Verification of Convection-Allowing WRF Model Forecasts of the Planetary Boundary Layer Using Sounding Observations

MICHAEL C. CONIGLIO
NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

JAMES CORREIA JR.
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, and NOAA/NWS/NCEP/Storm Prediction Center, Norman, Oklahoma

PATRICK T. MARSH
NOAA/OAR/National Severe Storms Laboratory, and Cooperative Institute for Mesoscale Meteorological Studies, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

FANYOU KONG
Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

(Manuscript received 1 October 2012, in final form 8 January 2013)

ABSTRACT

This study evaluates forecasts of thermodynamic variables from five convection-allowing configurations of the Weather Research and Forecasting Model (WRF) with the Advanced Research core (WRF-ARW). The forecasts vary only in their planetary boundary layer (PBL) scheme, including three “local” schemes [Mellor–Yamada–Janjić (MYJ), quasi-normal scale elimination (QNSE), and Mellor–Yamada–Nakanishi–Niino (MYNN)] and two schemes that include “nonlocal” mixing [the asymmetric cloud model version 2 (ACM2) and the Yonei University (YSU) scheme]. The forecasts are compared to springtime radiosonde observations upstream from deep convection to gain a better understanding of the thermodynamic characteristics of these PBL schemes in this regime. The morning PBLs are all too cool and dry despite having little bias in PBL depth (except for YSU). In the evening, the local schemes produce shallower PBLs that are often too shallow and too moist compared to nonlocal schemes. However, MYNN is nearly unbiased in PBL depth, moisture, and potential temperature, which is comparable to the background North American Mesoscale model (NAM) forecasts. This result gives confidence in the use of the MYNN scheme in convection-allowing configurations of WRF-ARW to alleviate the typical cool, moist bias of the MYJ scheme in convective boundary layers upstream from convection. The morning cool and dry biases lead to an underprediction of mixed-layer CAPE (MLCAPE) and an overprediction of mixed-layer convective inhibition (MLCIN) at that time in all schemes. MLCAPE and MLCIN forecasts improve in the evening, with MYJ, QNSE, and MYNN having small mean errors, but ACM2 and YSU having a somewhat low bias. Strong observed capping inversions tend to be associated with an underprediction of MLCIN in the evening, as the model profiles are too smooth. MLCAPE tends to be overpredicted (underpredicted) by MYJ and QNSE (MYNN, ACM2, and YSU) when the observed MLCAPE is relatively small (large).

1. Introduction

Planetary boundary layer (PBL) schemes that parameterize vertical turbulent fluxes of heat, moisture, and momentum in the atmosphere are required in current operational numerical weather prediction (NWP) models because their grid spacing is coarse relative to the spatial and temporal scales of turbulence. This includes convection-allowing NWP models because the horizontal length scales of the largest eddies are typically smaller than the grid spacing (1–4 km). Uncertainty and inaccuracy in forecasts of atmospheric state variables in the PBL that result from these schemes (Hacker 2010;
2. Data and methodology

a. Model description

For the HWT Spring Forecasting Experiment, the 2011 and 2012 CAPS ensemble used 4-km grid spacing, was initialized weekdays at 0000 UTC, and was integrated for 36 h over a domain covering the contiguous United States (see Kain et al. 2013 for more details). For the five members of the ensemble evaluated here, all physics options other than the PBL scheme were identical, including Thompson microphysics [Thompson et al. (2008), with updates provided by G. Thompson (2011, personal communication)], Rapid Radiative Transfer Model (RRTM) longwave radiation (Mlawer et al. 1997), and Goddard shortwave radiation (Chou and Suarez 1994). All five forecasts used the Noah land surface model (LSM; Chen and Dudhia, 2001; Ek et al. 2003), so any differences in profiles in the PBL can be attributed to the PBL schemes alone. In 2011 (2012), WRF-ARW version 3.2.1 (3.3.1) was used with 51 vertical levels stacked closest near the ground, with approximately twelve vertical levels below 1 km AGL and the lowest model level approximately 24 m above ground level (AGL; hereafter all references to height are AGL). Radial velocity and reflectivity data from Weather Surveillance Radars-1988 Doppler (WSR-88Ds) and surface observations were assimilated into the initial conditions of these members identically using a three-dimensional variational data assimilation (3DVAR; Xue et al. 2003; Gao et al. 2004) data and cloud analysis (Hu et al. 2006) system. NAM 12-km analyses were used as the 3DVAR analysis background, and corresponding NAM forecasts were used for lateral boundary conditions in all five members. The five convection-allowing model (CAM) forecasts are benchmarked against forecasts from the operational 12-km NAM forecasts from the grid point nearest to the observed sounding. Note that the NAM uses a version of the MYJ scheme, but the configurations of the NAM and the MYJ member of the WRF-ARW ensemble studied here are too diverse to isolate the causes of their differences in the PBL structures.

