Ensemble Calibration and Uncertainty Quantification of Precipitation Forecasts for a Risk-based UTM

Mounir Chrit[1]

1. Department of Atmospheric Sciences, University of North Dakota, Grand Forks, ND 4149, USA.

Abstract

Uncertainty on precipitation forecasts results in major high cancellation rate in Unmanned Aircraft Systems operations and reduces the benefits of BVLOS operations in terms of risk-based contingency planning. Hence, quantifying and reducing the uncertainty on precipitation forecasts will reduce mission uncertainties, avoid accidents and make the integration of UAS into the National Airspace System more efficient and reliable. To achieve this goal, the Member-By-Member post-processing technique is used to calibrate a probabilistic forecast composed of 20 members of precipitation rate over South Florida during summer period. The Continuous Ranked Probability Score (CRPS) of the ensemble is minimized to achieve the optimal regression between ensemble members without any assumption on the forecasted parameter. The radar data from the Multi-Radar/Multi-Sensor (MRMS) is used to correct ensemble spread every 10 min and reduce forecasting uncertainty. A multi-physics ensemble was used to generate high-resolution, convection resolving/allowing 48-hours forecasts. The calibration was obtained over a learning process over the simulated period over 3 years. The comparison between the raw and calibrated ensemble from unseen data is presented in terms of bias correction and ensemble reliability. The calibration was able to correct the bias found in raw probabilistic forecasts relative to MRMS data. The comparison with precipitation data from tipping buckets over four airports over South Florida revealed that the calibrated ensemble tends to overestimate the precipitation rates mainly because of the particles evaporation that is taking place under radar beam.

Plain Language Summary

The uncertainty on precipitation forecasts is a very important information for contingency planning within the framework of Beyond Visual Line Of Sight (BVLOS) Operations of Unmanned Aircraft Systems (UAS) and UAS Traffic Management (UTM) systems. In this article, forecasts uncertainty is reduced using ensemble calibration techniques using merged radar data over South Florida. This technique optimizes regression coefficients by learning from historical data and minimizing the difference between observations and forecasts and we show that thanks to the calibration, the ensemble becomes more reliable and the bias of the calibrated ensemble improved. The comparison between precipitation forecasts and ground-based data over airports revealed an improvement the forecasts, as the calibration is very sensitive to the used radar observations.

Introduction

Rain of different amplitudes accompanied by thunderstorms, reduced visibility and wind gusts are non-negligible threat to small Unmanned Aircraft Systems (sUAS). These precipitations are also a major cause of UAS flights cancellation as operators because flying during wet conditions is still conservative, which pose a genuine challenge to the business model of multiple companies relying on BVLOS operations (Campbell et al. 2017). Moreover, precipitations can cause lost-link hazard which also make the BVLOS operations less efficient. In addition, contingency planning related to lost-link hazards is highly impacted by deterministic assessment of precipitation because it requires weather uncertainty information, uncertainty on precipitation forecasts in particular. To solve this problem, Campbell et al. 2017 suggested two main recommendations: 1) quantify forecasts uncertainty and 2) investigate new solutions to reduce these uncertainties.

Within this risk-based planning approach, ensemble forecasting is widely used to provide more accurate forecasts and uncertainty information (Gneiting & Katzfuss, 2014). In fact, the ensemble mean is generally used as the forecast and the ensemble standard deviation or spread as the forecast uncertainty. However, systematic errors make forecasts 'certainty and accuracy strongly degrade and their reliability decreases as a function of lead times as the ensembles become very overconfident (under-dispersive) as shown in Nicolis et al. 2009 and Leutbecher and Palmer 2008.

Foretunately, these forecasting issues can be solved using ensemble post-processing and calibration. Multiple studies used different calibration techniques to improve probabilistic forecasts of vector or scalar variables (Pinson, 2012, (Vannitsem, 2009; Van Schaeybroeck and Vannitsem, 2011, 2012).

Two approaches exist today to calibrate an ensemble of forecasts. The first method is 'statistical' such as logistic distribution used in Wilks, 2009; Schmeits and Kok, 2010; Roulin and Vannitsem, 2012 or Non-homogeneous Gaussian Regression used in Gneiting et al., 2005; Hagedorn et al., 2008. However, these techniques are generally based on random sampling from assumed predictive distributions and ignore spatial and temporal correlations and cross-correlations as shown in Van Schaeybroeck and Vannitsem 2015. The second approach adopted in this work is member by member (MBM) independent calibration by which every member is individually calibrated in order to retain correlation correlations (Van Schaeybroeck and Vannitsem 2015).

In the MBM approach, Different cost functions and fitting procedures exist: Bayesian Model Averaging (BMA) used by Raftery et al. 2005; Sloughter et al. 2010. Other studies such as Bröcker and Smith 2007 used likelihood maximization with the logarithm loss but showed that this method fails to produce accurate calibrated members. However, the mentioned techniques are mainly based on strong assumptions and do not offer strong guarantees on ensemble improvement. The continuous ranked probability score (CRPS) is the squared difference between the cumulative distribution functions of the ensemble forecast and the observation was used by Thorey et al. 2018, Gneiting et al. 2005, Gebetsberger et al. 2017 as a cost function to minimize and obtain calibrated forecasts as it does not need a theoretical assumption regarding parameters distribution.

The goal of this paper is to show how ensemble-spread correction using CRPS minimization relative to the Multi-Radar Multi-Sensor (MRMS) precipitation data over multiple years yield to an improvement of the predictions and evaluate the performance of probabilistic forecasts of precipitation by comparison to precipitation observations over airports. In this study, we start with 20-members ensemble of precipitation forecasts and apply a MBM calibration approach developed by Schaeybroeck et al. 2015 to improve the probabilistic forecasts of a precipitation event in South Florida.

This paper is structured as follows: section 1 describes the calibration method. Section 2 discusses the simulated use case, the simulation setup and ensemble building, and the datasets used in the calibration and evaluation. Section 3 explains the evaluation method. Section 4 discusses the results and evaluation findings.

1. Ensemble Calibration

1.1. MBM post-processing method

Following Schaeybroeck and Vannitsem 2015, the calibrated ensemble of M members at time n $X_{C,n} = \left(X_{C,n}^{m}\right)_{1\leq m\leq M}$ can be expressed as a function of the raw ensemble $X_n = (X_n^m)_{1\leq m\leq M}$ as shown in Equation 1where $\overline{X}_n$ is the ensemble-mean, α is the bias parameter, β represents the ensemble-mean scale parameter. The parameter τ_n defined in Equation 2 is the spread tuner or adjuster of the corrected ensemble while ϵ_n defined in equation 4 represents the deviation from the mean of the uncorrected ensemble. $\langle . \rangle_m$ denotes the ensemble average. The standard deviation of the corrected ensembles is used as a spread measure of the corrected forecasts to quantify the uncertainty of the forecasts.

$$X_{C,n} = \alpha + \beta\,\overline{X}_n + \tau_n \epsilon_n \qquad (1)$$

$$\tau_n = \gamma_1 + \gamma_2 \delta_n^{-1} \qquad (2)$$

$$\delta_n = \left\langle \left\langle \left| X_n^{m_1} - X_n^{m_2} \right| \right\rangle_{m_1} \right\rangle_{m_2} \qquad (3)$$

$$\epsilon_n = X_n - \overline{X}_n \qquad (4)$$

1.2.CRPS minimization

The parameters (α, β, γ1, γ2) are estimated through regression learning through the same time over 3 years by the minimization of the associated Continuous Ranked Probability Score (CRPS) which is the squared difference between the Cumulative Distribution Functions (CDFs) of the ensemble forecasts and observations.

