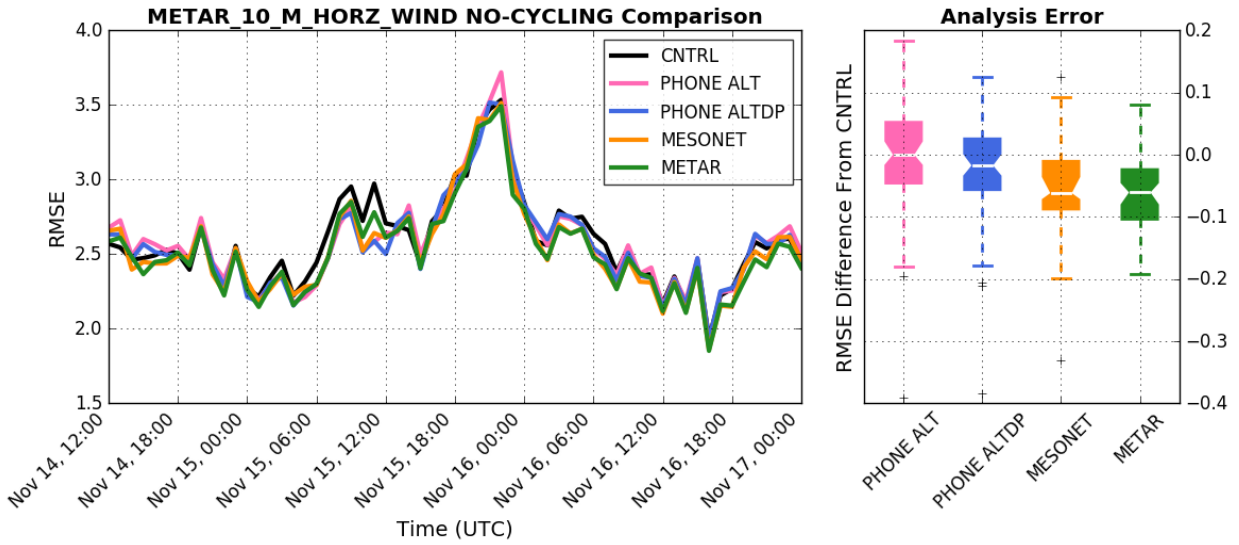


Figure 28: Time series of domain averaged 2-m temperature, 2-m dew point, and 10-m wind analysis RMSE relative to the background RMSE from the control ensemble (no assimilation). METAR observations used for verification.

Differences in error between the CNTRL and four unique assimilation experiments are also displayed (right panel). Assimilating SPOs (PHONE_ALT), consistently reduces dew point and temperature analyses errors by a small margin. The median reduction in analysis error for dew point and temperature is ~0.1 K and ~0.18 K respectively. No improvements are observed in 10-m wind analysis when SPOs are assimilated. Assimilating 1hr SPCOs in addition to SPOs does not produce any notable reductions in analysis error beyond that achieved by assimilating only SPOs. When MESONET and METAR pressures are assimilated individually, reductions in dew point and temperature analysis, similar in magnitude to the PHONE_ALT assimilation case, are observed. METAR pressure assimilation, and to a lesser degree MESONET pressure assimilation, does produce slight reductions in 10-meter wind analyses, but these reductions are small. The median wind analysis error reduction was a mere ~0.06 m/s when MESONET pressures were assimilated.

The ability of SPOs to constrain analysis of observed and unobserved variables suggests

that the quality of bias-corrected *uWx* SPOs may be comparable to existing MESONETs. SPOs can reproduce the reductions in analysis error achieved by assimilating traditional in-situ observations. This is a notable achievement considering that the vast majority of SPOs are concentrated in the western half of the domain. Only a few tens of SPOs are retrieved east of the Cascade crest during any given hour, nevertheless, significant reductions in domain averaged analysis error are achieved through SPO assimilation. This is partly due to the large covariance length scale of surface pressure which allows the SPOs in the western half of the domain to influence the pressure field downstream, east of the Cascades. Compared to MESONETs and METARs, the spatial distribution of SPO is largely inhomogeneous. The location of SPOs mimics population density to a greater degree than traditional in-situ pressure networks. This may explain why SPOs struggled to constrain wind analyses. Since the covariance between pressure and wind is often weak and the covariance length scale of wind is relatively small (< 100 km), improving the wind analyses often requires input from several pressure observations. While SPOs may have sufficient density in urban areas to constrain wind analyses, elsewhere SPOs are distributed like oases in a desert. This may explain why SPOs were less successful than MESONETs and METARs in constraining wind analyses.

4.3.2) Case I: Full Cycling Experiments

No-cycling experiments showed that SPOs were able to constrain analysis errors. To evaluate the impact of SPOs on forecasts, full-cycling experiments were examined. Figure 29, displays the time-series of domain averaged 1-hr altimeter forecast RMSE for three unique assimilation experiments relative to the background 1-hr forecast RMSE from the control ensemble. Compared to the METAR and MESONET pressure assimilation experiments, the PHONE_ALT experiment which assimilated only SPOs does remarkably well.

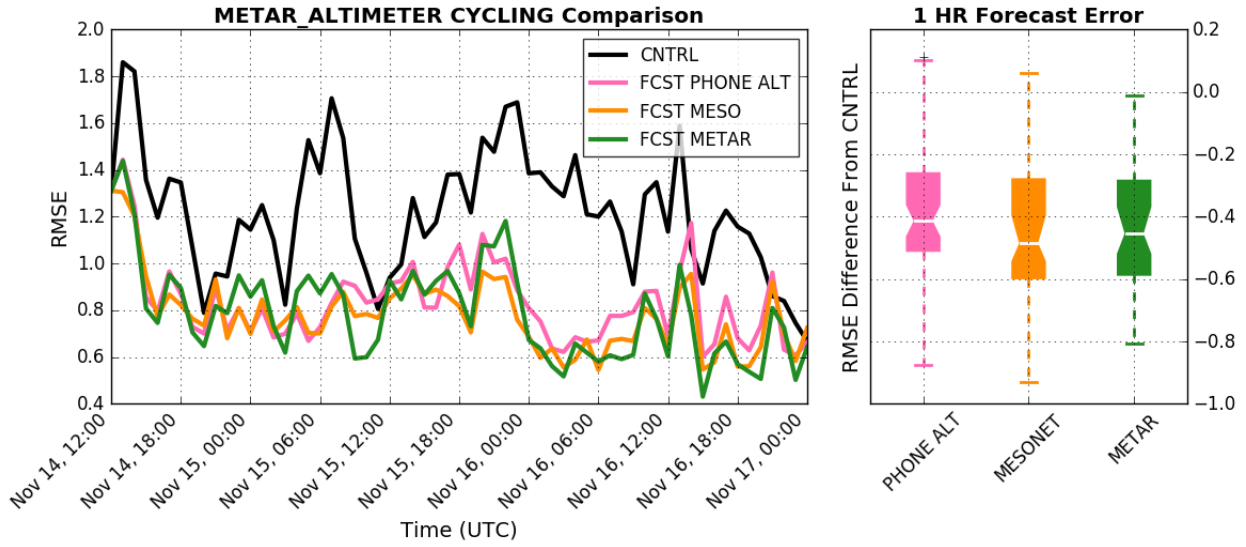
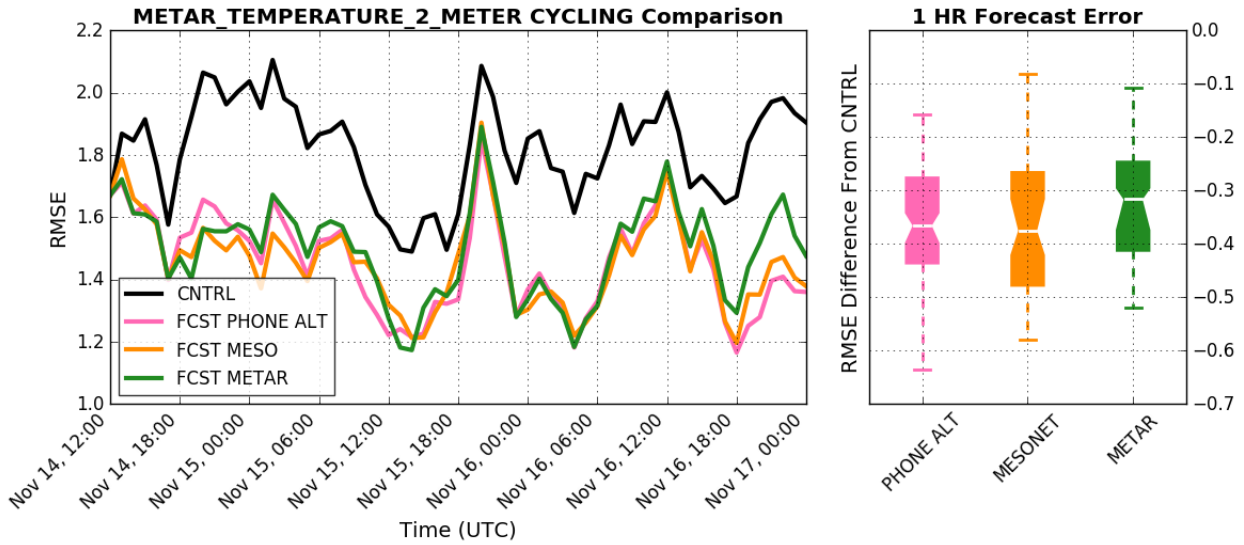


