Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL087839

Key Points:
- The localized probability-matched mean has distinct advantages over other ensemble consensus products for precipitation forecasts
- The computational performance of the localized probability-matched mean can be greatly improved by using a patch-based algorithm
- Localized probability-matched mean forecasts show good objective skill and earn high subjective ratings from human forecasters

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4

Correspondence to:
N. Snook,
nsnook@ou.edu

Citation:
Snook, N., Kong, F., Clark, A., Roberts, B., Brewster, K. A., & Xue, M. (2020). Comparison and verification of point-wise and patch-wise localized probability-matched mean algorithms for ensemble consensus precipitation forecasts. *Geophysical Research Letters*, *47*, e2020GL087839. https://doi.org/10.1029/2020GL087839

Received 5 MAR 2020
Accepted 25 MAY 2020
Accepted article online 30 MAY 2020

Comparison and Verification of Point-Wise and Patch-Wise Localized Probability-Matched Mean Algorithms for Ensemble Consensus Precipitation Forecasts

Nathan Snook[1], Fanyou Kong[1], Adam Clark[2], Brett Roberts[2,3,4], Keith A. Brewster[1], and Ming Xue[1,5]

[1]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK, USA, [2]NOAA/OAR/National Severe Storms Laboratory, Norman, OK, USA, [3]Cooperative Institute for Mesoscale Meteorological Studies, Norman, OK, USA, [4]NOAA/NCEP/Storm Prediction Center, Norman, OK, USA, [5]School of Meteorology, University of Oklahoma, Norman, OK, USA

Abstract When applied to precipitation on large forecast domains, the probability-matched ensemble mean (PM mean) can exhibit biases and artifacts due to using distributions from widely varying precipitation regimes. Recent studies have investigated localized PM (LPM) means, which apply the PM mean over local areas surrounding individual points or local patches, the latter requiring far fewer computational resources. In this study, point-wise and patch-wise LPM means are evaluated for 18–24-hr precipitation forecasts of a quasi-operational ensemble of 10 Finite-Volume Cubed-Sphere (FV3) forecast members. Point-wise and patch-wise LPM means exhibited similar forecast performance, outperforming PM and simple means in terms of fractions skill score and variance spectra while exhibiting superior bias characteristics when light smoothing was applied. Based on the results, an LPM mean using local patches of 60 × 60 km and calculation domains of 180 × 180 km is well suited for operational warm-season precipitation forecasting over the contiguous United States.

Plain Language Summary Weather and rainfall are often predicted using ensembles of numerical weather forecast models. The skill of the ensemble consensus is often better than any individual forecast, and valuable information about the range of possible outcomes and model uncertainty is gained. In this study, different methods for implementing a localized probability-matched mean (LPM mean) are examined. The LPM mean is designed to produce a more accurate consensus from a forecast ensemble while retaining local structures that other consensus methods fail to capture. Two LPM variations were examined for predicting accumulated precipitation—one computed at every model grid point and another on patches containing many nearby points. Both methods produced similar results and outperformed traditional ensemble consensus algorithms. The patch-based method took 1 to 2 orders of magnitude less time to compute. Operational weather providers should consider using the patch-based LPM mean algorithm to efficiently compute ensemble rainfall forecasts.

1. Introduction

An ensemble of numerical weather prediction (NWP) model forecasts generally provides a more accurate prediction than a single deterministic forecast, along with valuable information on the range of possible outcomes. Obtaining a judicious consensus from an ensemble can, however, be challenging, particularly for complex features with sharp boundaries such as convective storms, fronts, and areas of rainfall. One straightforward ensemble consensus method is to take the point-wise mean of the ensemble members. This simple mean often has lower error than individual ensemble members for large-scale features (e.g., Murphy, 1988; Snook et al., 2019; Snyder & Zhang, 2003). However, for forecasts containing complex small-scale features, or considerable variation and displacement of features among ensemble members (as often occurs for convective storms and precipitation), the simple mean is often overly smooth, high-biased in coverage, and low-biased in intensity, with very few extreme values (e.g., Clark, 2017; Ebert, 2001; Snook et al., 2019; Surcel et al., 2014).

To address the shortcomings of the simple mean for quantitative precipitation forecasts (QPF), Ebert (2001) introduced the probability-matched ensemble mean (PM mean), which calculates the simple mean but replaces its numerical values with the distribution of values found in the individual ensemble members. The resulting field has the spatial distribution of the simple mean without reduction in high-intensity values. In studies focusing on intense convective precipitation, the PM mean has consistently been found to produce more skillful forecasts than the simple mean (e.g., Clark et al., 2009; Schwartz et al., 2014; Snook et al., 2019; Xue et al., 2011).