b. PBL schemes

Two of the five CAM forecasts evaluated in this study use the MYJ and YSU schemes. The other three forecasts use the quasi-normal scale elimination (QNSE) scheme (Sukoriansky et al. 2005), the asymmetric cloud model version 2 (ACM2) scheme (Pleim 2007), and MYNN. In all of the schemes tested, mean turbulent quantities of the form \((w\,\bar{\phi})\) are estimated by

\[
-(w^f \phi') = K_\phi \frac{\partial \bar{\phi}}{\partial z},
\]

where \(K_\phi\) are coefficients determined by the scheme, primes denote perturbation quantities, \(w\) is the vertical velocity, and \(\phi\) is either temperature, water vapor mixing ratio, or the horizontal wind components. In MYJ, MYNN, and QNSE, \(K_\phi\) for each variable are functions of local turbulent kinetic energy (TKE), the specified mixing length, and the dimensionless stability functions that are specified for the scheme. All three schemes used here are variations of the MYJ level-2.5 scheme (Mellor and Yamada 1982) and diagnose potential temperature...
variance, water vapor mixing ratio variance, and their covariances to solve a prognostic equation for TKE. All three of these schemes are often referred to as “local” schemes since they use model fields only at adjacent levels to determine the unknown variables.

There are several differences between MYJ and MYNN. MYNN uses the liquid water potential temperature and total water content as the thermodynamic variables, considers the effects of buoyancy in the diagnosis of the pressure covariance terms, and uses closure constants in the stability functions and mixing length formulations that are based on large eddy simulation (LES) results instead of observational datasets. Furthermore, MYNN employs mixing length formulations that are more flexible across the stability spectrum compared to MYJ and QNSE (Nakanishi 2001; Nakanishi and Niino 2004, 2009). The QNSE scheme is similar to MYJ, except the diffusivities under stable conditions are calculated from spectral theory to account for internal wave generation in the presence of turbulence (Sukoriansky et al. 2005).

YSU and ACM2 are often referred to as “nonlocal” schemes because they attempt to simulate the effects of large eddies in the convective PBL using model fields at many vertical levels in the lower troposphere. In YSU, the $K_\phi$ terms in Eq. (1) are given by

$$K_\phi = k z \frac{u_*}{\varphi} \left[ 1 - \left( \frac{z}{h} \right)^2 \right],$$

where $k$ is the von Kármán constant (0.4), $z$ is the height above ground level (AGL), $u_*$ is the friction velocity, $h$ is the diagnosed height of the boundary layer, and $\varphi$ is an empirical function of $h$ and the Obukhov length scale. For heat and momentum in the YSU scheme, a nonlocal gradient correction term is added to Eq. (1) to represent the effects of large-scale eddies. YSU also adds a term to Eq. (1) to represent entrainment flux at the diagnosed boundary layer top where the stability is large. In ACM2, the mixing at any given level within the PBL is the sum of a local formulation given by Eqs. (1) and (2), as well as three terms representing nonlocal upward and downward mixing with the eddy diffusivity given by Eq. (2). A weighting factor is used to determine the portion of the mixing due to local and nonlocal transport. Under convective boundary layers, the nonlocal mixing term usually dominates the local mixing term through most of the depth of the PBL, but the local mixing term can be large near the top of the PBL where the wind shear can be large and turbulent mixing can continue despite a stable stratification. Under stable conditions, the portion of mixing due to nonlocal transport is set to zero.

![Fig. 1. Maps of the locations of the uncontaminated soundings used for verification.](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)
c. Evaluation methods

This analysis is part of an ongoing effort at CAPS and the HWT to evaluate the performance of convection-allowing configurations of the WRF-ARW for severe weather forecasting applications. The evaluation of the PBL scheme’s performance will be done in multiple parts, with this study focusing on the thermodynamic accuracy of the forecasts using radiosonde observations as truth. The set of observed soundings is mostly composed of National Weather Service (NWS) operational soundings, but a few soundings obtained during the Mideast Convective Clouds Experiment (MC3E) and the HWT in 2012 are used in the verification.

The goal of this study is to examine the performance characteristics of the PBL schemes in the potentially unstable inflow to convection. Only those soundings that occurred approximately within 400 km and 6 h of deep convection in either the model forecasts or in reality are considered here. To examine the environmental PBL profiles, it is important to exclude those soundings that were contaminated by convection or precipitation. For the observed soundings, contamination was determined using radar reflectivity and quantitative precipitation estimates from the National Mosaic and Multisensor Quantitative Precipitation Estimation project (NMQ; Zhang et al. 2011). For the model soundings, contamination was determined using the model 1-h accumulated precipitation fields and simulated composite reflectivity. A proximity sounding generally was not used if a location within 10 grid points (~40 km) of the sounding experienced ≥0.5 mm of precipitation or composite reflectivity ≥20 dBZ in the 3 h prior to the sounding. Furthermore, any sounding used must be uncontaminated from the analysis up to 15 h into the forecast. This was done to keep the dataset composed of the same soundings through the transition from the nocturnal to daytime boundary layer. Finally, soundings that were uncontaminated by convection, but were deemed to be too close to a meso-α- or larger-scale boundary were removed, as it is important to make sure that the air