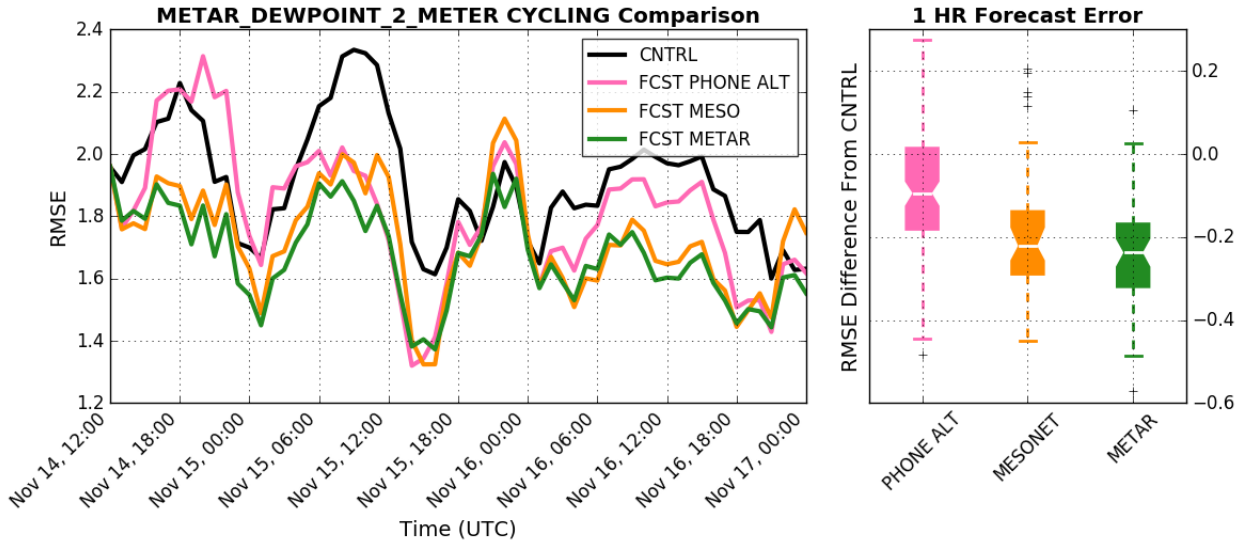
Figure 29: Time series of domain averaged altimeter 1-hr forecast RMSE relative to the background 1-hr forecast RMSE from the control ensemble (no assimilation). METAR altimeter observations used for verification.

Assimilating SPOs consistently reduces 1-hr forecast errors, to a magnitude similar to that observed when MESONET or METAR observations are assimilated. The right panel of figure 29, shows the difference in 1-hr forecast RMSE between each assimilation experiment and the control. In this panel, bootstrapped 95% confidence intervals for the median are displayed as notches in the boxplots. Since the boxplot notches overlap for all three cases, the median reduction in 1-hr forecast RMSE of altimeter setting observed in the PHONE_ALT experiment is not statistically significantly different than the median reduction observed in the METAR and MESONET experiments.

In no-cycling experiments, SPO assimilation was able to constrain analyses error for temperature and dew point but not wind. To determine whether improvements in these analyses were conveyed into forecasts, 1-hr forecasts of temperature, dew point, and wind were compared for three full-cycling assimilation experiments, in which pressures observations from smartphones (PHONE_ALT), METARs, and MESONETs were assimilated. Figure 30, displays the domain averaged 1-hr forecast RMSE for temperature, dew point, and wind relative to the

background control ensemble 1-hr forecast RMSE. The right panel displays the difference in the 1-hr forecast RMSE between each assimilation experiment and the control. SPOs can improve 1-hr temperature forecasts, and to a lesser degree 1-hr dew point forecasts. Overall, SPOs slightly degrade the performance of 1-hr wind forecasts.





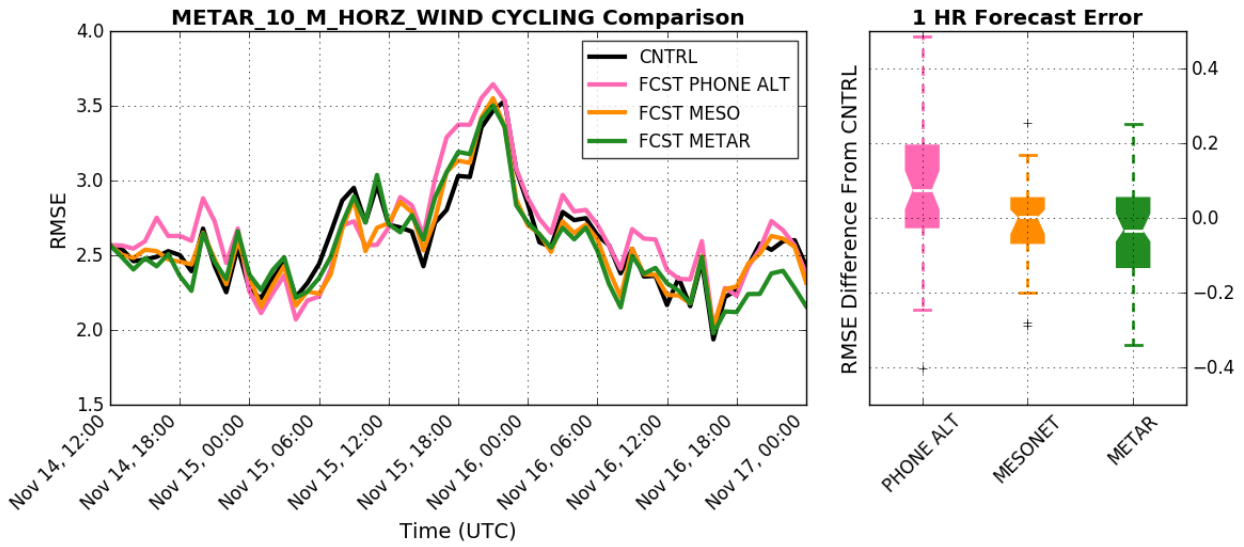


Figure 30: Time series of domain averaged 2-m temperature, 2-m dew point, and 10-m wind 1-hr forecast RMSE relative to the background RMSE from the control ensemble (no assimilation). METAR observations used for verification.

The ability of SPOs to constrain 1-hr forecasts of temperature is comparable to that of MESONET and METAR pressures. MESONET pressure assimilation produces slightly larger reductions in 1-hr forecasts of dew point relative to the PHONE_ALT case. On average, assimilating pressure from any observing network does little to improve or degrade 10-m wind forecasts. Even when METAR wind forecasts are evaluated at the same location that METAR pressures were assimilated an hour prior, no improvement in 1-hr wind forecasts is observed.