While the PM mean is generally viewed as more useful than the simple mean for QPF, it still has shortcomings. For example, Surcel et al. (2014) found that the gains in skill from the PM mean result from reduced variability at small scales relative to individual ensemble members. This variability is important to retain because it can indicate to forecasters the spatial scale and variability of potential hazards (e.g., flash flooding, severe weather) forecast by individual ensemble members. Also, Clark (2017) found that, over large domains, the PM mean systematically redistributes precipitation amounts from different regions, resulting in expansion of large heavy precipitation regions and shrinking or elimination of smaller ones. Thus, the bias characteristics over localized regions in the PM mean can be very unrepresentative of individual ensemble members. Potential therefore exists for improving the PM mean by limiting the domain from which the distribution is calculated. To this end, Clark (2017) introduced the localized probability-matched mean (LPM mean), which calculated the PM mean at each grid point, restricting the calculation to a specified influence radius and mitigating the loss of local structure common to the PM mean (Surcel et al., 2014).

While Clark's point-wise LPM mean outperformed the PM mean at large scales and improved the spatial structure of the resulting forecast, it requires calculation of the PM mean within the influence radius at every model grid point, making the method exceedingly computationally expensive. To confer the benefits of the LPM mean with reduced computational expense, Snook et al. (2019) developed a patch-wise LPM mean. Herein the relative forecast performance and computational costs of point-wise and patch-wise LPM algorithms are examined and compared to those of the PM mean and simple mean using data from the 2018 Center for Analysis and Prediction of Storms (CAPS) Finite-Volume Cubed-Sphere (FV3) storm-scale ensemble forecast (SSEF). The point-wise and patch-wise LPM algorithms are described in section 2, and a brief summary of the CAPS FV3 SSEF is provided in section 3. The relative performance of the ensemble consensus products is presented in section 4. Finally, conclusions and discussions are presented in section 5.

2. The LPM Mean

The LPM mean applies the PM mean locally within a radius of influence around either a point (point-wise LPM) or a collection of nearby points (patch-wise LPM), to spatially limit the forecast distribution considered for the mean. This is achieved using a local calculation domain centered on (but extending beyond) the point or patch in question; only forecast values within this domain are used to compute the LPM mean for that point or patch. While Clark (2017) used a circular domain, a rectangular domain can also be used. Rectangular domains were used by CAPS for calculating the patch-wise LPM of the CAPS SSEF during testbed experiments in 2017 and 2018 (e.g., Snook et al., 2019) and will be used in this study.

A conceptual illustration of the point-wise and patch-wise LPM means is shown in Figure 1 for diagonally adjacent grid cells (Figure 1a) or patches (Figure 1b). For rectangular calculation domains, the point-wise LPM (Figure 1a) is a special case of the more general patch-wise LPM (Figure 1b) with a patch size of 1×1 grid points. To calculate the LPM mean, the PM mean (Ebert, 2001) is calculated over the calculation domain, and values within the patch are used for the LPM mean field. For the point-wise LPM, this calculation is performed once for each grid point (Figure 1a); for the patch-wise LPM, this calculation is performed over the calculation domain of each patch (Figure 2b, lighter shading), with values corresponding to points within the patch (Figure 2b, darker shading) retained for use in the resulting LPM field. All patches are then stitched together, and a smoother is optionally applied. Calculation domains of nearby patches have considerable overlap; this, optionally combined with the smoother, minimizes patch edge discontinuities in the LPM mean field.

The LPM mean configuration can be defined in terms of three factors: LPM patch dimensions (1×1 for the point-wise LPM), calculation domain dimensions, and smoother settings. The dimensions of the calculation

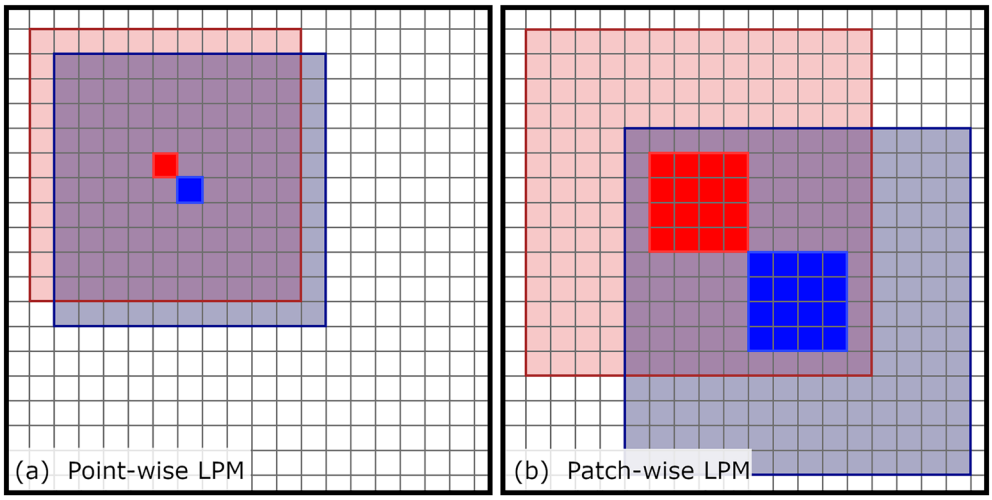

Figure 1. Conceptual illustration demonstrating (a) point-wise and (b) patch-wise LPM means. In (a), calculation domains (light red and blue shading) for the point-wise LPM mean are shown for two diagonally adjacent grid cells (bright red and blue shading). Panel (b) shows the same for the patch-wise LPM mean for two diagonally neighboring patches.