The loss function defined as the CRPS corresponding to the observations $X_{O,n}$ and the corrected-forecast members $X_{C,n}^{m}$ can be written as shown in Equation 5 (Gneiting and Raftery, 2007). The correction is used every 10 min during the two simulated summer days for 3 years: 2019, 2020 and 2021. The forecast ensemble used here covers three years 2019, 2020 and 2021 and the CDF of the observations was based on the radar MRMS observations over the same years at the same two days. Data from 2022 will be used as an independent test for the calibration. A short training period was chosen in this work (48 hours). In fact, there is a trade-off in selecting the length of the training period. Shorter training periods can be used to correct flow-dependent model biases that have rapid variations while longer training periods aim at reducing the statistical variability of different coefficients and hence the calibrated forecast.

$$CRPS\,(\alpha,\,\beta,\,\gamma_1,\,\gamma_2) \;=\; \left\langle \left\langle \left| X_{C,n}^{m} \;-\; X_{O,n} \right| \right\rangle_{m} \;-\; \frac{\delta_n}{2} \right\rangle_{n} \qquad (5)$$

2. Materials and Methods

2.1. Use case description

The simulated event is precipitation event that took place in South Florida that was visible in the MRMS data with scales of 200 km and small scales of 1-50 km as shown Figures 1 and 3. These events fall under Meso-β and Meso-γ features. In South Florida, particularly during the summer, mesoscale weather features (e.g., land–sea breezes, thermal troughs, outflow boundaries, etc.) have a significant impact on day-to-day weather forecasting, as they frequently represent the primary forcing for convection. During the simulated period,

These mesoscale features necessitate the use of high-resolution, convection resolving forecast tools in order to provide the detailed information needed to improve local forecasts and warnings. Moreover, Florida has recently emerged as a leader in autonomous vehicles including UAS through different investments in its Department of Transportation. Therefore, South Florida is a suitable area to study precipitation forecasting and its impact on UAS contingency planning.

During the simulation summer period, precipitations were of different types: mainly convective because sea breezes are often form on the west and east sides of Florida, and due to differences in temperature between the land (which heats quickly) and the ocean (which heats up more slowly) which enhance the convective lift and induce intense rainfall and thunderstorms. Convective and tropical convective precipitation are often embedded in areas of warm stratiform precipitation. Warm stratiform precipitations are also present in South Florida that result from frontal systems where the growth of hydrometeor particles occurs.

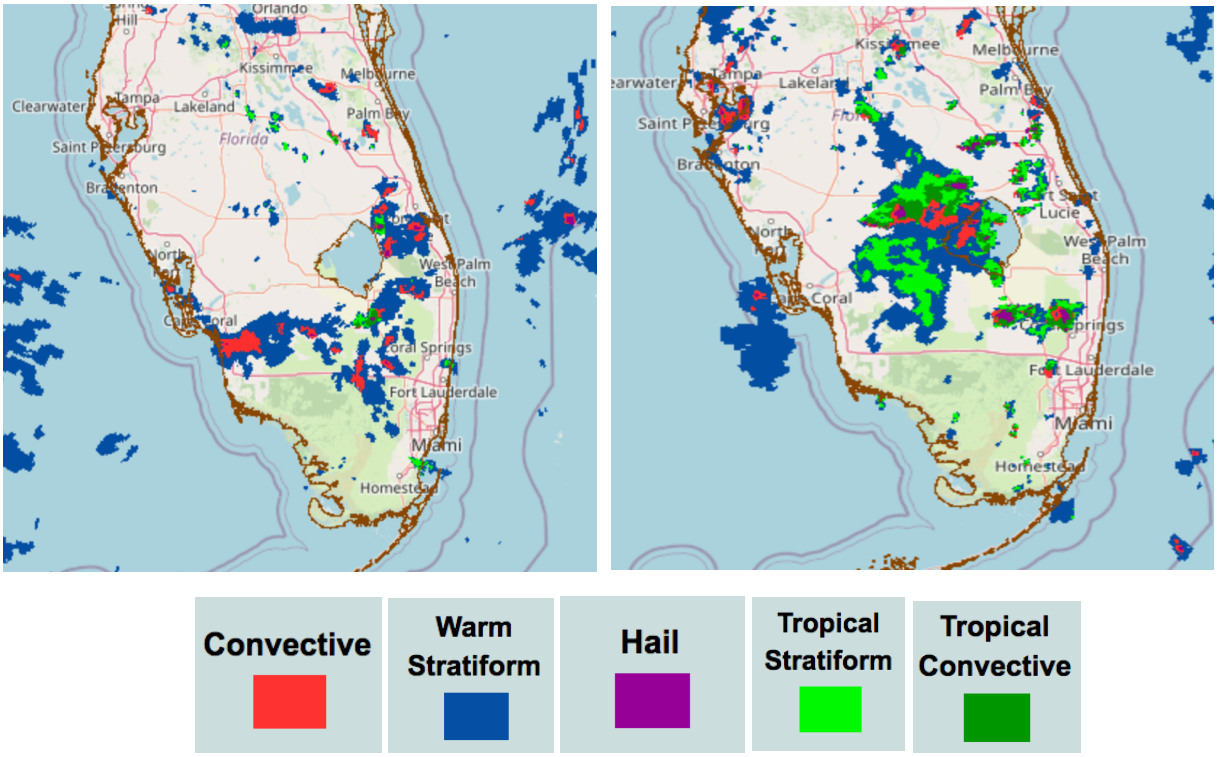

Figure 1: MRMS Precipitation type over South Florida on July 16th at 5:40pm (left panel) and on July 17th at 6:44 pm (right panel (https://mrms.nssl.noaa.gov/qvs/product_viewer/) .

2.2.Ensemble Forecasts

2.2.1. Simulations Setup

WRF (Sharmarock et al. 2005) was widely used in both academic research and industry (Chrit et al. 2022, Chrit et al. 2018, Chrit et al. 2017). The fully compressible and non-hydrostatic dynamic framework is used in the ARW module. The simulated domains D1 and D2 shown in Figure 2 represent the outermost and innermost domains respectively. The horizontal resolutions of D1, D2 are 3-km and 1-km. Vertically, 80 vertical levels are used with 30 vertical levels used below 1-km. The central point of the two domains is 80.74332 ºW, 26.40334 ºN.

The outermost D1 and innermost D2 domains have 560 x 720 and 460 x 400 grid points respectively in the south-north and east-west directions. In order to guarantee the numerical stability of the WRF model, the adaptive time stepping is used. The configuration and the physical parameterizations used in the simulations over D1 and D2 are shown in Table [1] of Appendix A.