It is telling that a modest smartphone network capable of assimilating just over a thousand pressures an hour, from a dense network of smartphones can nearly match the performance of pressure assimilation from existing in-situ pressure networks. In this case, SPOs were able to constrain both analysis and forecast errors of pressure, temperature, and to a lesser degree moisture. In the study region, there are likely well over a million smartphones capable of retrieving pressure. This would imply that in the experiments shown previously, less than one-tenth of one percent of potential SPOs were assimilated. Since a little over a thousand *uWx* SPOs

were able to match the performance of MESONET pressures in constraining forecasts of pressure and temperature it is natural to ponder whether a similar smartphone network with 10x, 100x or even 1000x the number of smartphones could provide additional forecasts improvements beyond that achieved by assimilating pressures from existing MESONET or METAR networks. Nearly 10x as many SPOs were utilized by Madaus and Mass (2017) in their study but over a much larger area; they found that full-cycling SPO assimilation degraded temperature forecasts and produced only marginal improvements in altimeter forecasts. This serves as a reminder that quality is more important than quantity. If SPOs are not bias-corrected and quality controlled, a lack of self-consistency among observations can inhibit EnKF assimilation by generating spurious analysis increments which hinder improvements in the ensemble mean analysis.

The fact that SPOs were unable to constrain wind analysis and forecasts is expected as previous work by Madaus et al., (2014) found that assimilating high-density pressure observations had no impact on 3-hr, 10-m wind forecasts in the Pacific Northwest. Nevertheless, one reason this case was selected was due to its importance for wind energy. Domain averaged 10-m wind forecast errors slightly increased on average after assimilation of SPOs. This does not imply that SPOs degraded wind forecasts throughout the entire domain. To evaluate how SPO assimilation impacted wind forecasts during the wind ramp event, 1-hr wind forecasts were temporally averaged over the six hours between 1200 and 1800 UTC on Nov. 18th, 2016. The average difference between the PHONE_ALT 1-hr wind forecasts and 1-hr forecasts from the control ensemble over this period is displayed in figure 31.

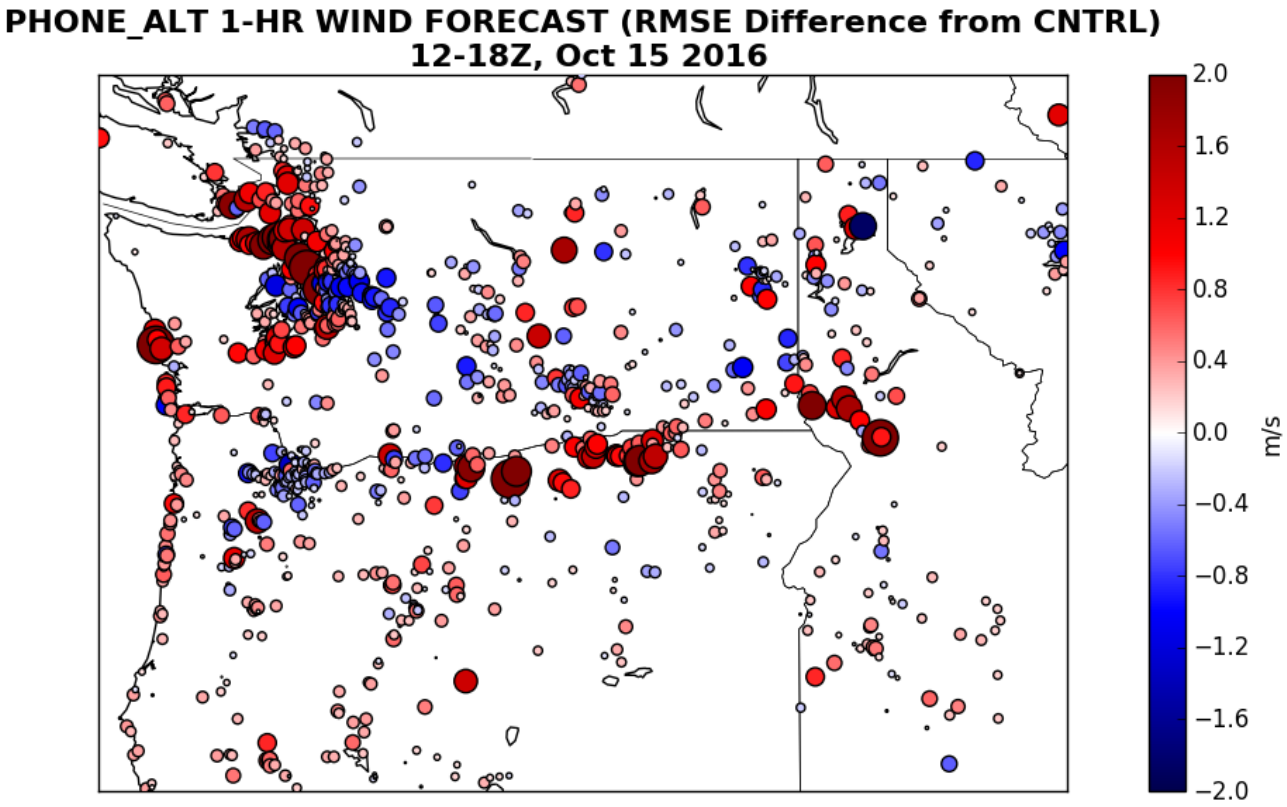


Figure 31: 1-hr wind forecast RMSE difference between the PHONE_ALT experiment and CNTRL experiment, averaged over the 6-hour period spanning 12-18Z, on Oct. 15, 2016. Verification was performed with MESONET observations.

To improve the readability of the figure, the size of each circle in the scatter plot was scaled linearly with the RMSE difference. 1-hr wind forecasts were verified with MESONET observations to get a better appreciation for how SPOs impact wind forecasts throughout the domain, especially in locations where SPO density is high. During the period examined in Figure 31, a frontal passage swept through the Columbia River Gorge resulting in a wind ramp event (Fig. 22). In the domain average, 1-hr wind forecast performance decreased when SPOs were assimilated during this period. Figure 31, reveals that changes in the performance of 1-hr wind forecasts are not uniform throughout the domain, with 1-hr wind forecast improvements appearing where the density of assimilated SPOs is high. Wind forecast improvements are observed in the Seattle and Portland metro areas, two locations where SPOs are abundant. The

RMSE difference downstream of the Columbia River Gorge, where wind energy assets are located, varies before becoming positive east of Boardman, Oregon.

To determine whether SPOs may have aided wind-energy stakeholders by reducing wind forecasts error in the near-surface layer, vertical wind profiles from WFIP2 SODARs were used to verify low-level wind forecasts during the six-hour (12-18UTC) wind ramp event. During this period, profiles of wind forecast RMSE between each assimilation experiment and the CNTRL experiment were temporally averaged. Figure 32, displays the location of wind farms and WFIP2 SODAR assets used to verify ensemble mean vertical wind profiles.

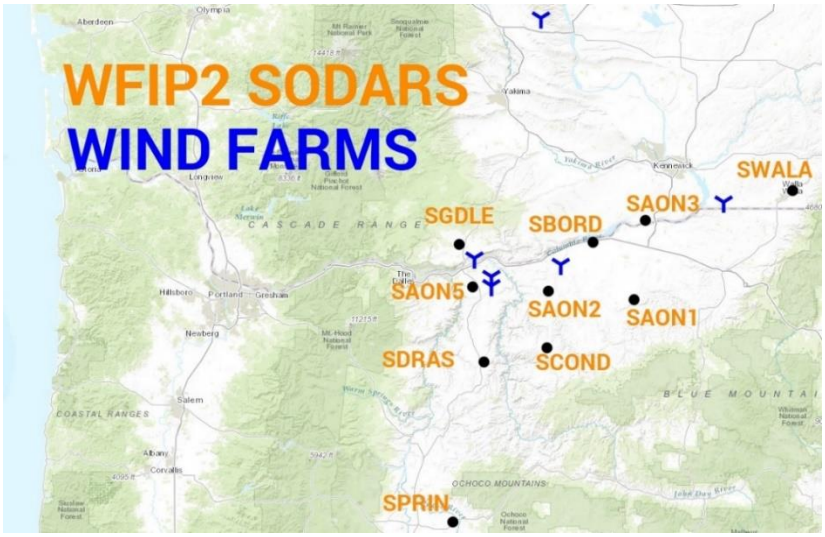


Figure 32: Location of wind forms and SODAR assets in Washington and Oregon.

The RMSE difference between vertical profiles of 1-hr wind forecasts for each assimilation experiment and the CNTRL experiment are displayed in figure 33.