domain control the degree of localization, while the patch size strongly impacts the computational requirements of the LPM mean, as each patch requires one PM mean calculation (over that patch's calculation domain). To reduce computational expense, it is thus desirable to use the largest patches possible without degrading forecast quality. Appropriate smoother settings, including the choice of smoothing function and how aggressively to smooth, can help minimize patch discontinuities. The optimal choice of smoother settings depends upon a variety of factors, including the size of the domain and the spatial scale of the forecast features of greatest interest. For this study, patch-wise LPM means are calculated following Snook et al. (2019), using 60 × 60 grid point calculation domains and a Gaussian smoother with a standard deviation equal to the grid spacing (3 km).

3. Data

Data used in this study are from the CAPS FV3 SSEF, which was run during the 2018 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) (Zhang et al., 2019). The SSEF consists of 10 FV3 members run on a two-way nested grid; the global outer grid has a 13-km horizontal grid spacing, and the inner grid covers the contiguous United States (CONUS) at a variable horizontal grid spacing averaging approximately 3.5 km. Forecasts were initialized at 0000 UTC, from the 0000 UTC run of the operational Global Forecast System on weekdays between 30 April and 1 June 2018 (25 days in total). Prior to LPM mean forecast generation, nested-grid data are interpolated onto the Cartesian 3-km CONUS grid of the Community Leveraged Unified Ensemble (CLUE) (Clark et al., 2018).

Ensemble diversity is provided by variation in microphysical and planetary boundary layer (PBL) parameterizations. Microphysical schemes used include the Thompson (Thompson et al., 2008) and National Severe Storms Laboratory (NSSL) (Mansell et al., 2010) schemes. PBL schemes used include the Yonsei University PBL scheme (Hong et al., 2006) and its scale-aware variant (Shin & Hong, 2015), the Mellor-Yamada-Nakanishi-Niino (MYNN) scheme (Nakanishi & Niino, 2006) and its scale-aware variant, and the Eddy-Diffusivity Mass-Flux scheme (Han et al., 2016). These five PBL schemes and two microphysical schemes allow each of the 10 members of the SSEF to have a unique combination of microphysics and PBL scheme. Further details regarding SSEF model settings are discussed in Zhang et al. (2019). In this study, in order to compare and evaluate the performance of different ensemble consensus products, we will focus on forecasts of 6-hr accumulated rainfall valid 24 hr after forecast initialization.