![Fig. 2. (top) Hourly evolution of the mean (solid lines, left axis) and standard deviation (dashed lines, right axis) of diagnosed PBL height (h_d; m AGL) through the full 36 forecast hours. Model forecasts are colored according to the legend in the top left and the mean and standard deviation values for the observed soundings are displayed by the open circles and asterisks, respectively. (bottom) The number of soundings in the dataset (dashed line) and the number of days with at least one sounding (solid) for each hour.](image-url)
The mass sampled by the radiosonde is analogous to the air mass sampled in all the model forecasts. The above procedure results in 211 locations with uncontaminated soundings in both the five model forecasts and in the observations, but the verification dataset at observation times was reduced further through manual quality control (QC). All observed soundings were inspected one by one on skew T–logp diagrams and those that displayed unusual or suspicious low-level thermodynamic profiles were removed. A few soundings with missing humidity data also were removed. After the above QC procedures were applied, 192 model and observed soundings remain for evaluating the model analyses (valid at 0000 UTC) and 191 soundings remain for evaluating the morning model forecasts (valid shortly after 1100 UTC, the hour closest to actual radiosonde release time; see Figs. 1a and 1b). The number of soundings retained in the dataset declines steadily from 15 through 36 h as convection develops in the models and contaminates the forecasts. There are 100 soundings in the verification dataset for the evening forecasts (mostly valid shortly after 2300 UTC, once again, the hour closest to actual radiosonde release time, although there are a few instances of radiosonde releases closer to 0000 UTC; see Fig. 1c). After applying the proximity criteria and removing the contaminated soundings, only 21 observed soundings from late-morning to midafternoon (1500–2100 UTC) remain. Therefore, the verification is restricted to the more data-rich times in the morning and evening when NWS radiosondes are routinely available.

One final note on the observed soundings that could be relevant to our results for the evening sounding comparisons concerns the work of Weckwerth et al. (1996), who show that potential temperature can vary by as much as 0.5 K between updraft and downdraft branches of horizontal convective rolls and mixing ratio can vary by as much at 1.5–2.5 g kg\(^{-1}\). While this could hinder the accuracy of comparisons on individual days, we believe a dataset of 100 soundings is sufficiently large to average out these variations that may result from boundary layer roll circulations.

Data for the observed soundings and from the NAM soundings are interpolated to the vertical levels used in the CAM forecasts prior to the computation of the variables and the aggregate statistics. Mean errors and mean absolute errors are computed for profiles of potential temperature and humidity, as well as for variables related to convective weather forecasting, including many of the variables displayed on the Storm Prediction Center (SPC) Mesoscale Analysis Web site (http://www.spc.noaa.gov/exper/mesoanalysis/).

Because the variable distributions are approximately Gaussian, significance in the aggregate statistics is estimated using a standard Student’s t test. The significance of a particular hypothesis is presented here as a confidence (%), which is calculated as 100 × (1 – \(p\)), where \(p\) is the two-tailed decimal probability of the hypothesis being false. In other words, the more statistically significant the result, the lower is the \(p\) value and the higher is the confidence in rejecting the null hypothesis. In most cases in this study, the null hypothesis is that the mean error of a particular quantity is equal to zero (unbiased) so a large degree of confidence (up to 100%) indicates that the mean error is significantly different than zero or that the forecast has a statistically significant bias.

The number of independent samples used in the t test (the effective sample size) is likely less than the raw sample size because of small but nonnegligible temporal and spatial correlations among the sounding locations.

---

1 Almost all of the soundings that were removed because of suspicious low-level profiles displayed unusual behavior in the low-level moisture profile under well-mixed conditions consisting of very significant drying just above the surface followed by a gradual return to a mixing ratio closer to a well-mixed value that would be expected given the surface mixing ratio value.
The locations of uncontaminated soundings vary day to day, so standard methods that compute the spatial correlation for fixed locations over long time periods (e.g., Elmore et al. 2006) are not readily applicable here. Because the effective sample size cannot be determined precisely, it is approximated here based on the distance between the regions sampled on a given day such that multiple soundings within ~500 km of each other form only one sample. If on a given day soundings from multiple locations in the southern plains are used (e.g., from Norman, Fort Worth, and Lamont, Oklahoma), this subset of three samples is reduced to one effective sample in the significance tests. If on that same day, soundings from the southeastern United States were used (e.g., from Peachtree City, Georgia, and Tallahassee, Florida), then the effective sample size for that day is increased to two. Although this method of determining the effective sample size is subjective, using a reduced sample size is preferable to using the full sample size in the estimation of statistical significance when spatial correlations are likely present in the data but are difficult to quantify.

3. Results

a. PBL height

A simple measure of the ability of PBL schemes to depict profiles of thermodynamic variables accurately in the lower troposphere is the PBL height. Computation of turbulent quantities in the schemes requires computation of the PBL height (denoted by $h$ in the equations given in section 2). Local and nonlocal schemes compute $h$ differently, with local schemes relying primarily on TKE predictions and nonlocal schemes using empirical formulas based on wind speed, vertical gradients of
virtual potential temperature, and the critical Richardson number (see Stensrud 2007). Since we aim to compare the resultant characteristics of the profiles within the PBL, we do not use the scheme-determined $h$ here. Rather, PBL height is diagnosed here using the same method for all model and observed profiles. We shall term this diagnosed PBL height ($h_d$) to distinguish it from the PBL height computation performed within the scheme itself ($h$). The first step in computing $h_d$ is to add 0.6 K to the maximum virtual potential temperature in the lowest three model levels. Then, $h_d$ becomes the first level at which the virtual potential temperature exceeds this value. The 0.6-K correction delays a diagnosis of a small PBL depth for evening convective boundary layers until the growing stable nocturnal inversion reaches a certain depth. This simple algorithm is very similar to that used for Rapid Update Cycle (RUC) output (C. Alexander 2011, personal communication).