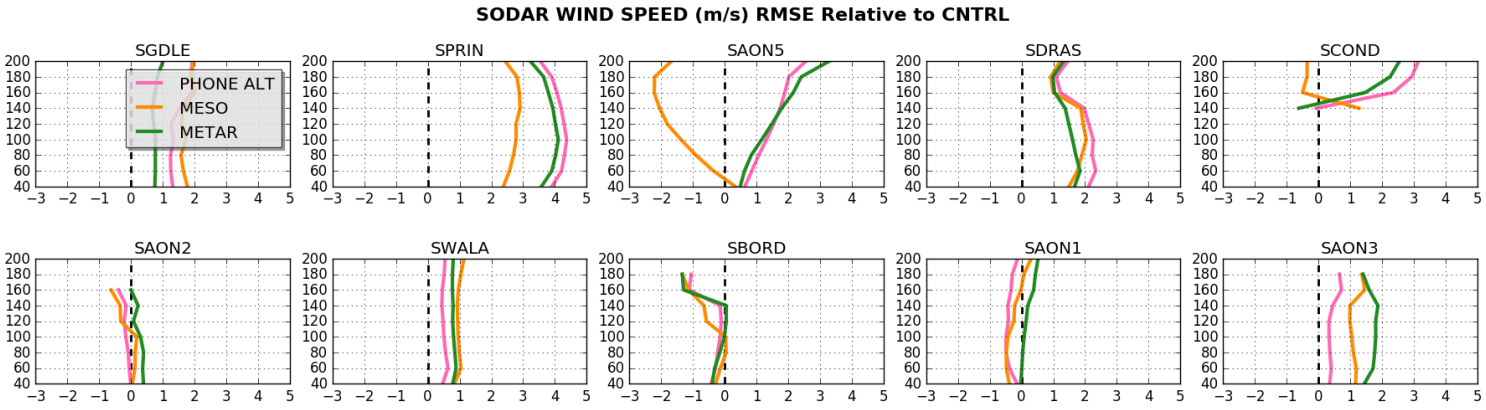


Figure 33: 1-hr wind profile forecast RMSE difference between assimilation experiments and the CNTRL experiment, averaged over the 6-hour period spanning 12-18Z, on Oct. 15 2016. Verification was performed with SODAR observations from WFIP2.

At three of the ten SODAR verification sites (AON1, AON2, and BORD) the assimilation of SPOs improved wind forecasts throughout the near-surface layer. These three sites are co-located south of the Columbia River between Condon, Oregon, and Hermiston, Washington. At all other locations assimilating SPOs increased the RMSE of near-surface wind forecasts. Except for AON5, assimilating MESONET pressures has a similar effect on wind forecast RMSE in the near-surface layer. When only METAR pressures are assimilated, wind forecasts RMSE at virtually all SODAR increases relative to the control experiment. In this case, pressure assimilation tended to have a detrimental effect on near-surface winds in the WFIP2 study region (Fig 32). A challenge often observed in regions of complex terrain is overmixing. When this occurs, the downward momentum flux is too large, resulting in larger than observed wind speeds in the PBL. In all pressure assimilation experiments the cross-barrier pressure gradient across the CRG was increased, resulting in an increase in the magnitude of near-surface winds within the WFIP2 study region. At most SODAR sites near-surface wind speeds were over-forecast. By increasing the cross-barrier flow, pressure assimilation likely enhanced overmixing in the PBL, resulting in near-surface wind errors that were larger than observed in the CNTRL experiment.

4.3.3) Case II: Full Cycling Experiments

In the previous case study, the covariance between wind and pressure was relatively weak and improvements to wind forecasts were mostly limited to urban centers where the density of SPOs was high. Since this case study focuses on a major PNW windstorm, the signal-to-noise ratio is considerably lower and the relationship between wind and pressure is much more pronounced. For this reason, it is not unreasonable to expect that if pressure forecasts can be constrained wind forecasts may also be constrained.

To determine the impact of SPOs on pressure forecasts, time series of 1-hr altimeter forecast RMSE is displayed in figure 34.

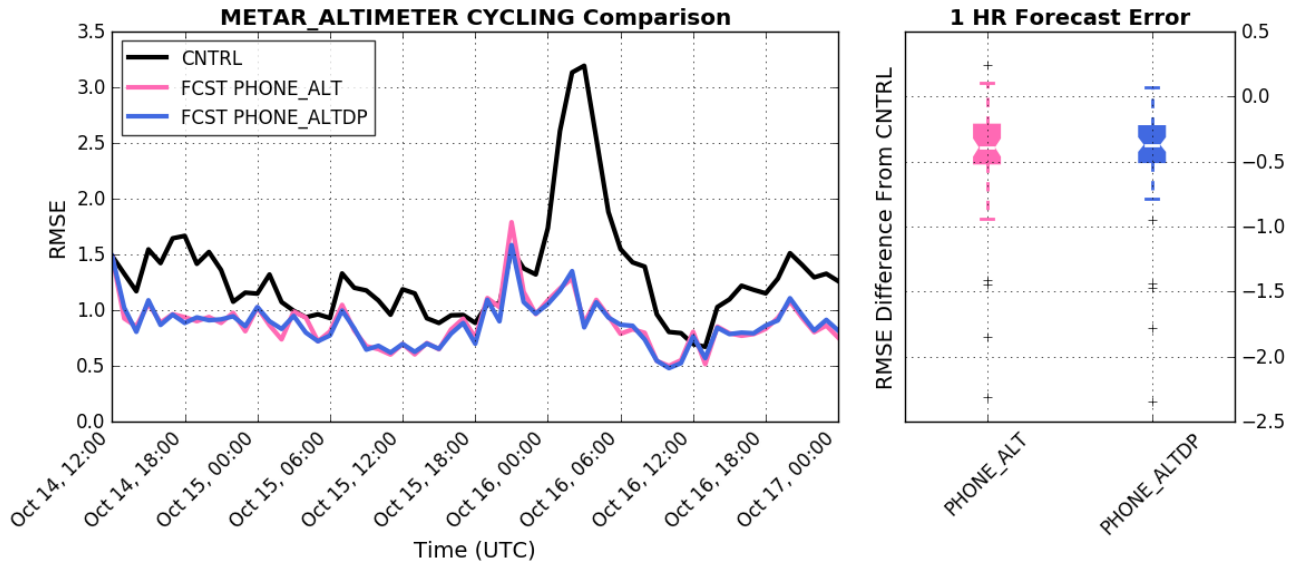


Figure 34: Time series of domain averaged 1-hr altimeter forecast RMSE relative to the background 1-hr forecast RMSE from the control ensemble (no assimilation). METAR altimeter observations used for verification.

The right panel of this figure shows 1-hr forecast RMSE differences from the CNTRL ensemble experiment, during which no assimilation was performed. For both PHONE_ALT (SPO only assimilation) and PHONE_ALTDP (SPO and 1h SPCO assimilation), the median reduction in 1-hr altimeter forecasts was approximately ~0.4 hPa. At 21UTC on Oct 15th, the assimilation of SPCOs in addition to SPOs appears to have constrained domain averaged altimeter forecast errors relative to the PHONE_ALT case. The largest reductions in domain averaged 1-hr altimeter forecast error occur during the period when the wind storm made landfall, between 00Z and 06Z on Oct. 16th. At 02 and 03UTC, the domain averaged 1-hr forecast altimeter RMSE decreased by 1.8 hPa and 2.3 hPa, respectively, when SPOs were assimilated. The magnitude of these reductions suggests that as the windstorm approached and made landfall SPOs had a major influence on the modeled track and intensity of the windstorm.

In this case study, the most important forecast variable is wind. The intense winds forecast for the Puget Sound lowlands late in the day on Oct. 15th had a major public impact. To evaluate how SPOs and SPCOs influenced 1-hr wind forecasts, wind forecasts from the PHONE_ALT and PHONE_ALTDP full-cycling experiments were verified with METAR wind observations. Figure 35, displays the time-series of 1-hr wind forecasts errors, relative to the CNTRL experiment. The right panel shows the 1-hr wind forecast RMSE difference from the CNTRL for each assimilation experiment. The median reduction in 1-hr wind forecast RMSE is very small over the entirety of the case; however, the time-series plot in Figure 35, reveals that substantial reductions in domain averaged wind forecast RMSE are observed during the peak of the windstorm between 00-06UTC on Oct. 16.