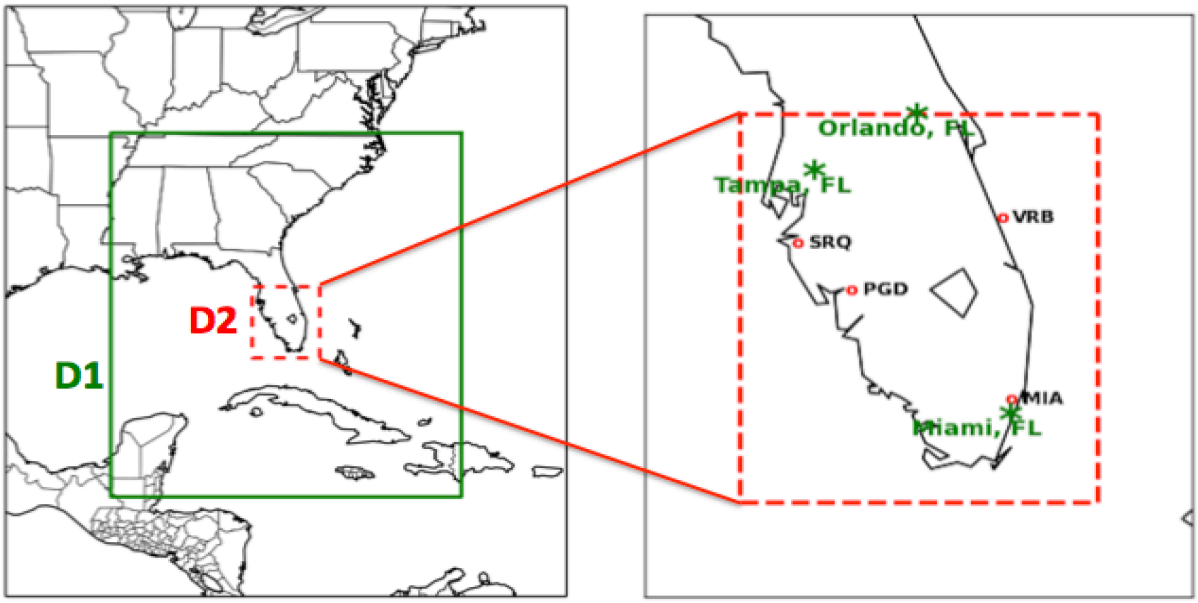

Figure 2: Left panel: Map of the simulated outermost and nested domains D1 and D2 delimited with green solid and red dashed rectangles, respectively. Right panel: Simulation domain is delimited with red dashed rectangle with the four ASOS stations used for evaluation shown with red points. The Three major cities in Florida (Miami, Orlando and Tampa) are shown in green stars.

2.2.2. Ensemble design

The ensemble used in the present study is a multi-physics ensemble with forecasts initialized with different initial and boundary conditions. In fact, multi-physics schemes have been very successful in generating reliable probabilistic forecasts, particularly for mesoscale prediction systems. Although obtaining these forecasts is computationally intensive, the ensemble results in members with physical interpretation comparing to members generated with perturbed initial conditions that poses difficulties for physical interpretation. On the other hand, precipitation forecasting is sensitive to the simulation setup namely the cumulus convection scheme (Vitart et al. 2001; Biswas et al. 2014), microphysics scheme (Liu et al. 2020), boundary layer parameterization (Taraphdar and Pauluis 2021) and radiations schemes (Li et al. 2014).

In this work, 20 distinct combinations of physics packages for parameterizing the microphysics (MP scheme), cumulus (C), Short Waves (SW) and Long Waves (LW) parameterization, planetary boundary layer (PBL), and land-surface models, (Table 1) are used to build four ensembles: three ensembles simulating the same 48 hours plus 12 hours as spin-up period (from July 15th, 2018 at 12 pm UTC to July 18th, 2021 at 12 am UTC) but over 2019, 2020, 2021 for the training and the fourth for testing simulating the same 48 hours during 2022. To maximize ensemble diversity, different boundary and initial conditions were used based on four models: the North American Model (NAM), RApid Refresh (RAP), North American Regional Reanalysis (NARR) and Global Forecast System (GFS). A total of 20 WRF simulations were performed to build the ensemble for each year.

Two MP parameterizations used are Microphysics schemes used are Thompson (Thom.; Thompson et al. 2008), WRF single-moment 6-class (WSM6; Hong and Lim . 2006). The C schemes used here are: Kain–Fritsch (Kain and Fritsch, 1993) cumulus parameterization, and Betts–Miller–Janjic cumulus parameterization (Betts & Miller, 1993). Two PBL parameterizations were used: Mellor–Yamada–Janjic (MYJ; Janjic 1994), Yonsei University (YSU; Noh et al. 2003). Two Land-Surface models were used: Rapid Update Cycle (RUC; Benjamin et al. 2004) or NOAH (NCEP–Oregon State University–Air Force–NWS Office of Hydrology; Ek et al. 2003). The SW parameterizations are Goddard (Tao et al. 2003) and Dudhia (Dudhia 1989), the LW radiations schemes are RRTM (Mlawer et al. 1997) and GFDL (Fels and Schwarzkopf 1981).

Member number	**ICs and LBCs**	**MP scheme (Thom and WSM6)**	**PBL parameterization (MYJ and YSU)**	**Land-Surface model (NOAH and RUC)**	**SW parameterization (GFDL and DUDHIA)**	**LW parameterization (GFDL and RRTM)**	**C parameterization (KAIN FRTISCH and BMJ)**
1	NAM	Thom	MYJ	NOAH	DUDHIA	RRTM	Kain
2	NAM	WSM6	MYJ	NOAH	DUDHIA	RRTM	Kain
3	NAM	Thom	YSU	NOAH	DUDHIA	RRTM	Kain
4	NAM	Thom	MYJ	RUC	DUDHIA	RRTM	Kain
5	NAM	Thom	MYJ	NOAH	GFDL	RRTM	Kain
6	NAM	Thom	MYJ	NOAH	DUDHIA	GFDL	Kain
7	NAM	Thom	MYJ	NOAH	DUDHIA	RRTM	BMJ
8	NAM	Thom	YSU	RUC	DUDHIA	RRTM	Kain
9	NAM	Thom	YSU	RUC	GFDL	RRTM	Kain
10	NAM	WSM6	YSU	RUC	DUDHIA	GFDL	BMJ
11	RAP	WSM6	YSU	RUC	DUDHIA	RRTM	Kain
12	NARR	Thom	MYJ	NOAH	DUDHIA	RRTM	Kain
13	GFS	Thom	MYJ	NOAH	DUDHIA	RRTM	Kain
14	NARR	Thom	MYJ	NOAH	DUDHIA	RRTM	Kain
15	RAP	Thom	MYJ	NOAH	DUDHIA	RRTM	Kain
16	RAP	Thom	YSU	RUC	GFDL	RRTM	Kain
17	NAM	WSM6	YSU	RUC	DUDHIA	GFDL	Kain
18	RAP	Thom	MYJ	RUC	DUDHIA	RRTM	Kain
19	GFS	Thom	MYJ	RUC	DUDHIA	RRTM	Kain
20	GFS	WSM6	MYJ	NOAH	DUDHIA	GFDL	Kain

Table 1: Physics packages for multi-physics ensemble: Parameterizations and schemes used for every ensemble member.

2.2.3. MRMS radar data

The Multi-Radar/Multi-Sensor (MRMS) system was created by the NOAA National Severe Storms Laboratory (NSSL) to produce severe weather and precipitation products for decision-making capabilities to improve severe weather forecasts and warnings, hydrology, aviation, and Numerical Weather Prediction. It currently integrates about 180 operational radars and creates a seamless 3D radar mosaic across the CONtiguous United States (CONUS) and southern Canada at very high spatial (1 km) and temporal (2 min) resolution.

The performance of the MRMS system over single radar-based Quantitative Precipitation Estimates (QPE) across CONUS was reasonable (Zhang et al., 2016). Chen et al. (2020) evaluated the MRMS and Global Precipitation Measurement Mission (GPM) products at 1-hr temporal resolution across Harris County and Spring Basin Texas. Their results showed that remote sensing technologies could detect and estimate the unprecedented extreme rainfall associated with Hurricane Harvey. Among the remote sensing products they used in their study, MRMS had the best agreement with the network rain gauge observations.