Mean errors of $h_d$ in our study are consistent with many past comparisons of local and nonlocal PBL schemes (Hong and Pan 1996; Bright and Mullen 2002; Stensrud and Weiss 2002; Kain et al. 2005; Hu et al. 2010); the local schemes produce shallower PBLs than the nonlocal schemes (ACM2 and YSU) during the daytime (Fig. 2). During the early morning hours, all of the forecasts of $h_d$ are very similar except for the YSU scheme, which predicts significantly higher PBL heights than the other four PBL schemes (at $>$99% confidence). This result suggests that the use of local mixing under more stable regimes in the ACM2 scheme (Pleim 2007) consistently produces more accurate nocturnal PBL depths than the YSU scheme. During the late morning the PBL deepens earliest in the two schemes that include nonlocal mixing under daytime conditions, with the ACM2 scheme separating from the local schemes by about 1500 UTC (Fig. 2). All schemes predict the peak in $h_d$ at either 2100 or 2200 UTC and rapidly return to relatively shallow depths by 0200 UTC.

In comparison to the observations, the NAM and, especially, the YSU forecasts overpredict $h_d$ at 1100 UTC (at $>$99% confidence; see Fig. 2). In the daytime, the local schemes often underpredict $h_d$ and the nonlocal schemes often overpredict $h_d$ (Fig. 2) in agreement with many past studies. However, the amount of the daytime
overprediction (underprediction) is not the same between the local and nonlocal schemes. The magnitudes of the negative mean errors for MYJ and QNSE are larger than the magnitudes of the positive mean errors for ACM2 and YSU for the evening forecasts (Fig. 3). All of the mean absolute errors are similar, but MYNN has the smallest mean error in $h_d$ among the five WRF-ARW forecasts, which is not statistically distinguishable from zero. The smallest mean absolute error in $h_d$ is found for the NAM forecasts, followed by that for MYNN.

**b. Potential temperature profiles**

The mean error profiles for the forecasts of potential temperature are shown in Figs. 4–6. The morning (11-h) forecasts of potential temperature for the WRF-ARW runs are all very similar and show (with high confidence) a cool bias in the lowest 0.5 km and a warm bias in the 1.5–3-km layer (with highest confidence near 2 km; see Fig. 5). The morning NAM forecasts are nearly identical to the WRF-ARW forecasts below 1 km, but retain a nearly unbiased profile above this level (Fig. 5).

Because of the stable boundary layer, the PBL scheme likely is not affecting conditions above 1.5 km significantly through 11 h, so the reason for the warm bias above this level for the WRF-ARW runs could be related to a warm bias that is introduced into the initial analysis (Fig. 4). (All five PBL members use the same initial conditions, so the mean error and mean absolute error profiles for the WRF-ARW 0-h forecasts shown in Fig. 4 are identical.) In the analysis, a warm bias of $\sim$0.4 K is seen for the WRF-ARW forecasts at the lowest model level that gradually increases with height to over $+1$ K above 1 km (Fig. 4). The warm bias for the NAM analysis is even larger near the ground, but the mean error decreases with height and generally has a magnitude $<0.5$ K above 1 km (Fig. 4). This comparison between the WRF-ARW analysis and the NAM analysis indicates that the CAPS 3DVAR and cloud analysis system alleviates some of the warm bias in the

![Fig. 6. As in Fig. 4, but for evening forecasts. The horizontal lines in the left panel indicate the mean forecasted PBL heights for each model. The mean observed PBL height is shown in black.](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)
NAM background near the ground but worsens the NAM biases above 1 km.

The results for the evening (23–24 h) forecasts (Fig. 6) show a tendency to undermix in MYJ and QNSE. Inspection of the MYJ and QNSE error profiles shows that there is a significant warm bias near the ground that switches to a significant cool bias starting near 0.75 km and reaching above the boundary layer. This change in sign of the bias across the depth of the boundary layer is indicative of a lack of mixing (too warm at low levels and too cool in the upper boundary layer), which is typical in a convective boundary layer in a local scheme.

The mean error profiles are different for the ACM2 and YSU schemes (Fig. 6), which show a tendency to overmix. Inspection of the ACM2 and YSU error profiles shows that there is very little temperature bias near the ground, but a significant warm bias in the upper part of the boundary layer and above the boundary layer (1–2 km). This warmth near the PBL top is indicative of too much mixing in the upper portions of the boundary layer, which may indicate that the nonlocal eddy terms are entraining too much air from the free atmosphere above.