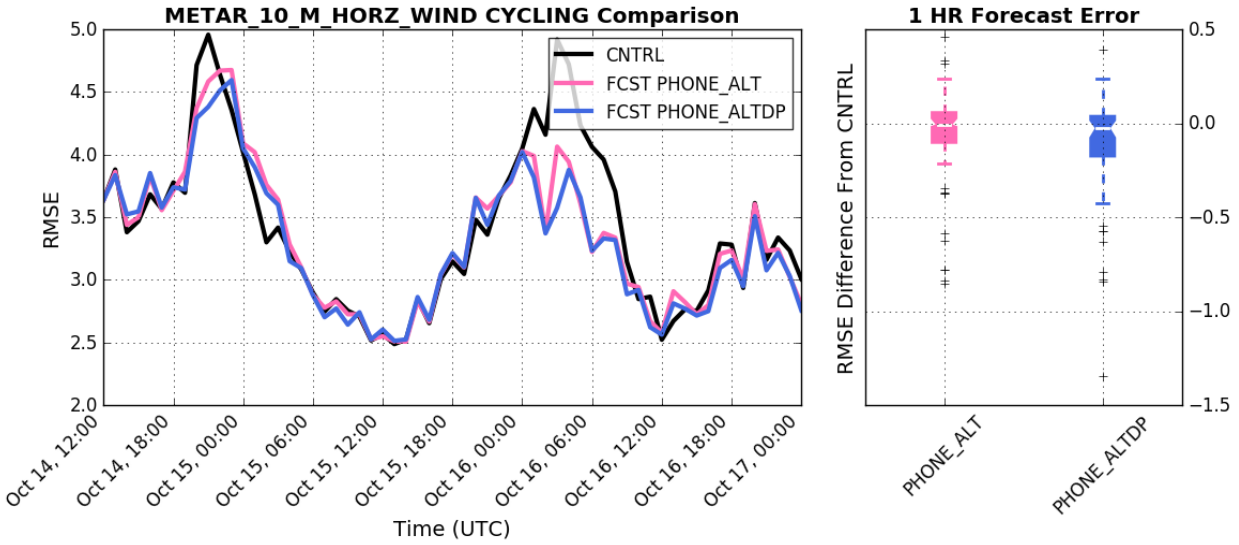


Figure 35: Time series of domain averaged 1-hr wind forecast RMSE relative to the background 1-hr forecast RMSE from the control ensemble (no assimilation). METAR altimeter observations used for verification.

During this period, the 1-hr wind forecast RMSE is decreased by up to 0.8 m/s when SPOs are assimilated and up to 1.4 m/s when SPO and SPCOs are assimilated. Interestingly, assimilating SPCOs provides additional improvements over assimilating SPOs alone. Recall that in the short

term, pressure minima tend to propagate toward the region where pressure falls are greatest. Thus, by assimilating SPCOs, the propagation of the surface low can be constrained. Since in this case, strong winds are collocated with the surface low, improvements in the low's track forecast can produce improvements in wind forecasts.

To examine how SPOs and SPCOs influenced the forecast track of the windstorm, the location of the minimum pressure in 1-hr forecasts of mean sea level pressure was analyzed for the CNTRL, PHONE_ALT, and PHONE_ALTDP experiments. 1-hr track forecasts for each of these experiments, starting at 12UTC on Oct 15 and ending at 04UTC on Oct 16, are displayed in figure 36. For reference, an estimate of the true storm track from the Real-Time Mesoscale Analysis (DePondeca et al., 2011) was also included in this figure. At 18UTC, the assimilation of SPOs and SPCOs began to produce small but notable changes in the 1-hr forecast track of the storm. Large changes in the forecast track appeared at 00UTC when the storm track in both assimilation experiments abruptly shifted westward relative to the control track. The PHONE_ALT experiment produces the most realistic forecast track. In the RTMA the storm tracks approximately sixty miles west of the entrance to the Strait of Juan de Fuca before making

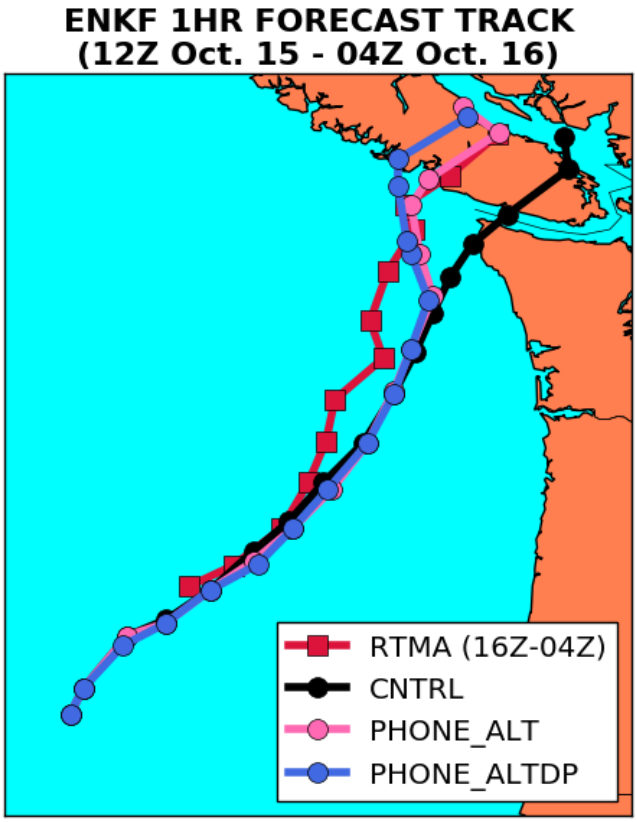


Figure 36: 1-hr forecast track for CNTRL, PHONE_ALT, and PHONE_ALTDP full cycling experiments. Each point on the forecast track represents the location of the minimum observed pressure in 1-hr cycled forecasts.

landfall on Vancouver Island near the western edge of Canada's Pacific Rim National Park. While the shift in the 1-hr forecast track was not observed until 00UTC, this shift still occurred three hours before the time when the strongest winds were anticipated to impact the Puget Sound lowlands.

4.3.4) Case II: Partial Cycling Experiments

To determine whether the western shift of the windstorm track could have been forecasted at earlier lead times, two partial cycling experiments were performed. In these experiments, a 48-member WRF ensemble was advanced freely with SKEBS perturbed RAP boundary conditions and initial conditions from either the CNTRL ensemble or the PHONE_ALT ensemble. The first partial cycling experiment was initialized at 19 UTC, while the second experiment was initialized at 21 UTC, on Oct. 15th. At each of these times, the surface low was a few hundred kilometers offshore, allowing its structure and location to be influenced by SPOs along the coast. The forecast tracks for each of these experiments initialized at 19 and 21UTC, respectively, are displayed in Figure 37. For reference, an estimate of the true storm track, from the RTMA, is also displayed. In both experiments, the initial surface low is slightly weaker in the PHONE_ALT run than in the CNTRL run. In the 21 UTC run, the initial difference in the magnitude of the surface low is more pronounced as by this time several SPO assimilation cycles have progressively weakened the surface low in the analysis. In both runs, the PHONE_ALT track diverges from the CNTRL track around 00 UTC. While the westward shift of the PHONE_ALT track is a fraction of the shift observed in the full-cycling experiments, it is still notable. In both experiments, the track and the intensity of the wind storm are nudged in the right direction by the assimilation of SPOs. In the 19Z run, the surface low is marginally weaker in the PHONE_ALT case than in the CNTRL case between 00Z and 01Z.