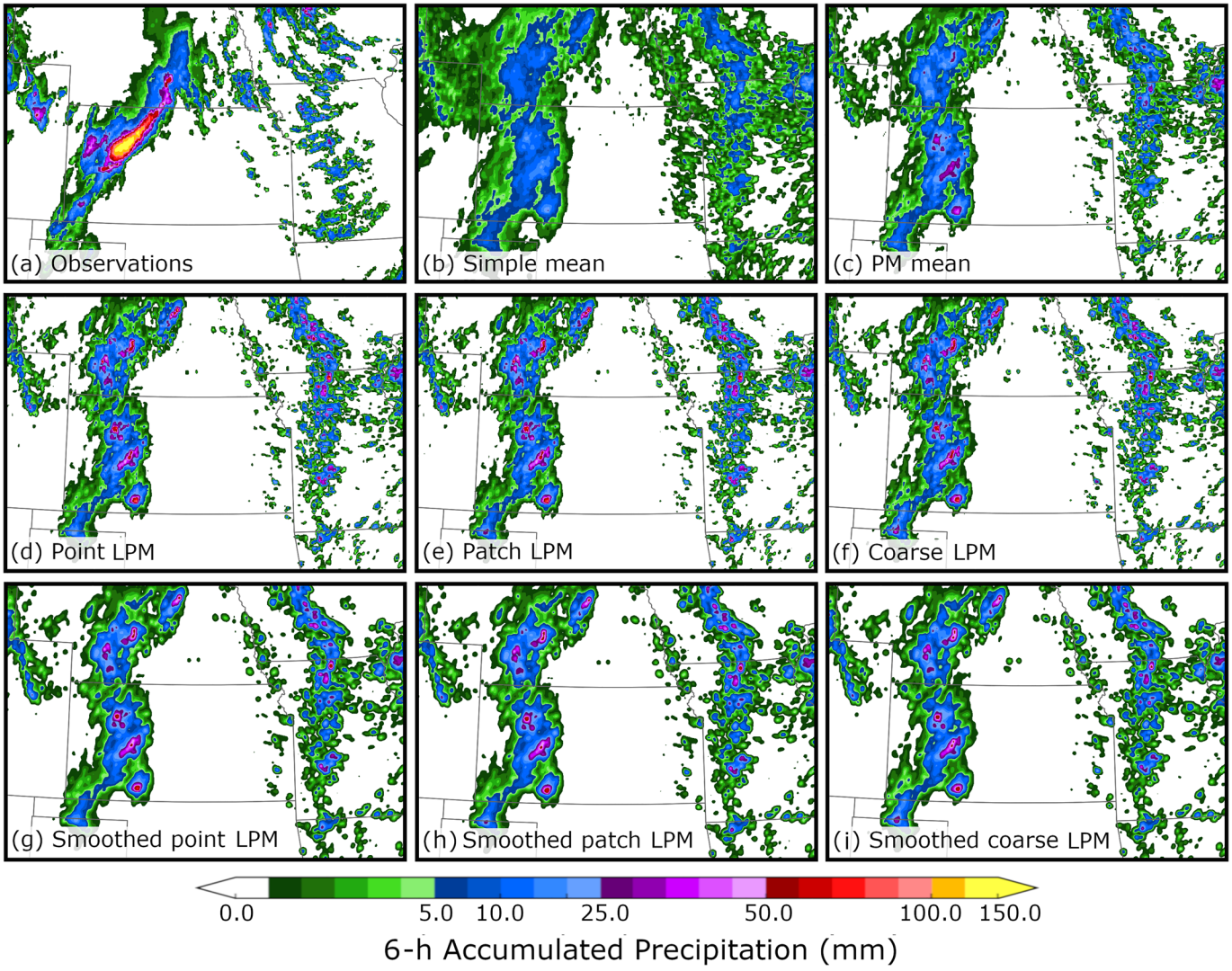

Figure 2. Comparison of (a) observed accumulated precipitation for 1800 UTC 28 May to 0000 UTC 29 May 2018, as well as ensemble consensus precipitation accumulation over the same period using (b) the simple mean, (c) the PM mean, (d) the point-wise LPM, (e) the patch-wise LPM, (f) a patch-wise LPM using coarse 40 × 40 patches, and (g–i) versions of the LPM variants shown in panels (d)–(f) smoothed using a Gaussian smoother with a standard deviation of 3 km.

4. Results

Verification of QPF is challenging, as point-by-point metrics such as the equitable threat score (ETS) are strongly impacted by phase and timing errors (e.g., Ebert, 2008). Neighborhood methods are thus generally preferred for verifying convection-allowing forecasts. We use several skill metrics to robustly evaluate model performance, including ETS, frequency bias, fractions skill score (FSS) (Roberts & Lean, 2008), and two-dimensional variance spectra (Denis et al., 2002; Surcel et al., 2014). The latter two methods are intended to objectively compare the structure of observed and forecast precipitation at different spatial scales. Because improved computational efficiency is a primary objective of the patch-wise LPM algorithm, statistics on execution time for point- and patch-wise LPM algorithms are also presented. Stage IV precipitation estimates (Lin & Mitchell, 2005) are chosen as the observation data set for QPF verification. Hourly Stage IV data, originally on a 4-km grid, are interpolated to the 3-km CLUE grid and summed to correspond to the 6-hr precipitation forecasts being verified.

4.1. Spatial Structure of LPM Mean Forecasts

To illustrate the spatial structure of different ensemble consensus products, forecasts of 6-hr accumulated precipitation from 1800 UTC 28 May to 0000 UTC 29 May 2018 are compared in Figure 2 for the simple

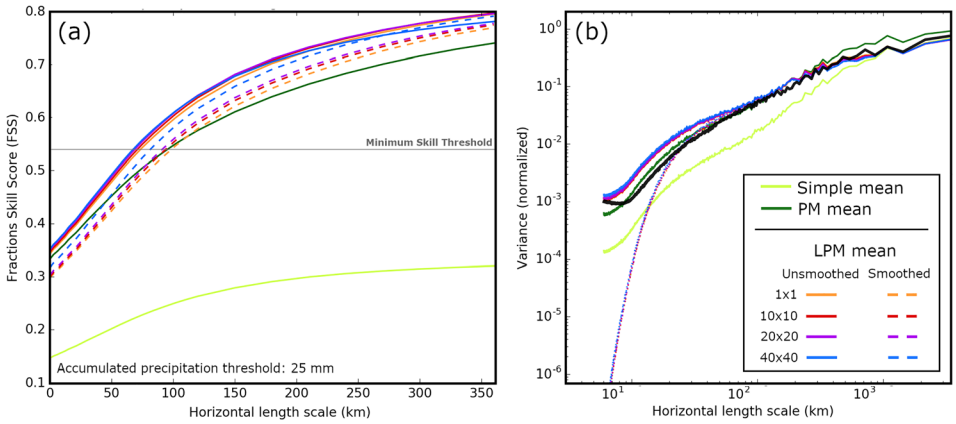

Figure 3. Mean (a) fractions skill score (FSS) and (b) normalized variance spectra, plotted as a function of horizontal length scale, calculated over the 2018 HWT SFE period, for 6-hr accumulated precipitation in 18–24-hr CAPS SSEF forecasts valid 0000 UTC the day after initialization. Shown are the simple mean (light green), PM mean (dark green), and various LPM mean configurations. LPM means include smoothed (dotted lines) and unsmoothed (solid lines) variants using both point-wise (orange) and patch-wise algorithms, including patches of three sizes (red: 10 × 10; purple: 20 × 20; blue: 40 × 40). The minimum skillful FSS value is indicated by the horizontal gray line in panel (a). The mean normalized variance spectrum of Stage IV rainfall accumulation is shown (thick black line) for comparison in panel (b).