It is noteworthy that the mean potential temperature error profiles for the MYNN scheme are nearly unbiased up to 1.5 km AGL (up to and slightly above the mean PBL height) and show a relatively small bias (<1 K) up to 4 km (Fig. 6). At the levels where there is relatively high confidence in a warm bias for MYJ and QNSE but little bias for the ACM2 and YSU schemes (below 0.5 km), the MYNN bias is nearly zero. At the levels where there is relatively high confidence in a warm bias

---

2 It should be noted that much of the warm bias below 1 km in the NAM analyses is from the 2012 cases (the changes from 2011 to 2012 in the NAM bias above 1 km are smaller). In October 2011, the operational NAM implemented numerous changes, including the use of an Arakawa staggered B grid instead of the Arakawa E grid and the change to update the first guess 2-m temperature and humidity fields in the Gridpoint Statistical Interpolation (GSI) analysis code (NOAA 2011). Because of the numerous changes made, it is difficult to determine the cause of this warm bias. The reader is simply made aware that the results here are representative of the NAM analysis in 2012 and, presumably, the current state of the NAM at the time of this writing.
for ACM2 and YSU (1.75–3 km), the MYNN warm bias is less than half as large (Fig. 6). These comparisons show a relatively good degree of performance in terms of temperature for the MYNN scheme compared to the other four PBL schemes run at 4 km in WRF-ARW. However, it is also noteworthy that the biases for the NAM forecasts are as good as MYNN, and even slightly better than MYNN in the 1.75–2.5-km layer (Fig. 6).

c. Humidity profiles

The WRF-ARW initial condition has a small positive moist bias below ~1.5 km (Fig. 7). However, as for the warm bias in the initial conditions (Fig. 4), the CAPS 3DVAR and cloud analysis system alleviates some of the moist bias that is present in the NAM analysis near the ground. The moist bias in the NAM analysis is worse than that for the WRF-ARW analyses near the ground, approaches zero more quickly with height, and then becomes negative above 1.5 km (Fig. 7).

The morning (11-h) forecasts of mixing ratio for the WRF-ARW runs are all very similar (except for YSU) and show (with high confidence) a dry bias in the lowest 0.75 km (Fig. 8). The NAM 11-h forecasts also have a dry bias that is more pronounced than the WRF-ARW forecasts up to 3 km. The different ways of handling turbulence under stable conditions among the PBL schemes discussed in section 2 appear to be evident in the relative humidity (RH) profiles (Fig. 9). YSU (a nonlocal scheme) shows a pronounced dry RH bias near the ground while QNSE (a local scheme with diffusivities determined from spectral theory to account for internal wave generation within the nocturnal inversion) is significantly different than YSU (with >99% confidence) with only a slight dry RH bias, and has the smallest bias among all the forecasts at 11 h. The largest differences in mean errors between QNSE and its close counterpart, MYJ, also is seen for RH at 11 h, although the confidence in these differences is only moderate (~70%).

The evening 23–24-h forecasts of mixing ratio (Fig. 10) are consistent with the PBL-depth results and the evening potential temperature forecasts. In the evening, the MYJ and QNSE schemes are too moist and the ACM2 and YSU schemes are too dry (Fig. 10). The MYJ

![Figure 8](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)

Fig. 8. As in Fig. 4, but for mixing ratio errors (g kg⁻¹) for 11-h forecasts.
and QNSE profiles become too dry above ~1 km and the ACM2 and YSU profiles become too moist above ~1.5 km. These mean mixing ratio error profiles are expected for too little (much) mixing in the convective PBL. Furthermore, the mean absolute errors in mixing ratio in the evening are largest for MYJ and QNSE in the 0.5–1.5-km layer (Fig. 10), reflecting errors in PBL height at the same level, although the confidence that the mean absolute errors for MYJ and QNSE are larger than those for the other models is only moderate (60%–85%).

Again, it is noteworthy that among the five PBL schemes run at 4-km grid spacing, the MYNN mean error profiles show the smallest overall biases in the lowest 3 km (Fig. 10). The MYNN profile generally falls somewhere between the profiles for the other schemes and the magnitude of the mixing ratio bias never exceeds 0.4 g kg\(^{-1}\) anywhere in the lowest 3 km. However, the mean MYNN profile, again, is not any better than the NAM forecasts statistically in this dataset (Fig. 10). Furthermore, when viewed as a mean mixing ratio over the PBL depth, MYNN is statistically very similar to the ACM2 and YSU schemes for both the mean errors and mean absolute errors in the evening (Fig. 11). However, the tendency for MYJ and QNSE to be too moist in low levels resulting from too little mixing also appears in this display with mean errors approaching +1 g kg\(^{-1}\) in the QNSE scheme.

d. Derived sounding variables

Many of the variables used for severe weather forecasting that are available at the time of this writing on the SPC Mesoscale Analysis Web site (http://www.spc.noaa.gov/exper/mesoanalysis/) did not show any significant differences in mean errors or mean absolute errors among the PBL schemes and the NAM forecasts, but there were some exceptions and interesting results to share. The morning cool and dry bias (Figs. 5 and 8) creates a significant low bias in the lowest 100-hPa mixed layer (ML) convective available potential energy (CAPE) that is significantly different from zero at high confidence for all forecasts (Fig. 12a), with mean errors \(-400\) J kg\(^{-1}\) [CAPE and convective inhibition (CIN) are only computed if all the models and the observation