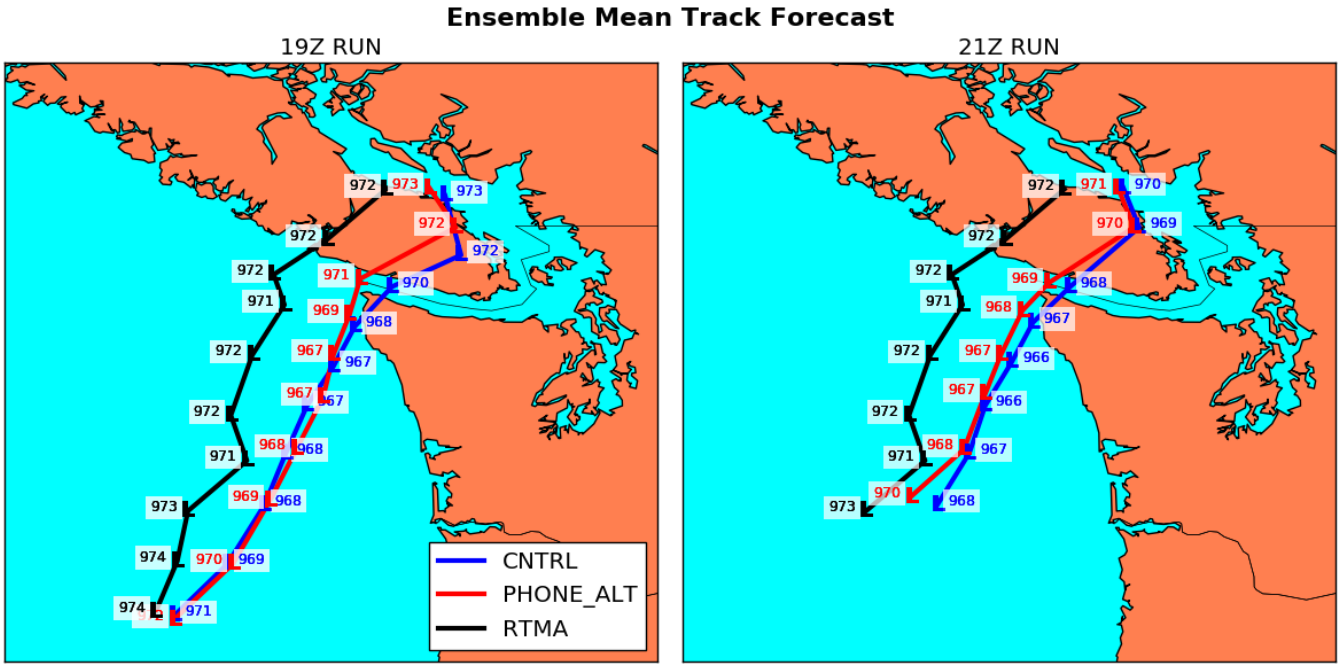


Figure 37: Forecast tracks for CNTRL and PHONE_ALT free runs initialized at 19Z and 21Z, respectively.

In the 21Z run, the minimum pressure is higher in the PHONE_ALT case than in the CNTRL case, at all forecast lead times. By constraining both the track and intensity of the surface low as it approached the PNW coast, SPOs were able to provide more accurate initial conditions to the WRF ensemble. This is a notable achievement as it suggests that SPOs could be incorporated into operational models, many of which utilize partial cycling, to improve initial conditions and short-term forecast skill.

Improvements in the track and intensity forecast of the windstorm were observed in full-cycling and partial cycling experiments, yet it remains unclear how these forecast adjustments affect the spatial distribution of wind forecast skill. The location of wind forecast skill is crucial. If the forecast adjustments produced by SPOs only improved offshore wind forecast skill this would be less satisfactory than if wind forecast skill was improved throughout the Puget Sound lowlands, where strong winds were expected to cause damage to property and infrastructure. The Brier Skill Score (BSS) metric was used to compare wind forecast skill between assimilation and

CNTRL experiments. The BSS is defined by the following equation:

$$B_{SS} = 1 - \frac{BS_{test}}{BS_{ref}}, \;\; where\; BS = \frac{1}{n}\sum_{t=0}^{n}(F_t - o_t)^2$$

This metric utilizes the Brier score (Brier, 1950) which essentially estimates the MSE for probabilistic forecasts. In the Brier score, f_t represents the fraction of the ensemble members that forecast an event to occur at time t, while o_t is a boolean that defines whether an event occurred or not at time t. In this case study, the Brier score is averaged in time and the reference forecast is the CNTRL forecast. When the BSS is negative (positive) the test forecast is less (more) skillful than the reference.

To evaluate how SPOs affected wind forecast skill in full-cycling experiments the BSS was used to compare wind forecast skill in the PHONE_ALT and CNTRL experiments. In each of these experiments, gridded verification of 1-hr wind forecasts was performed with wind analysis from the RTMA. The period used to compute Brier scores for the PHONE_ALT and CNTRL experiments extended from 12Z Oct 15th to 06Z Oct. 6th. The BSS was computed for wind forecasts exceeding 5, 10, 15, and 25 m/s (Fig. 38). For all thresholds examined, 1-hr wind forecasts from the PHONE_ALT experiment were more skillful than CNTRL forecasts. The largest increases in wind forecast skill relative to the CNTRL experiment are observed throughout Puget Sound and the Strait of Juan de Fuca. The BSS for lower threshold wind probability forecasts (5, 10 m/s) was positive over a significant portion of offshore waters. This may be a consequence of SPOs limiting the spatial extent of winds around the surface low by constraining its structure. The substantial improvements in wind forecast skill, achieved by fully cycling SPOs every hour, is not unexpected since the position and the intensity of the surface low dictated the location and strength of the winds. Since SPOs were able to better constrain the

forecast track and intensity of the windstorm the forecast wind probabilities were more skillful.

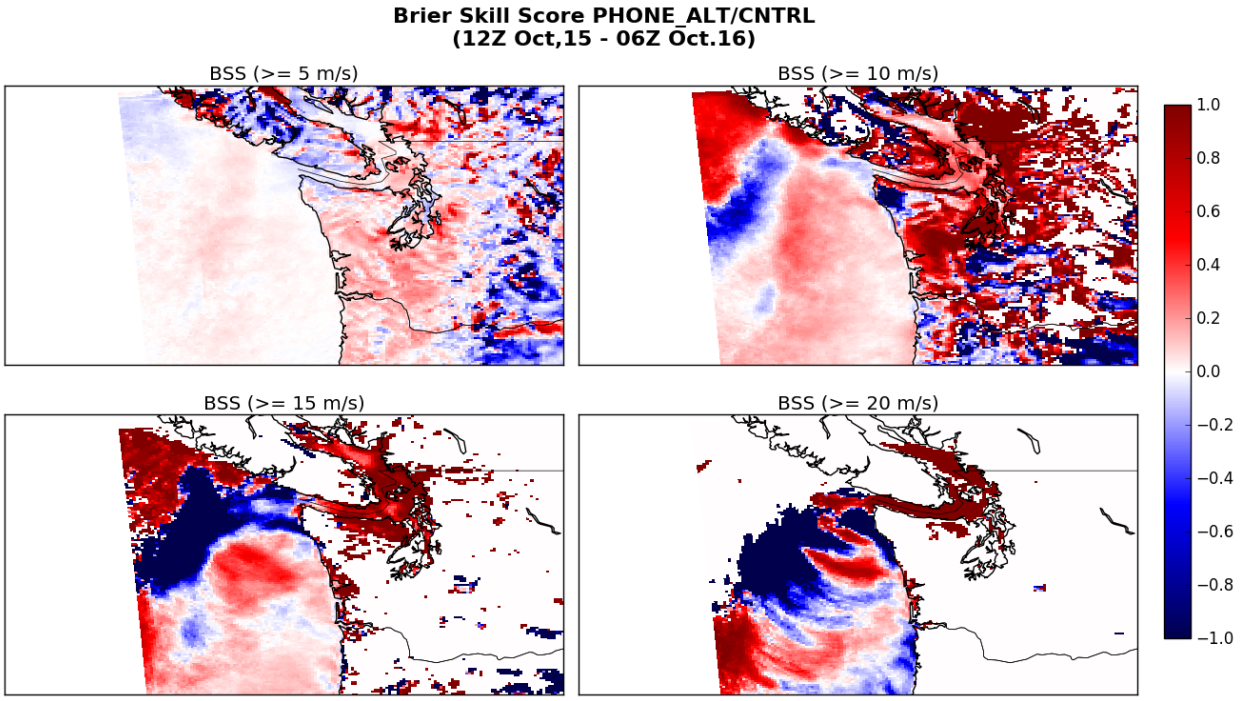


Figure 38: BSS of fully-cycling 1-hr wind forecasts greater than 5, 10, 15, and 20 m/s between 12Z Oct. 15 and 06Z Oct. 16. Red shading indicates that the PHONE_ALT forecast was more skillful, while blue shading indicates the CNTRL forecast was more skillful.