The MRMS surface precipitation rate used in this paper is currently calculated using multiple R–Z relationships. Polarimetric variables are not used because various polarimetric radar QPE schemes are still under evaluation across CONUS and an optimal approach for all seasons and all geographic regions has yet to be developed. The following empirical R–Z relationships are used in MRMS to compute surface precipitation rate for each precipitation type: convective rain, hail, warm and cold stratiform rain, snow and tropical stratiform mixed rain. More information about the MRMS system can be found at NSSL's MRMS webpage (ASOS user guide), the MRMS Fact Sheet (https://www.nssl.noaa.gov/news/factsheets/MRMS_2015.March.16.pdf), and Kirstetter et al., 2012. The MRMS data for the two simulated days were re-gridded to the same WRF grid over D2 with a 1-km resolution for every year of the learning and testing years.

2.2.4. ASOS data

The Automated Surface Observing System (ASOS) network provides most of the basic hydrometeorological observations at different airports, including 1-hour accumulated precipitation. The data is reported every 5 min in the majority of the stations. One hour precipitation for the period from the observation time to the time of the previous hourly precipitation reset. The precipitation accumulation algorithm obtains precipitation accumulation data from the Heated Tipping Bucket (HTB) precipitation gauge once each minute (ASOS user guide). The trace reports are considered as 0.1 mm. The detection threshold specified for the ASOS HTB is 0.01 inch per hour (0.254 mm per hour), and the precipitation rate accuracy is the larger of 10 percent or 0.01 inches per hour (0.254 mm per hour).

For this study, four METAR observation sites located over South Florida were used for the evaluation of the different forecasts, and these sites are shown in Figure 1. Table [1] of Appendix B shows the characteristics of the four stations that will be used for comparison and evaluation. Additional stations are available, but either no precipitation is recorded, or most data is missing.

3. Evaluation method

The probabilistic evaluation will be based on the rank histogram score and the reliability diagram. The rank-histogram score δ defined in Equation (6) is a tool used to measure the spread and hence the reliability of the ensemble.

$$\delta = \frac{N+1}{N\,M} \sum_{j=0}^{N} \left(r_j - \overline{r}\right)^2 \qquad (6)$$

$$\overline{r} = \frac{M}{N+1} \qquad (7)$$

The rank-histogram score measures the deviation from a perfect and flat rank histogram (Talagrand et al., 1999; Candille and Talagrand, 2005). In Equation (6), N is the number of ensemble members, M is the number of observations, r_j the number of observations of rank j, and $\overline{r}$ is the expectation of r_j defined in Equation (7). The optimal ensemble with a flat rank histogram has a score of 1. A score lower than 1 would indicate overconfidence in the results.

Reliability is also assessed with the reliability diagram. This diagram provides a probabilistic interpretation in terms of frequency of occurrence of precipitation events. The x-axis represents the predicted probability of occurrence (p) of an event in a time when the y-axis represents the relative frequency which is defined as the proportion of the observed event that really occurred among events with a predicted probability of p. A reliability curve overlaying the first bisector shows a perfectly reliable ensemble.

The statistical evaluation of the forecasted PR against the ASOS data uses classical skill metrics namely the simulated mean ($\overline{s}$), the Root Mean Square Error (RMSE), the correlation coefficient (R) and the Mean Bias Error (MBE). These scores are defined in Table 1 in Appendix D.

4. Results and discussions

4.1.Performance evaluation

In this section, we evaluate the two ensembles: the "Raw Ensemble" and the "Calibrated Ensemble" during the two simulated days of 2022 against the corresponding MRMS observations. Figure 3 compares the PR measured by MRMS data and the simulated data using the Raw and calibrated ensembles on July 16th , 2022 at 10 pm.

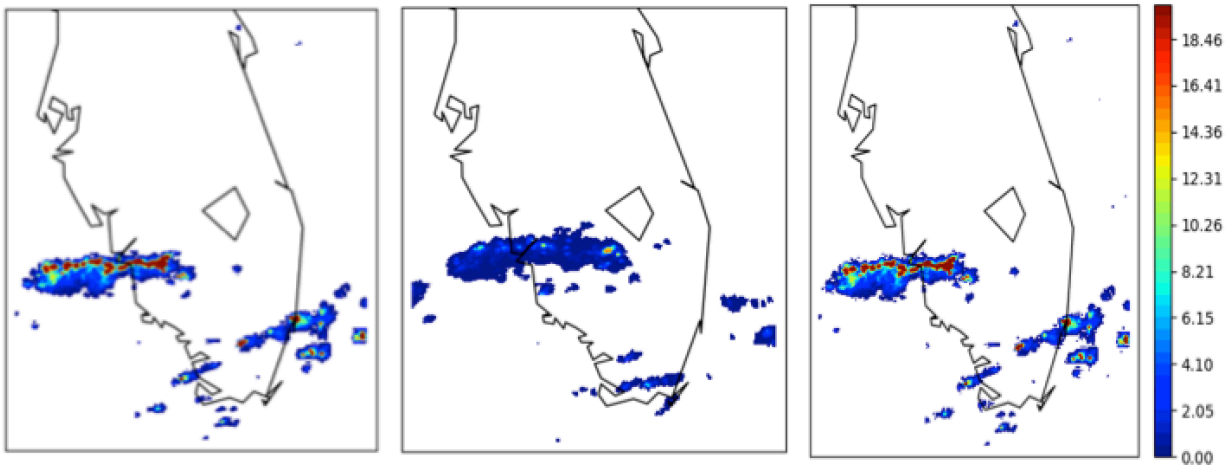

Figure 3: Left panel: PR from MRMS data on July 16th at 10 pm. Middle panel: Simulated PR

with Raw Ensemble mean at the same time and date as the left panel. Right panel: Simulated PR with Calibrated Ensemble mean at the same time and date as the left panel. The white area represents areas with zero PR.

Figure 3 shows clear discrepancies between the means of the Raw and Calibrated Ensembles. The Raw Ensemble was able to predict the location and timing of the meso-β precipitation system but was not able to reproduce the meso-γ precipitation systems over the south-eastern part of the simulation domain. However, the raw prediction of the PR is underestimated by a factor of 2. In fact, 75 % of the raw ensemble members underestimate the PR mainly because 75% of the simulated members use the Thompson microphysical scheme that produces less liquid condensate which results is lower precipitation amount. Similar results were found by Guo et al. 2019 by comparing four MP parameterizations over Eastern China over a six-year summer period (2009-2014). They concluded that the Thompson scheme creates more snow articles than other schemes which produces less graupel and precipitations during warm times. The prediction of PR using the calibrated ensemble substantially improved the PR forecasts as the predicted mean is closer to the MRMS observations and the predicted mean increased from 1 with the raw mean to 19 $mm.h^{-1}$ with the calibrated mean. Furthermore, the calibration improved the timing and the location of this simulated precipitation event and the meso-γ precipitations.