![Image](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)
have positive CAPE]. Disappointingly, all the models (including the NAM) underforecast MLCAPE in the morning more than 75% of the time, and the morning MLCAPE is too low by 700–800 J kg\(^{-1}\) 25% of the time among all the forecasts. An example of a typical situation where the WRF-ARW forecasts all underpredict the morning MLCAPE is shown in Fig. 13. The mixed layer is defined to be the lowest 100 hPa, which is typically \(\sim 1 \text{ km AGL}\), so these estimates of MLCAPE include air above the nocturnal inversion in the computation of the mean parcel quantities. However, surface-based CAPE and CAPE for shallower mixed layer parcels are also underpredicted to a similar magnitude (not shown).

The above result of an underprediction of MLCAPE in the morning may be a significant factor in the difficulty that CAMs have in forecasting the intensity of convection overnight. Stratman et al. (2013) show that 0–12-h forecasts of composite reflectivity \(\geq 35 \text{ dBZ}\) from a 0000 UTC initialized MYJ member of the CAPS ensemble showed very little skill beyond a few hours even when enlarging the neighborhoods for the verification. Adding to this hypothesis that poor forecasts of instability overnight lead to poor forecasts of the intensity of convection overnight is a tendency for the models to over forecast MLCIN,\(^3\) meaning the models tend to produce too much convective inhibition in the early morning hours (Fig. 12b). The mean absolute errors of MLCIN are \(\sim 100 \text{ J kg}^{-1}\) for all the models. Absolute errors of MLCAPE and MLCIN of 800 and 100 J kg\(^{-1}\), respectively, can make a significant difference in the character of forecasted convection (Rasmussen and Blanchard 1998; Davies 2004; Ziegler et al. 2010; Nowotarski et al. 2011), or even whether or not convection develops (Fabry 2006; Wilson and Roberts 2006), particularly when the observed values of MLCAPE and MLCIN are approximately the same as the errors. This could have significant implications for the ability of CAMs to forecast severe weather at night.

In the evening (23–24-h forecasts), the predictions of MLCAPE and MLCIN improve overall with only \(\text{FIG. 10. As in Fig. 6, but for mixing ratio errors (g kg}^{-1}\) for evening forecasts.\)

\(^3\) An overprediction of CIN means that the absolute magnitude of CIN was larger in the model forecasts than in the observations.
ACM2 and YSU showing a tendency to underpredict MLCAPE (confidence in the 70%–95% range) (Fig. 14a). The mean MLCAPE error for MYNN also is negative but is not significantly indistinguishable from zero. Mean MLCAPE errors for MYJ and QNSE are very close to zero showing that their tendency to produce PBLs that are too shallow, and subsequently with too much moisture in low levels, does not translate into an overall MLCAPE bias. This is because the calculation of CAPE using a parcel with mean conditions over the lowest 100 hPa, which is typically up to ~1 km AGL, effectively averages out the mean conditions in the PBL for the MYJ and QNSE schemes (however, this is not the case when the observed MLCAPE is large, which is shown later). This explanation is supported by surface-based CAPE values that show a slight overforecast for the MYJ and QNSE schemes and somewhat of an underforecast for the ACM2 and YSU scheme (not shown). Furthermore, the MLCAPE errors for MYJ and QNSE have more variance compared to the other schemes (see the vertical extent of the boxes in Fig. 14a) resulting in mean absolute errors that are statistically very similar among all the forecasts with values ~600 J kg\(^{-1}\).

The mean absolute errors in MLCIN in the evening are again ~100 J kg\(^{-1}\) for all forecasts, but all of the MLCIN forecasts are statistically unbiased in the evening when averaged over all soundings with positive CAPE (Fig. 14b). However, a difference in the datasets is seen when stratifying based on the existence of an inversion atop the PBL. The observed soundings in the set of 75 soundings valid at 2300 UTC for which MLCAPE is nonzero in all the forecasts and in the observation were each inspected manually for the existence of an inversion atop the convective boundary layer and then were separated into two groups: one with a substantial capping inversion and one with either a weak inversion or no inversion at all. The maximum temperature lapse rate between a level 500 m below the PBL height (\(h_d\)) (or

![Fig. 12](image-url). As in Fig. 3, but for MLCAPE errors for 11-h forecasts (valid at 1100 UTC). Only those soundings that have positive CAPE in both the model forecasts and the observation are included here.
2 m, whichever came first) and a level 2500 m above $h_d$ (or 5000 m, whichever came first) was computed. All of the soundings in the “substantial capping inversion” group have a positive maximum lapse rate (positive meaning temperature increases with height) and most of the soundings in the “weak or no capping inversion” group have either negative or small maximum lapse rates. For the few soundings that were placed in the substantial capping inversion group that have relatively small maximum lapse rates, the layer that contains the positive lapse rates is deep (>25 hPa). Likewise, the few soundings that were placed in the weak capping
inversion group that have a positive (but small) maximum lapse rate have shallow inversion layers (<25 hPa).