Since the value of 1-hr forecasts is limited, two partial cycling experiments were performed. For each of these experiments, the BSS for wind forecasts exceeding 5, 10, 15, and 20 m/s was evaluated at a single forecast lead time, valid at 02 Z, on Oct 16th. The BSS for seven-hour wind forecasts initialized at 19Z is displayed in Figure 39. Recall that during the first partial cycling experiment the surface low was slightly weaker and more westward in the PHONE_ALT run than in the CNTRL run at 02Z. The slight decrease in magnitude and westward shift of the surface low was enough to produce improvements in wind forecast skill in the Strait of Juan de Fuca and the lower Puget Sound. Offshore, positive BSS for wind forecasts exceeding 10 and 15 m/s are observed west of the Olympic Peninsula near the location of the surface low in the seven-hour forecast.

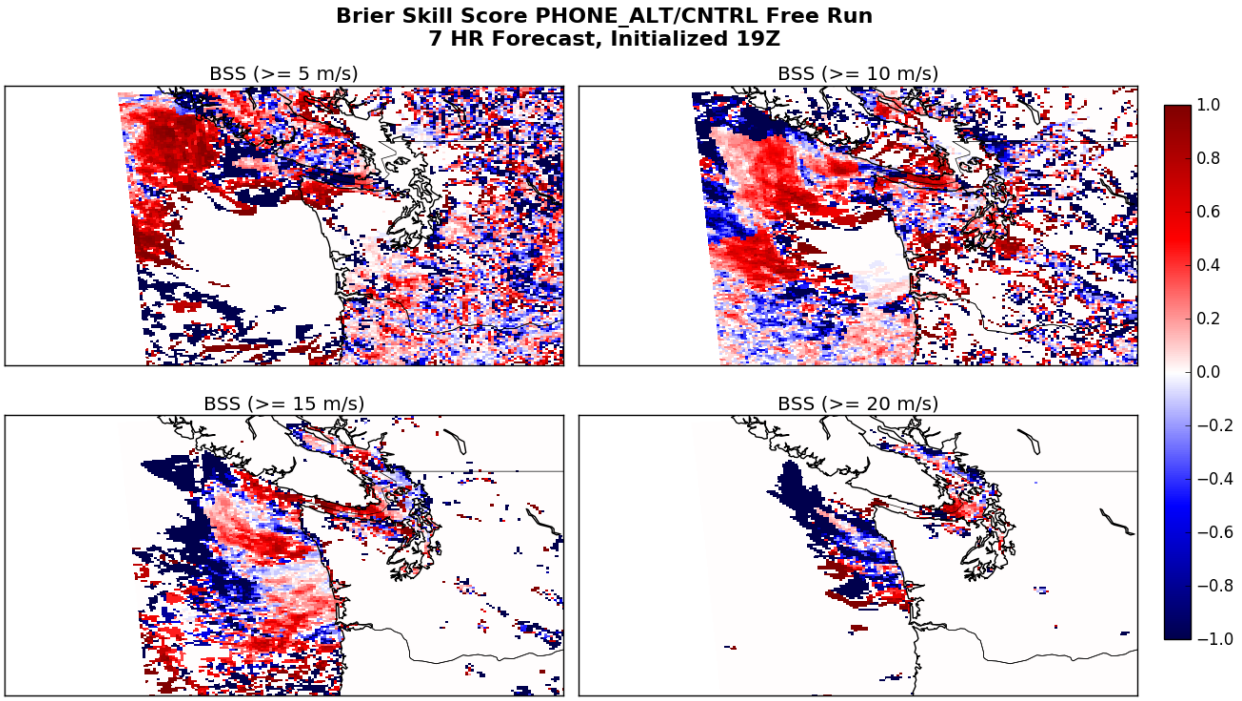


Figure 39: BSS of 7-hr wind forecasts greater than 5, 10, 15, and 20 m/s valid at 02Z Oct. 16. Red shading indicates that the PHONE_ALT forecast was more skillful, while blue shading indicates the CNTRL forecast was more skillful.

At the highest threshold of 20 m/s, the PHONE_ALT run is less skillful than the CNTRL. The 19Z free run was unable to achieve widespread improvements in wind forecast skill throughout the domain. Nevertheless, the improvement to initial conditions, provided by the assimilation of SPOs, was able to produce modest improvements in wind forecast skill at a forecast lead time of seven hours.

Wind forecast skill in the second partial cycling experiment is examined in figure 40. At 02UTC the forecast position of the surface low in the PHONE_ALT run was just off the coast of Neah Bay at the tip of the Olympic Peninsula. Relative to the CNTRL run, wind forecast skill for the 15 m/s threshold is markedly increased to the west and southwest of the Olympic Peninsula. Improvements in wind forecasts for the 10 m/s threshold are also observed in southern portions of Puget Sound and the Olympic Peninsula.

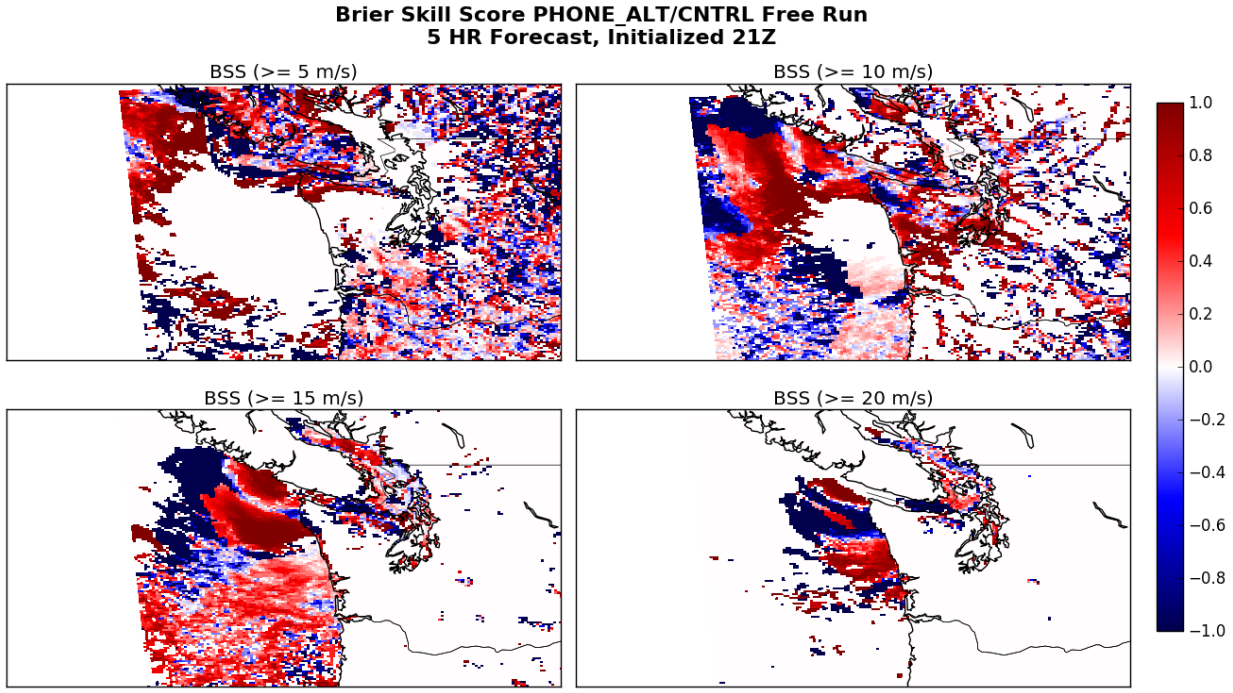


Figure 40: BSS of 5-hr wind forecasts greater than 5, 10, 15, and 20 m/s valid at 02Z Oct, 16. Red shading indicates that the PHONE_ALT forecast was more skillful, while blue shading indicates the CNTRL forecast was more skillful.

At the highest wind forecast threshold, the PHONE_ALT run is less skillful than the CNTRL run in the Strait of Juan de Fuca and off the northwest Olympic coast. It is noted that the spatial distribution of wind forecast skill improvements is markedly similar for both partial cycling experiments. In both experiments, the PHONE_ALT wind forecast skill exceeded the CNTRL forecast skill in the southern Puget Sound lowlands and the vicinity of the surface low. At these locations improvements in wind forecast skill were larger in magnitude in the second partial cycling case. This is the case because in the second partial cycling experiment free forecasts were initialized at a later hour when the surface low was closer to shore and more strongly influenced by the assimilation of SPOs. Since the forecast intensity of the windstorm was better constrained in the second partial cycling experiment improvements in wind forecast skill were more pronounced.