and PM means, as well as point-wise and patch-wise LPM means, including smoothed variants and patch sizes of 10 × 10 (Figures 2e and 2h) and 40 × 40 (Figures 2f and 2i) grid points, for precipitation associated with convective storms over the Great Plains. Observed precipitation is shown for comparison (Figure 2a). The characteristic smoothed features and limited magnitude of the simple mean is evident, particularly over western Kansas (Figure 2b). The PM mean includes some more intense precipitation (Figure 2c) but, like the simple mean, smooths out localized precipitation maxima in, for example, southern Nebraska. All LPM mean variants (Figures 2d–2i) exhibit very similar spatial structure over this subdomain; the LPM means (Figures 2d–2i) exhibit greater localized variation in precipitation magnitude compared to the PM mean (Figure 2c). The higher maxima in the LPM means better represent the forecasts from individual ensemble members (see supporting information Figures S1 and S2) than the simple mean (Figure 2b) and the PM mean (Figure 2c). However, because none of the individual ensemble members forecast the extremely high observed precipitation amounts in northwest Kansas, none of the methods capture the most intense observed maxima. Patch boundary discontinuities are largely absent in the LPM mean fields, evident only in the unsmoothed coarse LPM (e.g., in southwestern Kansas), suggesting that relatively large patches can feasibly be used. These general patterns hold true in other portions of the domain, including for a landfalling tropical cyclone along the gulf coast (see Figures S2–S4).

Verification from the full 2018 HWT SFE period using FSS with an accumulated precipitation threshold of 25 mm is presented in Figure 3a. FSS measures the similarity between a forecast and observations in terms of fractional coverage of an event at different spatial scales and can be used to determine the minimum scale at which the forecast has useful skill (Roberts & Lean, 2008). The simple mean performs quite poorly in terms of FSS, failing to attain a skillful FSS at any scale. The PM mean exhibits a minimum skillful scale (the smallest scale for which the FSS exceeds the useful skill threshold of 0.54) of approximately 95 km, compared to 65–75 km (unsmoothed) and 80–100 km (smoothed) for the LPM means. The unsmoothed LPM outperforms the PM in terms of FSS at all scales; smoothing reduces FSS performance of the LPM slightly, though even the smoothed LPM outperforms PM at scales larger than the minimum skillful scale. The point-wise and patch-wise LPM configurations perform similarly, although when weak smoothing is applied, LPM calculated using larger local patches exhibits slightly higher FSS (Figure 3a).

The spatial structure of 6-hr accumulated precipitation in LPM, PM, and simple mean forecasts is compared to that of observed 6-hr precipitation using variance spectra in Figure 3b. As for the FSS in Figure 3a, spectra are averaged for 18- to 24-hr forecasts of 6-hr accumulated precipitation valid 0000 UTC the day after forecast initialization over the full 2018 HWT SFE period. The mean spectrum of Stage IV estimated 6-hr