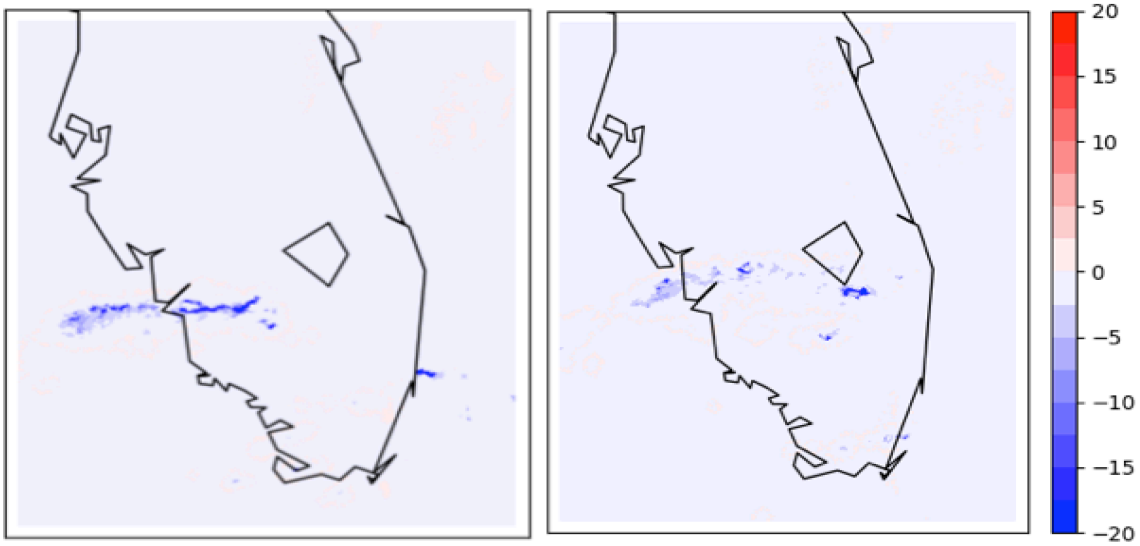

Figure 4: Left panel: Absolute difference ($mm.h^{-1}$) between the CRPS of the raw ensemble and the calibrated ensemble on July 16 at 10pm. Right panel: Similar to the left panel on July 17th at 00 am.

Figure 4 shows the impact of the calibration of the CRPS of the PR forecasts. The calibration was successful in reducing the CRPS of the calibrated ensemble by a 90 % approximately over the high PR areas, hence improving accuracy relative to MRMS observations. This improvement was guaranteed by the MBM method as it was based on learning the minimization of the CRPS.

This is indicative that the weighting coefficients were able to accurately learn temporal features during the two simulated days and correct the raw forecasts.

Figure 5 shows the bias of the means of the Raw and Calibrated Ensembles relative to the MRMS data at two different times. The mean of the raw ensemble has a high bias significant over the precipitation areas that can be as high as 20% against the MRMS data. Figure 5 shows also the impact of ensemble calibration on bias and CRPS of the probabilistic forecasts. The calibration had a significant impact over the forecasted PR as the bias of the calibrated mean decreased by 20% relative to the MRMS observed PR.

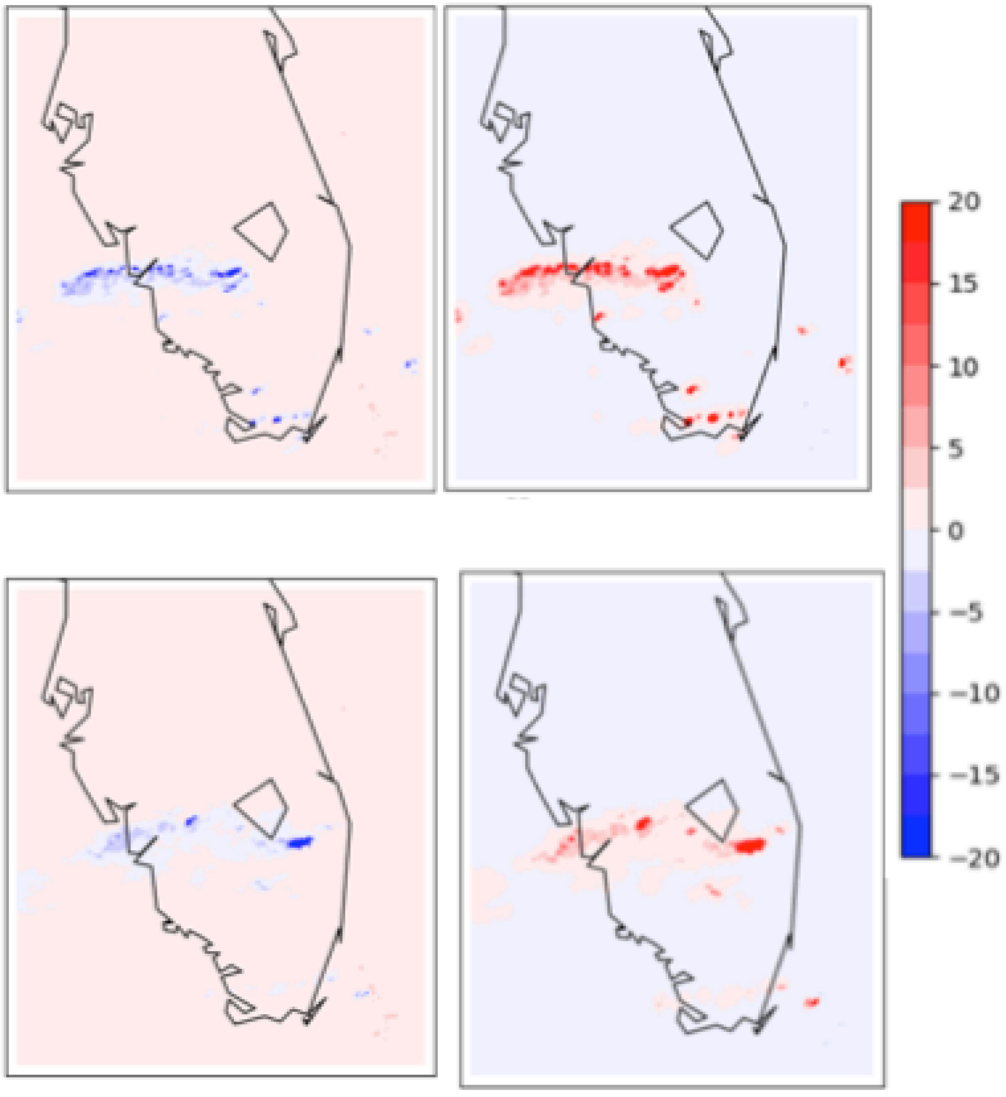

Figure 5: Top left panel: Bias error ($mm.h^{-1}$) of the Raw Ensemble mean at July 16th 2022 at 10pm . Top right panel: Absolute difference between Bias errors of the Calibrated and Raw ensemble means at 8pm. Bottom left panel: Similar to top left panel at July 17th 00pm . Bottom right panel: Similar to top right panel at July 17th at 00 am.

The reliability diagram of the Raw and Calibrated ensembles are shown in Figure 6. Raw PR forecasts tend to over-forecast both high and low probability events. When considering the calibrated ensemble, the reliability increased for both low and high frequency events. In addition, there is a better reliability for low frequency precipitation events, but the calibrated ensemble is still over forecasting the high-frequency precipitation events. The calibrated ensemble was not able to reproduce the high-frequency event because of biases related to the location and spatial extent of the precipitation events of different scales. The rank-histogram scores of the raw and calibrated ensembles are 15.9 and 4.1 respectively. The rank-histogram score decreased but still more than the optimal score confirming that the calibration improved the spread of the ensemble but still do not have optimal spread in our ensemble. Training on more years such as a decade, although very resource-intensive may further improve the reliability of the calibrated ensemble.