When a substantial capping inversion exists in the observations, the corresponding capping inversion in the model forecasts tends to be damped considerably (examples are shown in Fig. 15). The MLCIN was calculated for the subset of 44 locations that showed a substantial capping inversion in the observed sounding but a smoothed representation of the inversion in the model soundings. In every model, the mean MLCIN errors for these 44 soundings was positive, indicating too little convective inhibition tends to be forecast in these situations (although the reduced sample size reduces the confidence in this result; see Fig. 16a). In every model, the MLCIN error distribution for the 31 soundings that showed a weak or no capping inversion is shifted downward (toward more negative errors; see Fig. 16b). This result indicates that the MLCIN tends to be underforecast more often when a substantial capping inversion is present in the observations.

Finally, an interesting difference emerges when stratifying by observed MLCAPE. The 23-h forecast errors in sounding parameters are computed for two sets of soundings; those with observed MLCAPE ≤1500 J kg⁻¹ and those with observed MLCAPE >1500 J kg⁻¹ (Fig. 17). The MYJ and QNSE slightly overestimate the MLCAPE when the observed value is low (Fig. 17a), but the more significant result is that the MYNN, ACM2, and YSU schemes all underestimate the MLCAPE when the observed MLCAPE value is large (Fig. 17b). This results from the mixing ratio being too low over the lowest 100 hPa in the MYNN, ACM2, and YSU schemes when the MLCAPE is large (Fig. 18). Inspection of the soundings reveals that when the environment has a relatively high mixing ratio near the ground and is relatively dry in the free atmosphere above the PBL—an environment often associated with large CAPE—the enhanced mixing–entrainment in the MYNN, ACM2, and YSU schemes are bringing too much dry air from the free atmosphere into the PBL and, subsequently, bring too much moisture aloft into the free atmosphere, as can be seen in Fig. 18. However, this excess moisture is typically above the level that is likely to be used to calculate the CAPE for a surface-based mixed layer parcel (the lowest 100 hPa in this study), so the 100-hPa MLCAPE is underestimated in these situations.

4. Summary and concluding remarks

This study examines forecasts of thermodynamic state variables from five WRF-ARW forecasts over two spring seasons (2011 and 2012) in regions favorable for deep convection. This study is part of a larger effort of CAPS and the HWT to evaluate the performance of convection-allowing models for severe weather forecasting applications. The purpose of this study is to gain a better understanding of the characteristics of the newer PBL scheme options (ACM2, MYNN, and QNSE) in convection-allowing configurations of the WRF-ARW model as they pertain to convective weather forecasting.

![Figure 14](https://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)

**Fig. 14.** As in Fig. 3, but for (a) MLCAPE and (b) MLCIN errors (J kg⁻¹). Only those soundings that have positive CAPE in both the model forecasts and the observation are included here.
and compare them to the well-known characteristics of the widely used MYJ and YSU schemes.

A total of 191 observed soundings are used to evaluate forecasts valid in the morning (11-h forecasts) and 100 soundings are available to evaluate 23–24-h forecasts valid in the evening (usually 23-h forecasts compared to radiosondes released shortly after 2300 UTC). In the morning, all PBL schemes are too cool and dry near the
ground despite having little bias in PBL depth. The exception is for YSU, which produces profiles of temperature and moisture significantly different than the other PBL schemes in the morning. Because turbulent mixing is less often a primary physical mechanism in stable boundary layers, and because the cool and dry bias is seen for all PBL members, it is likely that something other than the turbulent mixing parameterization—perhaps the LSM or initial analysis—is responsible for this bias in the morning. Further work should detail these sensitivities in the LSM or initial conditions in convection-allowing configurations of the WRF-ARW model in regions favorable for deep convection.

The cool and dry biases in the morning lead to a significant underprediction (overprediction) of MLCAPE (MLCIN) at that time in all schemes, with disturbingly

![Fig. 16](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)

**Fig. 16.** As in Fig. 3, but for MLCIN for (a) the 44 observed soundings with a substantial capping inversion and (b) the 31 observed soundings with a weak to no capping inversion. Only those soundings that have positive CAPE in both the model forecasts and the observations are included here.

![Fig. 17](http://journals.ametsoc.org/waf/article-pdf/28/3/842/4653589/waf-d-12-00103_1.pdf)

**Fig. 17.** As in Fig. 3, but for MLCAPE errors for (a) the 40 observed soundings with MLCAPE $\leq 1500$ J kg$^{-1}$ and (b) the 35 observed soundings with MLCAPE $>1500$ J kg$^{-1}$. Only those soundings that have positive CAPE in both the model forecasts and the observations are included here.
FIG. 18. As in Fig. 4, but for mixing ratio errors (g kg$^{-1}$) for (a) the 45 observed soundings with MLCAPE $\leq 1500$ J kg$^{-1}$ and (b) the 35 observed soundings with MLCAPE $> 1500$ J kg$^{-1}$. For reference, the lowest 100-hPa layer typically reaches $\sim 1$ km AGL.
large mean errors in MLCAPE of \(\sim (500-600) \text{ J kg}^{-1}\) and errors of \(\sim (700-800) \text{ J kg}^{-1}\) occurring nearly a quarter of the time. This is disturbing because the magnitude of the MLCAPE often is not much larger than these values overnight, and significant severe weather episodes often occur overnight in these values of MLCAPE (Dean and Schneider 2008). This could be a significant factor in the difficulty of convection-allowing models to forecast the intensity of convection accurately at night.