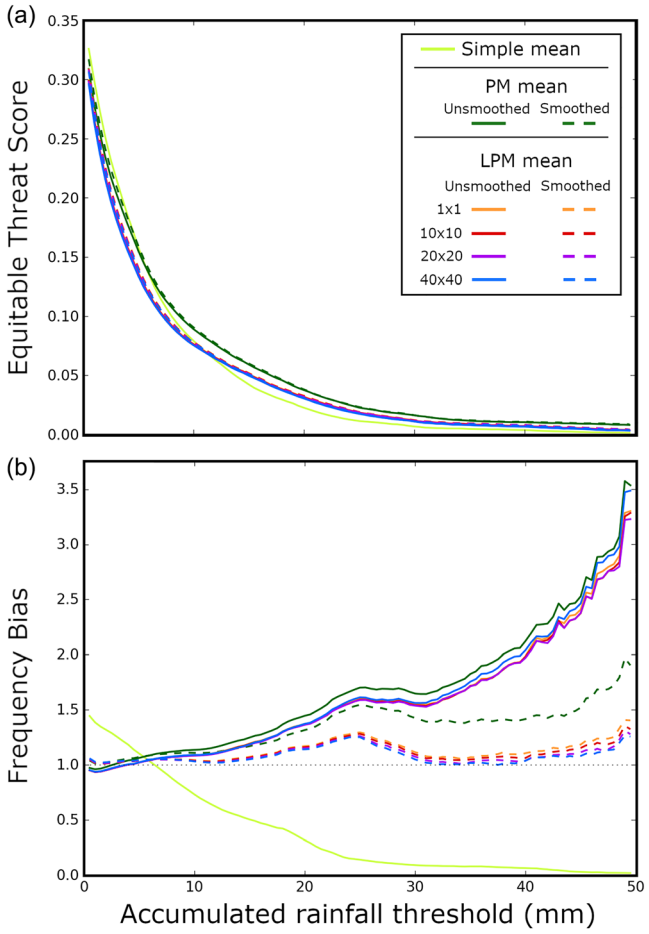

Figure 4. (a) Equitable threat score (ETS) and (b) frequency bias (BIAS), calculated over the 2018 HWT SFE period, for 18–24-hr accumulated precipitation forecasts valid 0000 UTC the day after initialization. Shown are the simple mean (light green), PM mean (dark green), smoothed PM mean (dashed dark green), and various LPM means. LPM means include smoothed (dotted lines) and unsmoothed (solid lines) variants using point-wise (orange) and patch-wise algorithms with patches of several sizes (red, purple, and blue). In panel (b), the horizontal dotted gray line indicates an unbiased forecast (BIAS of 1.0).

precipitation accumulation is also shown. The overly smooth simple mean exhibits much lower variance than the observations at all scales smaller than 1,000 km, while the PM and LPM means remain much closer to the observed spectrum. However, the PM mean overpredicts variance at scales larger than about 300 km, while the spectra of the LPM means remain very close to the observed spectrum even at the largest scales. This is consistent with behavior reported for 2016 and 2017 ensemble forecast results using the patch-wise LPM with a WRF-ARW ensemble in Snook et al. (2019). LPM patch size has little impact on variance spectra for patches of up to 40 × 40 grid points (Figure 3b). Difference in variance spectra among LPM configurations is largely attributable to smoothing; when smoothing is applied, variance is drastically reduced for scales smaller than about 40 km (this is expected, as smoothing using a Gaussian filter with a standard deviation of one grid point is designed to minimize near-grid-scale noise and patch boundary discontinuities).

4.2. LPM Mean Forecast Skill

While our main focus is on spatial structure, as the LPM mean is designed to retain local features, it is also important to consider forecast skill using more traditional skill metrics. In Figure 4, forecast skill is evaluated using ETS (Figure 4a) and frequency bias (Figure 4b). ETS of 18- to 24-hr rainfall forecasts is greatest for light precipitation, with little skill for thresholds exceeding 20 mm. ETS is similar among all ensemble consensus products, though it is slightly higher for the PM than for the LPM for all thresholds. The simple mean exhibits the poorest performance for thresholds greater than approximately 10 mm (Figure 4a). Applying smoothing results in very little change to ETS for both PM and LPM products. Note that, at the grid-scale, Clark (2017) also found that the PM mean performed better than the LPM mean, but when increasingly large neighborhoods were used to compute ETS, this reversed with the LPM mean having higher scores.

For frequency bias (Figure 4b), the simple mean exhibits a severe low bias for thresholds exceeding 8 mm, while the PM and the unsmoothed LPM means exhibit a high bias for larger thresholds; this high bias is largest in the PM mean. LPM mean bias behavior is greatly improved when weak Gaussian smoothing is applied, with bias of near 1.0 over all thresholds. When the same smoothing is applied to the PM mean, a substantial high bias remains, particularly for thresholds exceeding 20 mm (Figure 4b). In terms of both bias and ETS, the LPM patch size has little impact on forecast performance (Figure 4b). Note that, although both smoothed and unsmoothed LPM means have biases closer to 1.0 relative to the PM mean, this indicates the biases of the LPM means are generally lower than individual ensemble members. Clark (2017) noted similar behavior for the LPM mean but applied a function to slightly inflate precipitation amounts so that they were more in line with the members. Biases consistent with ensemble members may or may not be desirable depending on preferences of the user and/or ensemble characteristics.

4.3. Computational Expense of the LPM Mean

Because the LPM mean requires performing multiple PM mean calculations on overlapping subdomains, it is much more computationally expensive than the PM mean and simple means. To examine differences in computational expense in detail, PM and simple means and point-wise and patch-wise LPM means were calculated 100 times for a 240 × 240 subdomain. The average runtime is shown in Figure 5. Runtime varied by at most 2% for any configuration during testing. Timing tests were performed using Python (including NumPy) implementations of the products on one processor of a Linux (CentOS 6) cluster. Figure 5 uses a logarithmic scale for wallclock time.