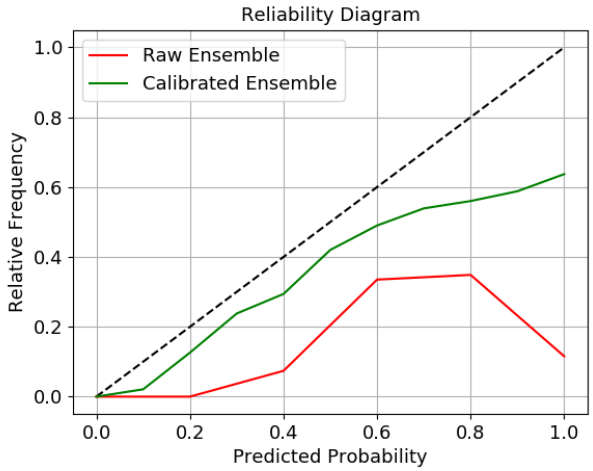

Figure 6: Reliability diagram of the Raw and Calibrated ensembles over the simulated time and over the precipitation areas of the D2 domain.

4.2.Comparison with ASOS data

The calibration is evaluated against the measured PR over the four ASOS stations shown in Figure 7. Table 2 shows the statistical scores of both raw and calibrated means. Tables 1, 2, 3 and 4 in Appendix C show the statistical evaluation of PR over the stations PGD, MIA, SRQ and VRB respectively.

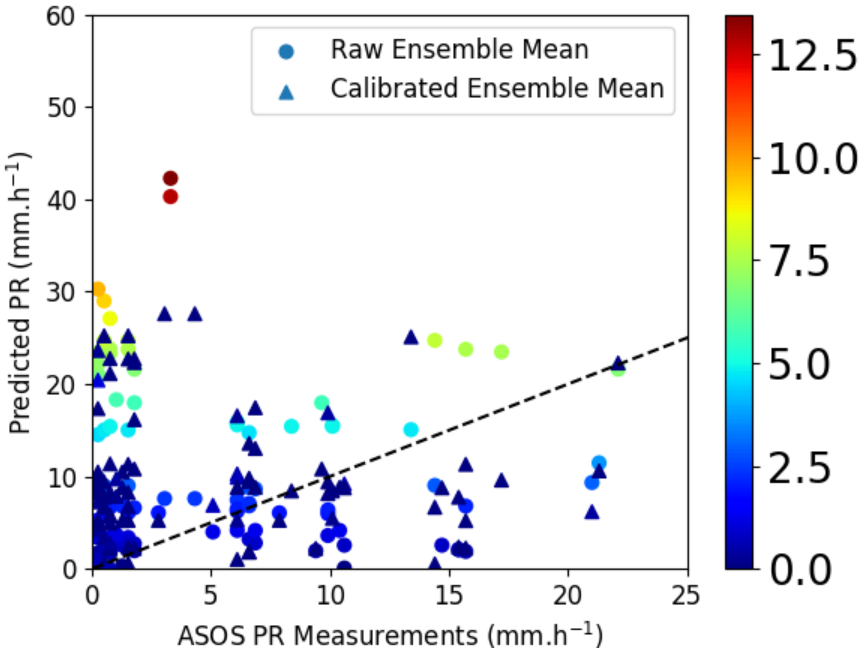

Figure 7: Scatter plot of the simulated PR using the means Raw and Calibrated ensembles. The colors are the uncertainty of the forecasts.

$\bar{o}$= 4.82 mm.h^{-1}	$\bar{s}$ (mm.h^{-1})	RMSE (mm.h^{-1})	R (%)	MBE (%)
Raw Ensemble Mean	19.07	31.60	16.70	1807.09
Calibrated Ensemble Mean	10.31	11.07	23.15	615.91

Table 2: Statistics of the means of the Raw and Calibrated ensembles against data over the four ASOS.

The scatter plot in Figure 5 shows that both raw and calibrated means overestimate the observed PR over the four ASOS stations with a simulated means of 19.07 and 10.31 mm.h^{-1} for raw and calibrated ensemble respectively against 4.82 mm.h^{-1} .The slopes of the lines of best fit are 3.87 and 1.57 for the raw and calibrated means respectively. The calibration improved the forecasts as the RMSE decreased from 31.60 mm.h^{-1} to 11.07 mm.h^{-1} and the MBE decreased from 1807.09% to 615.91 %. Besides, the uncertainty of the forecasted PR was reduced because of the calibration as the uncertainty of the calibrated mean decreased from 14 to 4 mm.h^{-1}.

The calibrated ensemble still has high bias and significantly overestimates the PR by a factor of 2. This overestimation may be due to the overestimation of PR during summertime by the MRMS data compared to ground based ASOS data because of the evaporation process occurring under the radar beam. In fact, both raw and calibrated PR forecasts overestimate the light precipitation (particularly PR less than 2 mm.h^{-1} because the MRMS data also overestimates the

light precipitations. Similar result was found by Gao et al. 2018 by evaluating the MRMS data against the NEXt generation weather RADar (NEXRAD) data over TEXAS, USA and a dense rain gauge network covering Harris County, Texas, USA. Santer and Grams 2020 evaluated the MRMS Quantitative precipitation estimation (QPE) and PR during 18-months period against rain gauges from the National Centers for Environmental Prediction Meteorological Assimilation Data Ingest System (MADIS) over CONUS and showed that, during warm times, an important systematic overestimation exist because of sub-radar beam evaporation. They also quantified the uncertainty of a MRMS radar measurement based on distance from the radar and partial radar beam blockage

Conclusion

In this study, we have applied the MBM calibration technique by minimizing CRPS in order to improve the probabilistic forecasting of precipitation as part of a risk-based approach to integrate UAS into the NAS. The algorithm does not depend on any assumptions on distributions such as gaussianity or uniformity and comes with theoretical guarantee of performance.

The case study examined the impact of ensemble calibration on precipitation forecasts accuracy and uncertainty over South Florida. The MRMS radar data was used to calibrate a 20-members ensemble that was underestimating the PR. This paper showed that CRPS minimization brings improvement on classical scores for the ensemble mean and probabilistic diagnostic tools. Indeed, the forecasting capability measured by classical scores (RMSE, MBE and bias) are improved by the algorithm used during the two simulated summer days. Besides, this spread correction provides a bias correction, improved the reliability of the ensemble and reduced forecasts' uncertainty although the comparison with ASOS data shows a persistent overestimation because of the inherent bias of the MRMS data.

In addition, the selection of more predictors such as relative humidity, cloud cover and vertical wind velocity may further enhance the skill of probabilistic post-processing for near-real-time precipitation estimates. Besides, using satellite data along with radar data as used here may also improve the evaluation against ground-based validation. The use of deep learning methods such as distributional regression network, Bernstein Quantile Network and Histogram Estimation Network is a promising as demonstrated in Schulz and Lerch 2022.

Future work should investigate the validation of the impact of the calibration and weights on other use cases and the assessment of the performance of the calibrated ensemble over longer lead times and different testing periods. The validation against denser rain gauges network is also necessary as it will show the accuracy of the calibration over off-airport areas which is important for weather-risk assessment and contingency planning during BVLOS operations.

Acknowledgments

This work was funded by the University of North Dakota.

The authors declare no conflict of interest.

Data Availability Statement

The MRMS data used in this paper are publicly available in https://www.nssl.noaa.gov/projects/mrms/ . The WRF outputs are available upon request from the corresponding author. The code used to calibrate the ensemble is available in the open source python library available "pythie" here: https://github.com/Climdyn/pythie. The ASOS data are publicly available in https://mesonet.agron.iastate.edu/request/download.phtml?network=JP__ASOS.