In the evening, the local schemes produce shallower PBLs that are often too shallow and too moist compared to nonlocal schemes, a familiar result. However, MYNN (a local scheme) is nearly unbiased in PBL depth, moisture, and potential temperature, which is comparable to the background NAM forecasts. This result increases confidence that the MYNN scheme is an improvement over the MYJ or YSU scheme in deterministic convection-allowing model forecasts of temperature and moisture near the ground in relatively warm, moist conditions.

Likewise, MLCAPE and MLCIN forecasts improve in the evening, with MYJ, QNSE, and MYNN having small mean errors, but ACM2 and YSU have a somewhat low bias. Evening MLCIN tends to be underpredicted in areas with strong capping inversions, as the model inversions are too smooth. Evening MLCAPE tends to be overpredicted by MYJ and QNSE when the observed MLCAPE is relatively small \((\lesssim 1500 \text{ J kg}^{-1})\) and tends to be underpredicted by MYNN, ACM2, and YSU when the observed MLCAPE is large \((>1500 \text{ J kg}^{-1})\). This suggests that CAPE calculations could be improved when the depth over which mean parcel quantities are computed is varied according to the PBL depth.

Forecasts from the operational NAM are presented along with those from the five WRF-ARW forecasts to provide a general comparison of forecasts between convection-allowing and non-convection-allowing models. Many differences in the configurations exist between the NAM and the WRF-ARW forecast models presented here, including grid spacing, vertical levels \((60 \text{ in the NAM versus } 51 \text{ in the WRF-ARW simulations used here})\), effective resolution, and model numerics. These differences prevent any direct diagnoses of the reasons for the differences in the NAM forecasts and the WRF-ARW forecasts. However, the NAM forecasts are presented to provide a baseline to a widely used operational model. The NAM forecasts were quite comparable to the best-performing schemes for all variables examined. In fact, the NAM forecasts often had significantly lower biases compared to the 4-km forecasts. This result suggests that the combination of physics parameterization schemes used in convection-allowing configurations of the WRF-ARW produces (excluding MYNN) PBL profiles that tend to degrade the background NAM profiles.

A significant result from this study is the relatively good performance of MYNN in terms of biases in temperature and moisture in the lower troposphere compared to the other PBL schemes in WRF-ARW. This is noteworthy because the MYNN appears to perform well in areas where only nonlocal schemes have performed well in the past. This suggests that a commonly accepted criticism of local schemes, that they are underdispersive and cannot build a convective PBL properly, may not be the case with the improvements made in MYNN. This seems to suggest that the local-scheme design is not flawed, but rather that it requires a more appropriate set of closure constants and a mixing length formulation appropriate for the spectrum of stabilities.

Finally, while it is certainly important to accurately predict PBL structures, it is not the full story for convective forecasting applications. The reader is reminded that the results here are valid only for environments that are near deep convection in space and time \((\sim 400 \text{ km} \text{ and } 6 \text{ h})\), but are convectively uncontaminated. Furthermore, although many locations east of the Rocky Mountains are included in the dataset, the results are most applicable to central U.S. convective environments (see Fig. 1) in mid- to late spring. The fact that MYNN performed the best for mean thermodynamic variables in moist low-level conditions upstream from convection does not necessarily mean it will perform the best when evaluated for explicit forecasts of convection or for other characteristics of the simulations that depend on turbulent mixing (e.g., the position of drylines and fronts). Efforts to better understand the performance of these schemes in predicting convective storms and precipitation is ongoing.

Acknowledgments. We are very appreciative of the hard work and dedication of CAPS scientists, particularly Kevin Thomas, and the staff at the National Institute of Computational Science at the University of Tennessee who made the 2011 and 2012 CAPS ensemble possible. The authors thank Joseph Olson from NOAA/GSD for his help in understanding the WRF-ARW implementation of PBL schemes, and Adam Clark of NSSL for helpful suggestions. We also thank Dr. Mike Douglas of NSSL, Dr. Don Conlee and the sounding team from Texas A&M, Scott Giangrande for help with the MC3E sounding data, and Kimberly Elmore for helpful discussions on significance testing. Funding was provided by NOAA/Organization for Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The CAPS ensemble forecasts were primarily supported by the NOAA CSTAR program, and were produced on Kraken at the National Institute of
Computational Science at the University of Tennessee, with some postprocessing done at OSCER at the University of Oklahoma. Supplementary support was provided by NSF-ITR Project LEAD (ATM-0331594), NSF ATM-0802888, and other NSF grants to CAPS.

REFERENCES


——, 2001: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso Model. NCEP Office Note 437, 61 pp.