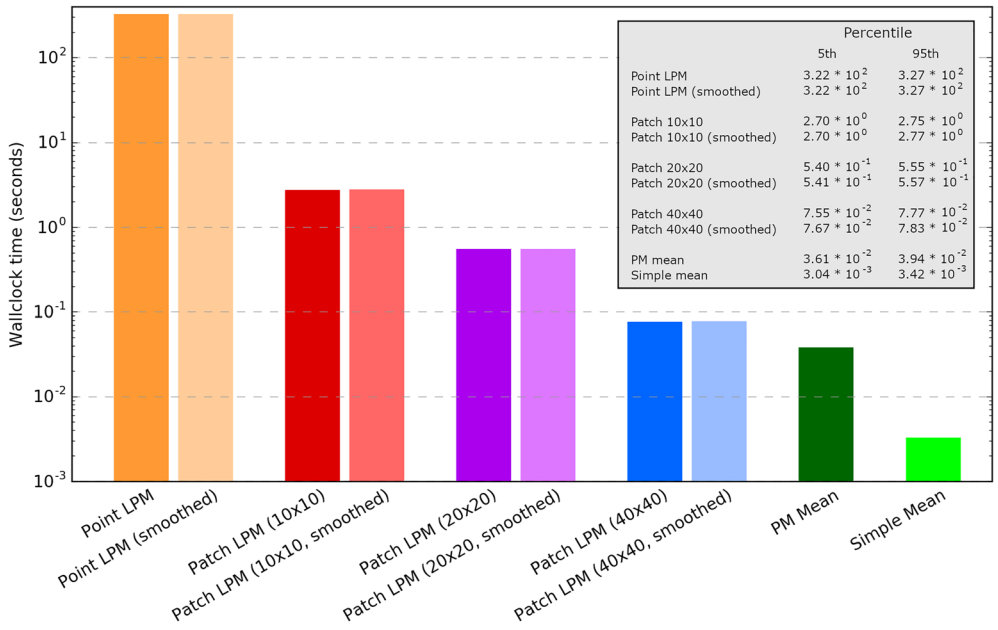

Figure 5. Wallclock runtime for Python implementations of the simple mean, PM mean, and various configurations of the point-wise and patch-wise LPM mean over a 240 × 240 grid point subdomain for 10 members at one forecast time. The 5th and 95th percentiles for timings for each configuration are shown in the table in the upper right.

The LPM mean is substantially more computationally expensive than the PM and simple means, with computation time depending strongly on patch size. The point-wise LPM mean requires 2 orders of magnitude more computation time than the 10 × 10 patch-wise LPM mean; this disparity increases to more than 3 orders of magnitude compared to the 40 × 40 patch-wise LPM mean. The computational cost of smoothing is negligible compared to the impacts of patch size. Because of the high computational cost of using the point-wise LPM or patch-wise LPM with small patches, relatively large LPM patches are likely necessary for real-time implementations.

5. Summary and Conclusions

Point-wise and patch-wise implementations of the LPM mean are compared using the 2018 CAPS FV3 SSEF. The LPM mean applies probability matching over localized domains, either around a grid point or a local patch, to generate an ensemble consensus forecast which retains small-scale structures and uses only local information. For prediction of 6-hr accumulated precipitation using a convection-allowing model on a continental scale, patches of up to 20 × 20 grid points can be used without introducing noticeable discontinuities along patch boundaries.

The LPM mean performs similarly to the PM mean in terms of ETS and exhibits superior bias behavior, particularly when weak smoothing is applied, though for ensembles with different bias properties tuning of the smoother might be required. When spatial structure is considered, the LPM mean outperforms the PM and simple means, exhibiting a substantially smaller minimum skillful scale in terms of FSS and higher FSS overall for scales greater than around 100 km. The LPM mean also exhibits superior precipitation variance spectra compared to the simple mean (which underpredicts variance at small scales) and the PM mean (which overpredicts variance at large scales). Varying the size of the LPM patch from a single point to 40 × 40 points resulted in only small changes to precipitation forecast performance for the objective skill metrics considered.

While the LPM mean, even in its patch-wise form, is more computationally expensive than the PM and simple means, its ability to retain local structures and magnitudes is valuable. For example, precipitation forecasts using the patch-wise LPM mean produced by CAPS during the 2017 and 2018 Hydrometeorology Testbed Flash Flood and Intense Rainfall (FFaIR) experiments were subjectively rated as more useful

than simple or PM mean forecasts by FFaIR participants (Albright & Perfater, 2018). Based upon these tests, the patch-wise LPM mean was recommended for operational implementation and has been tested in a prototype version 3 of the High Resolution Ensemble Forecast (HREFv3), where it was found to exhibit good performance, particularly in terms of bias (M. Pyle, personal communication, 18 July 2019), though it underperformed PM for ETS at low precipitation amounts.

It should be noted that CAPS SSEF consists of only 10 members; for larger ensembles, differences among the ensemble mean, the PM mean, and the LPM mean might be larger due to increased diversity among a larger number of ensemble members. We also note that the impact of using the LPM mean instead of the PM mean is greatest when the size of the local domains used in the LPM are much smaller than the full forecast domain, so the potential difference between the LPM mean and the PM mean will be less for smaller regional forecast domains or when very large local LPM domains are used. For operational implementations, using LPM patches which are as large as possible without introducing patch boundary discontinuities is generally optimal given the strong inverse correlation between LPM patch size and LPM mean runtime. Optimal smoother choice depends upon a variety of factors, including the size of the forecast domain and the spatial scale of the features of greatest interest. Overall, given the extreme computational cost of the point-wise LPM mean (or a patch-wise LPM mean with small patches) and the relative similarity in objective forecast performance of different LPM configurations, a patch-wise LPM mean using relatively large patches (~20 × 20 grid points), with weak smoothing to improve bias performance, is recommended for precipitation forecasting with convection-allowing NWP model ensembles.