References

Automated Surface Observing System : ASOS User's Guide. [Washington, D.C.] :U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration : Federal Aviation Administration : U.S. Navy : U.S. Dept. of the Air Force, 1998.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, andR. Bleck, 2004: Mesoscale weather prediction with the RUChybrid isentropic-terrain-following coordinate model.Mon.Wea. Rev.,132,473–494.

Betts, A., & Miller, M. (1993). The Betts-Miller scheme. In The representation of cumulus convection in numerical models (pp. 107–121). American Meteorological Society. https://doi.org/10.1007/978-1-935704-13-3_9

Biswas, M., Bernardet, L., & Dudhia, J. (2014). Sensitivity of hurricane forecasts to cumulus parameterizations in the HWRF model. Geophysical Research Letters, 41, 9113-9119. doi:10.1002/2014GL062071

Bröcker, J., Smith, L.A., 2007b. Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting 22, 382–388.

Campbell, S. D., Clark, D. A., Evans, J. E., 2017, Preliminary UAS Weather Research Roadmap, Project Report ATC-438, MIT Lincoln Laboratory, Lexington, MA.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. Quart. J. Roy. Meteor. Soc., 131, 2131–2150, doi:10.1256/qj.04.71.

Chen, M.; Nabih, S.; Brauer, N.S.; Gao, S.; Gourley, J.J.; Hong, Z.; Kolar, R.L.; Hong, Y. Can Remote Sensing Technologies Capture the Extreme Precipitation Event and Its Cascading Hydrological Response? A Case Study of Hurricane Harvey Using EF5 Modeling Framework. Remote Sens. 2020, 12, 445. https://doi.org/10.3390/rs12030445

Chrit, M.; Majdi, M. Using Objective Analysis for the Assimilation of Satellite-Derived Aerosol Products to Improve PM2.5 Predictions over Europe. Atmosphere 2022, 13, 763. https://doi.org/10.3390/atmos13050763

Chrit, M.; Sartelet, K.; Sciare, J.; Pey, J.; Nicolas, J.B.; Marchand, N.; Freney, E.; Sellegri, K.; Beekmann, M.; Dulac, F. Aerosol sources in the western Mediterranean during summertime: A model-based approach. Atmos. Chem. Phys. 2018, 18, 9631–9659.

Chrit, M.; Sartelet, K.; Sciare, J.; Majdi, M.; Nicolas, J.; Petit, J.E.; Dulac, F. Modeling organic aerosol concentrations and properties during winter 2014 in the northwestern Mediterranean region. Atmos. Chem. Phys. Discuss. 2018, 2018, 1–28.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. J. Atmos. Sci., 46 , 3077–3107.

Ek, M. B., K. E. Mitchell, Y. Lin, P. Grunmann, E. Rogers, G. Gayno, and V. Koren, 2003: Implementation of upgraded Noah land-surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J. Geophys. Res.,108,8851, doi:10.1029/2002JD003296.

Fels, S.B.; Schwarzkopf, M.D. An efficient, accurate algorithm for calculating CO2 15 micron band cooling rates. J. Geophys. Res. Ocean 1981, 86, 1205–1232.

Gao, S., Zhang, J., Li, D., Jiang, H., and Fang, N. Z., "Evaluation of Multi-Radar Multi-Sensor (MRMS) and Stage IV Gauge-adjusted Quantitative Precipitation Estimate (QPE) During Hurricane Harvey", vol. 2018, 2018.

Gebetsberger, M.; Messner, J. W.; Mayr, G. J.; Zeileis, A. (2017) : Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood, Working Papers in Economics and Statistics, No. 2017-23, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

Gneiting, T., and M. Katzfuss (2014): "Probabilistic Forecasting," Annual Review of Statistics and Its Application, 1, 125–151.

Gneiting T, Raftery AE, Westveld A, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133: 1098–1118.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Statist. Assoc. 102: 359–378.

Hagedorn R, Hamill TM, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Mon. Weather Rev. 136: 2608–2619.

Hong, S-Y., and J-O. J. Lim, 2006: The WRF Single-Moment 6-Class Microphysics Scheme (WSM6). J. Korean Meteor. Soc., 42 , 129–151.

Janjic, Z. I., 1994: The step-mountain eta coordinate model: Further development of the convection, viscous sublayer, and turbulence closure schemes. Mon. Wea. Rev., 122, 927-945.

Johnson C, Bowler N. 2009. On the reliability and calibration of ensemble forecasts. Mon. Weather Rev. 137: 1717–1720.

Kain, J.S., Fritsch, J.M., 1993. Convective parameterization for mesoscale models: the Kain–Fritsch scheme. The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr., No. 46. Amer. Meteor. Soc., pp. 165–170

Kirstetter, P. E., Y. Hong, J. J. Gourley, S. Chen, Z. Flamig, J. Zhang, M. Schwaller, W. Petersen, and E. Amitai (2012), Toward a framework for systematic error modeling of spaceborne precipitation radar with NOAA/NSSL ground radar-based national mosaic QPE, ,J. Hydrometeorol.,13(4), 1285–1300

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. J. Comput. Phys. 227: 3515–3539.

Lin Liu, Chunze Lin, Yongqing Bai, Dengxin He, "Assessing the Effects of Microphysical Scheme on Convective and Stratiform Characteristics in a Mei-Yu Rainfall Combining WRF Simulation and Field Campaign Observations", Advances in Meteorology, vol. 2020, Article ID 8231320, 17 pages, 2020. https://doi.org/10.1155/2020/8231320

Li J-L F, Forbes R M, Waliser D E, Stephens G and Lee S W 2014b Characterizing impacts of precipitating snow hydrometeors in the radiation using the ECMWF IFS global model J. Geophys. Res. Atmos. 119 10981–85

Mlawer, E.J.; Taubman, S.J.; Brown, P.D.; Iacono, M.J.; Clough, S.A. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. J. Geophys. Res. Atmos. 1997, 102, 16663–16682.

Noh, Y., W. G. Cheon, S-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. Bound.-Layer Meteor., 107 , 401–427.

Pinson P. 2012. Adaptive calibration of (u, v) wind ensemble forecasts. Q. J. R. Meteorol. Soc. 138: 1273–1284.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles 133, 1,155–1,174.

Roulin E, Vannitsem S. 2012. Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. Mon. Weather Rev. 140: 874–888.

Santer, H. M. and Grams H. M., Evaluation and Uncertainty of MRMS v12 Dual-polarized Radar Quantitative Precipitation Estimation Product, National Weather Center Research Experience for undergraduates, summer 2020, https://caps.ou.edu/reu/reu20/finalpapers/Santer-finalpaper.pdf.

Schmeits MJ, Kok KJ. 2010. A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. Mon. Weather Rev. 138: 4199–4211.

Schulz, B., and Lerch, S. (2022). Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison. Monthly Weather Review 150, 1, 235-257, https://doi.org/10.1175/MWR-D-21-0150.1

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2005), A description of the Advanced Research WRF version 2 (No. NCAR/TN-468+ STR). National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div.

Sloughter, J.M., Gneiting, T., Raftery, A.E., 2010. Probabilistic wind speed forecasting using ensembles and bayesian model averaging. Journal of the American Statistical Association 105, 25–35.

Sloughter, J.M.L., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using bayesian model averaging. Monthly Weather Review 135, 3209–3220.