4.4) Summary

To date, poor data quality has plagued commercial smartphone pressure networks, undermining their use in NWP. If data quality can be improved, smartphone pressures could enhance NWP by providing unprecedented observational coverage and density. This thesis attempted to resolve challenges to mobile data quality by developing new approaches to quality control and bias correction. A novel crowdsourcing pressure app, *uWx*, was developed per this goal. In this app, a set of best practices for mobile pressure collection were defined. To limit sources of uncertainty the period over which pressure is measured was extended to avoid dynamic lag errors. Location accuracy was also improved by forcing the use of the GPS during location retrieval. This decreased the magnitude of elevation errors which contributed to inconsistencies among smartphone pressures.

Although pseudo-random sources of error related to user behavior, like dynamic lag, were largely eliminated by these procedures large pressure biases persisted. Since the traditional QC techniques utilized by MADIS failed to produce the desired results, a machine learning approach to post-processing smartphone pressures was developed. Random forests, trained on a plethora of sensor and location data, were largely successful in predicting and thus reducing phone bias. By identifying subtle relationships between ancillary phone data and phone bias, random forests were able to learn how phone bias related to user behavior. This is demonstrated by the fact that even phones in moving vehicles were able to be bias-corrected.

Improved pressure collection protocols and machine learning approaches to bias correction largely mitigated the data quality problem. With the data quality challenge met, two case studies were examined to test whether SPOs could be used to improve NWP. In the first experiment, a wind ramp event downstream of the Columbia River Gorge was examined due to

the availability of verification assets from a local field project, WFIP2. In this experiment, SPOs were able to consistently reduce analysis error of altimeter setting, temperature, and dew point. Reductions in 1-hr forecasts for these surface variables were also achieved when SPOs were assimilated in full-cycling mode. When compared to experiments that assimilated pressures from traditional MESONETs or METARs, SPOs were able to achieve nearly the same reductions in forecast and analyses error of altimeter and temperature. This result suggests that when properly collected and bias-corrected, the quality of SPOs is comparable to MESONET observations. The fact that this result is achieved by assimilating observations from less than 2,000 phones is very promising, given that there are hundreds of millions of phones capable of collecting pressures. When wind forecasts were examined during this case study, improvements were confined spatially to regions where the density of SPOs was greatest. In general, pressure assimilation from all networks did little to improve forecasts of the wind ramp event observed during this case. If substantial observational density could be achieved in regions where wind ramp forecasts are desired, pressure assimilation could provide some value to wind energy stakeholders.

In the second case study, a windstorm forecast bust was examined. The assimilation of SPOs in this case produced substantial improvements in altimeter forecasts during the climax of the event when the windstorm made landfall. By fully cycling SPOs, errors in the forecast track of the windstorm were markedly reduced. Running free forecasts at early lead times revealed that partially cycling SPOs could constrain the forecast track and intensity of the windstorm at forecast lead times out several hours. When the skill of probabilistic wind forecasts was evaluated for fully cycled experiments, widespread improvements in skill were observed. In partially cycled experiments, freely run forecasts at six-hour and eight-hour lead times produced modest improvements in wind forecast skill in the southern Puget Sound lowlands and the

vicinity of the surface low.

4.5) Implications

Many of the techniques and methods developed to quality control and bias correct SPOs in this study could be applied to other unique mobile observing platforms. The same sensors used to quality control and bias correct *uWx* observations are found in wearable devices For example, both the Samsung Gear smartwatch and Microsoft Band fitness tracker have network connectivity (Wi-Fi) and an embedded ambient light sensor, barometer, and GPS. The software developed to collect and bias correct *uWx* pressures could be applied to wearable devices since many already collect the same ancillary data as smartphones.

With the aid of machine learning, smartphone pressures measured from within moving vehicles were successfully bias-corrected. This suggests that the machine learning approach to quality control, applied in this study, could be expanded to vehicle-based sensors. The wealth of ancillary data available in vehicles, from GPS data to engine temperature, could be utilized by machine learning algorithms to identify subtle patterns related to biases in surface variables measured by onboard temperature, pressure, and humidity sensors. Meteorological conditions such as visibility and precipitation rate could even be inferred through the collection of wiper setting and headlight usage data. A machine learning approach to quality control of inferred observations would be ideal as such observations would be inherently linked to driver behavior.

As more meteorological sensors are embedded in devices, vehicles, buildings, and other items the potential for crowdsourcing weather observations will grow. This thesis lays the groundwork for this future, by demonstrating that crowdsourced meteorological data from mobile platforms can improve NWP.

References

Anderson J., 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.

Anderson J., B. Wyman, S. Zhang, and T. Hoar, 2005: Assimilation of PS observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.*, **62**, 2925–2938.

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.

Bochkanov S., Bystritsky. V., 2009: AlgLib. http://www.alglib.net

Bosch Sensortec, 2016: BMP280 Digital Pressure Sensor Datasheet, https:/ae-bst.resource.bosch.com/media/_tech/media/datasheets/BST-BMP280-DS001-18.pdf

Breiman, L. 1996: Bagging Predictors. *Machine Learning.* **24**, 123-140.

Breiman, L, 2001: Random Forests. *Machine Learning. **45**, 5–32.*

Brier, G., 1950: Verification of Forecasts Expressed in Terms of Probability. *Mon. Wea. Rev.*, **78**, 1–3.

Compo, G. P., and Coauthors, 2011: The twentieth-century reanalysis project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28.

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.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612.

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.

Duchon, 1976, Splines minimizing rotation-invariant semi-norms in Sobolev spaces. pp 85–100, In: Constructive Theory of Functions of Several Variables, Oberwolfach 1976, W. Schempp and K. Zeller, eds., Lecture Notes in Math., Vol. 571, Springer, Berlin, 1977.

Gesch, D.B., Oimoen, M.J., and Evans, G.A., 2014, Accuracy assessment of the U.S. Geological Survey National Elevation Dataset, and comparison with other large-area elevation datasets—SRTM and ASTER: U.S. Geological Survey Open-File Report 2014–1008, 10 p.

Hanson G. S., Greybush S. J, 2016: Impact of Assimilating Surface Pressure Observations from Smartphones on Regional, Convective-Allowing Ensemble Forecasts: Observing System Simulation Experiments. *Mon. Wea. Rev.*

Ho, T. K., 1998: The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844.

Lei, L., and J. Anderson, 2014: Impacts of frequent assimilation of surface pressure observations on atmospheric analysis. *Mon. Wea. Rev.*, **142**, 4477–4483.

Madaus, L., G. Hakim, and C. Mass, 2014: Utility of Dense Pressure Observations for Improving Mesoscale Analyses and Forecasts. *Mon. Wea. Rev.*, **142**, 2398–2413.

Madaus, L. and C. Mass, 2017: Evaluating Smartphone Pressure Observations for Mesoscale Analyses and Forecasts. *Wea. Forecasting,* **32**, 511–531.

Marquis, M., Olson, J., Kenyon, J., Benjamin, S., Wilczak, J., Bianco, L., Djalalova, I., McCaffrey, K., Pichugina, Y., Banta, R., Choukulkar, A., Echman, R., Clifton, A., Carley, J., & Cline, J. (2015, June). Wind forecast improvement project-2, improving model physics in complex terrain, NOAA's plans for improving the rapid refresh and high-resolution rapid refresh models. Paper presented at the North American Wind Energy Academy 2015 Symposium, Blacksburg, VA.

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

McNicholas, C., and D.D. Turner. 2014: Characterizing the convective boundary layer turbulence with a high spectral resolution lidar. *J Geophys Res Atmos*, **119**, 12910-12927Miller, P. A., M. F. Barth, and L. A. Benjamin, 2005: An update on MADIS observation ingest, integration, quality control, and distribution capabilities. Preprints, 21st Int. Conf. on Interactive Information and Processing Systems, San Diego, CA, *Amer. Meteor. Soc.*, J7.12.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. Tech. Rep. NCAR/TN- 4751STR, National Center for Atmospheric Research, Boulder, CO, 125 pp.

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.

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.