Data Availability Statement

Precipitation data used in this study are available online (at ftp://ftp.caps.ou.edu/snook_2020_grl).

Acknowledgments

This work was supported by NOAA Testbed Grants NA17OAR4590120 and NA19OAR4590141 with supplementary support from NOAA CSTAR Grant NA16NWS4680002. Supercomputing resources were provided by NSF XSEDE, primarily at the University of Texas Advanced Computing Center. Some postprocessing was performed using University of Oklahoma Supercomputing Center for Education and Research resources.

References

Albright, B., & Perfater, S., (2018). 2018 Flash Flood and Intense Rainfall Experiment findings and results. Retrieved from https://www.wpc.ncep.noaa.gov/hmt/2018_FFaIR_final_report.pdf

Clark, A. J. (2017). Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Weather and Forecasting*, *32*(4), 1569–1583. https://doi.org/10.1175/WAF-D-16-0199.1

Clark, A. J., Gallus, W. A. Jr., Xue, M., & Kong, F. (2009). A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather and Forecasting*, *24*(4), 1121–1140. https://doi.org/10.1175/2009WAF2222222.1

Clark, A. J., Jirak, I. L., Dembek, S. R., Creager, G. J., Kong, F., Thomas, K. W., et al. (2018). The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting experiment. *Bulletin of the American Meteorological Society*, *99*(7), 1433–1448. https://doi.org/10.1175/BAMS-D-16-0309.1

Denis, B., Cote, J., & Laprise, R. (2002). Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, *130*(7), 1812–1829. https://doi.org/10.1175/1520-0493(2002)1302.0.CO;2

Ebert, E. E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, *129*(10), 2461–2480. https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, *15*(1), 51–64. https://doi.org/10.1002/met.25

Han, J., Witek, M. L., Teixeira, J., Sun, R., Pan, H. L., Fletcher, J. K., & Bretherton, C. S. (2016). Implementation in the NCEP GFS of a hybrid eddy-diffusivity mass-flux (EDMF) boundary layer parameterization with dissipative heating and modified stable boundary layer mixing. *Weather and Forecasting*, *31*(1), 341–352. https://doi.org/10.1175/WAF-D-15-0053.1

Hong, S.-Y., Noh, Y., & Dudhia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, *134*(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

Lin, Y., & Mitchell, K. E., (2005). The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. Hydrology*, San Diego, CA: American Meteorological Society

Mansell, E. R., Ziegler, C. L., & Bruning, E. C. (2010). Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *Journal of the Atmospheric Sciences*, *67*(1), 171–194. https://doi.org/10.1175/2009JAS2965.1

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Nakanishi, M., & Niino, H. (2006). An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, *119*(2), 397–407. https://doi.org/10.1007/s10546-005-9030-8

Pyle, M., (2019). Personal communication. 18 Jul. 2019.

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Schwartz, C. S., Romine, G. S., Smith, K. R., & Weisman, M. L. (2014). Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Weather and Forecasting*, *29*(6), 1295–1318. https://doi.org/10.1175/WAF-D-13-00145.1

Shin, H. H., & Hong, S.-Y. (2015). Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Monthly Weather Review*, *143*(1), 250–271. https://doi.org/10.1175/MWR-D-14-00116.1

Snook, N., Kong, F., Brewster, K., Xue, M., Albright, B., Perfater, S., et al. (2019). Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–2017 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall experiments. *Weather and Forecasting*, *34*(3), 781–804. https://doi.org/10.1175/WAF-D-18-0155.1

Snyder, C., & Zhang, F. (2003). Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Monthly Weather Review*, *131*(8), 1663–1677. https://doi.org/10.1175//2555.1

Surcel, M., Zawadzki, I., & Yau, M. K. (2014). On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Monthly Weather Review*, *142*(3), 1093–1105. https://doi.org/10.1175/MWR-D-13-00134.1

Thompson, G., Field, P. R., Rasmussen, R. M., & Hall, W. D. (2008). Explicit forecasts of winter precipitation using an improved bulkmicrophysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, *136*(12), 5095–5115. https://doi.org/10.1175/2008MWR2387.1

Xue, M., Kong, F., & Tomas, K. W., Wang, Y., Brewster, K. A., Gao, J., et al. (2011). Realtime convection-permitting ensemble and convection-resolving deterministic forecasts of CAPS for the Hazardous Weather Testbed 2010 Spring Experiment. *25th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 9A.2.

Zhang, C., Xue, M., Supinie, T. A., Kong, F., Snook, N., Thomas, K. W., et al. (2019). How well does the FV3 model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA Hazardous Weather Testbed with different physics combinations. *Geophysical Research Letters*, *46*, 3523–3531. https://doi.org/10.1029/2018GL081702