Talagrand, O., R. Vautard and B. Strauss, 1999, Evaluation of Probabilistic Prediction Systems, in Proceedings of Workshop on Predictability (October 1997), ECMWF, Reading, England, 1-25, available at the address http://www.ecmwf.int/publications/library/do/references/list/16233.

Tao, Wei, Joanne Simpson, Deborah Baker, Scott A. Braun, Ming Dah Chou, Brad S. Ferrier, Daniel E. Johnson, Alexander Khain, Stephen E Lang, Barry H. Lynn, Chung-lin Shie, David Starr, C-H. Sui, Yansen Wang and Peter J. Wetzel. "Microphysics, Radiation and Surface Processes in the Goddard Cumulus Ensemble (GCE) Model." Meteorology and Atmospheric Physics 82 (2003): 97-137.

Taraphdar, S., & Pauluis, O. M. (2021). Impact of planetary boundary layer and cloud microphysics on the sensitivity of monsoon precipitation using a gray-zone regional model. Earth and Space Science, 8, e2020EA001535. https://doi.org/10.1029/2020EA001535

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. Mon. Wea. Rev., 136 , 5095–5115.

Thorarinsdottir, T.L., Gneiting, T., 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society: Series A (Statistics in Society) 173, 371–388.

Thorey, J., Chaussin, C., Mallet, V.. Ensemble forecast of photovoltaic power with online CRPS learning. International Journal of Forecasting, Elsevier, 2018, 34 (4), pp.762-773. ff10.1016/j.ijforecast.2018.05.007ff.

Van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. Quart. J. Roy. Meteor. Soc., 141, 807–818, https://doi.org/10.1002/qj.2397.

Van Schaeybroeck, B., and S. Vannitsem, 2011: Post-processing through linear regression. Nonlinear Processes Geophys., 18, 147–160, https://doi.org/10.5194/npg-18-147-2011.

Vannitsem S. 2009. A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts. Q. J. R. Meteorol. Soc. 135: 1801–1815.

Vannitsem S, Hagedorn R. 2011. Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods. Meteorol. Appl. 18: 94–104.

Vitart, F., Anderson, J.L., Sirutis, J., Tuleya, R.E.: Sensitivity of tropical storms simulated by a general circulation model to changes in cumulus parameterization. Quart. J. Roy. Meteor. Soc. 127, 25–51 (2001)

Wilks, D.S., 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications 16, 361–368.

Zhang, J., Howard, K., Langston, C., Vasiloff, S., Kaney, B., Arthur, A., Van Cooten, S., Kelleher, K., Kitzmiller, D., Ding, F., Seo, D., Wells, E., & Dempsey, C. (2011). National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans, Bulletin of the American Meteorological Society, 92(10), 1321-1338.

Appendix

Appendix A

Model parameter	**Used configuration**
Model and domains	
Model version	ARWv4.0 (Skarmarock et al. 2008)
Time step	Adaptative time step (36 s for D1)
Map projection	Lambert
Pressure top	50 hPa
Vertical levels	80 (*)
Time integration scheme	Third order Runge-Kutta scheme
Time integration scheme for acoustic and gravity-wave modes	Second order scheme

Horizontal/vertical advection	Fifth order upwind
Scalar advection	Positive definite
Upper-level damping (for vertical propagating gravity waves)	Rayleigh damping
Computational horizontal diffusion	6th-order numerical diffusion
Forecast period	60 h (from July 15th, 2018 at 12 pm UTC to July 18th, 2021 at 12 am UTC)

*Table [1]: WRF model configuration and input physics parameterizations. * η levels are 1, 0.99938147, 0.9918859506, 0.9860143, 0.9835575, 0.97480931, 0.9691238, 0.95061912, 0.938789424, 0.91847208, 0.89114445, 0.87771024, 0.8344125, 0.807124586, 0.76820505, 0.71652851, 0.6848121, 0.615978875, 0.5720332, 0.5472062, 0.5233661, 0.5004734, 0.4784906, 0.4573815, 0.4371113, 0.4176468, 0.3989559, 0.3810079, 0.3637731, 0.3472234, 0.3313315, 0.316071, 0.3014172, 0.2873457, 0.2738335, 0.2608584, 0.2483989, 0.2364347, 0.2249459, 0.2139138, 0.2033201, 0.1931475, 0.1833792, 0.173999, 0.1649918, 0.1563425, 0.1480369, 0.1400615, 0.132403, 0.1250489, 0.1179871, 0.111206, 0.1046944, 0.09844154, 0.09243726, 0.08667168, 0.08113512, 0.07581868, 0.07071351, 0.06581128, 0.06110381, 0.0565835, 0.05224282, 0.04807468, 0.04407217, 0.04022875, 0.0365381, 0.03299413, 0.02959097, 0.02632311, 0.0231851, 0.02017184, 0.01727832, 0.0144998, 0.01183172, 0.00926967, 0.006809457, 0.004447003, 0.002178475, 0.*

Appendix B

Station ID	Latitude(ºN)	Longitude(ºW)	Height ASL (m)
VRB	27.6556	80.4179	8.00
PGD	26.9172	81.9914	8.00
MIA	25.7880	80.3169	4.00
SRQ	27.4014	82.5586	9.00

Table 1: List of the four ASOS stations in South Florida and their corresponding latitude, longitude and above sea-level (ASL) height.

Appendix C

$\bar{o}$=5.02 mm.h^{-1}	$\bar{s}$ (mm.h^{-1})	RMSE (mm.h^{-1})	R (%)	MBE (%)
Raw Ensemble Mean	29.71	53.92	14.49	2469.24

Calibrated Ensemble Mean	10.77	14.24	20.81	575.37

568 Table 1: Statistics of the raw and calibrated means over the PGD ASOS station

569

$\overline{o}$= **4.34 mm.h^{-1}**	$\overline{s}$ **(mm.h^{-1})**	**RMSE (mm.h^{-1})**	**R (%)**	**MBE (%)**
Raw Ensemble Mean	14.57	16.91	4.96	1023,87
Calibrated Ensemble Mean	13.79	15.09	8.39	940.71

570 Table 2: Statistics of the raw and calibrated means over the MIA ASOS station

571

$\overline{o}$= **2.54 mm.h^{-1}**	$\overline{s}$ **(mm.h^{-1})**	**RMSE (mm.h^{-1})**	**R (%)**	**MBE (%)**
Raw Ensemble Mean	15.57	17.07	8.30	1302.87
Calibrated Ensemble Mean	8.60	9.77	29.36	605.77

572 Table 3: Statistics of the raw and calibrated means over the SRQ ASOS station

573

=7.39 mm.h^{-1}	$\overline{s}$**(mm.h^{-1})**	**RMSE (mm.h^{-1})**	**R (%)**	**MBE (%)**
Raw Ensemble Mean	16.43	15.61	19.72	904.30
Calibrated Ensemble Mean	8.06	7.04	30.06	30.06

574 Table 4: Statistics of the raw and calibrated means over the VRB ASOS station

575

576 **Appendix D**

Statistical indicator	Definition
$\overline{s}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i}$
$\overline{o}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} o_i}$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i - o_i)^2}$

Correlation	$\frac{\sum_{i=1}^{n} (s_i - \bar{s})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n} (s_i - \bar{s})^2} \sqrt{\sum_{i=1}^{n} (o_i - \bar{o})^2}}$
MBE	$\frac{1}{n}\sum_{i=1}^{n}(c_i - o_i)$

Table 1: Definition of the statistics used in this work. o_i and s_i are the observed and simulated wind speeds at time i. n is the number of data.