Surface Turbulent Fluxes during Persistent Cold-Air Pool Events in the Salt Lake Valley, Utah. Part II: Simulations

XIA SUN AND HEATHER A. HOLMES

Atmospheric Sciences Program, Department of Physics, University of Nevada, Reno, Nevada

HUI XIAO

Institute of Tropical and Marine Meteorology, China Meteorological Administration, Guangzhou, China

(Manuscript received 17 October 2019, in final form 21 April 2020)

ABSTRACT

Realistically representing the land–atmosphere interactions during persistent cold-air pools (PCAPs) is critical in simulating the strength of PCAPs, where uncertainties in simulating the PCAP strength will impact the ability to model the poor air quality. To quantify the model performance for land–atmosphere exchange, measurements of surface turbulent and radiative energy fluxes during two PCAPs, one weak and one strong, in Utah were compared with simulations from the Weather Research and Forecasting (WRF) Model. The results show that the WRF Model simulated the surface energy fluxes well in the weak PCAP case and that the performance degraded in the strong PCAP case. The significantly overestimated surface sensible heat flux $H$ and latent heat flux (LE) in the strong PCAP were related, in part, to the overestimated net radiation and soil moisture and unsuitable turbulence parameterizations. The simulation using the Mellor–Yamada–Nakanishi–Niino planetary boundary layer scheme produced the least bias in both net radiation and surface turbulent fluxes for the strong PCAP case, which is expected because of the local higher-order (2.5) turbulence closure scheme. The surface exchange coefficient ($CH$), a crucial variable used to calculate $H$, was overall overestimated by the WRF Model. The underestimation of the nondimensional vertical temperature gradient in the Monin–Obukhov stability function was responsible for the overestimated $CH$, where the stability functions deviate significantly from expected values from observations for the stable atmospheric boundary layer. Our study highlights the need to improve the flux–profile parameterizations under stable conditions over complex terrain by including impacts due to mountainous terrain, such as surface radiative flux divergence and the diurnal mountain wind system.

1. Introduction

Poor air quality is of great concern because of its impact on human health and ecosystems worldwide. Persistent cold-air pools (PCAPs), characterized by topographically confined cold air, are usually accompanied by high levels of surface air pollution concentrations (Silcox et al. 2012) due to suppressed vertical mixing and shallow boundary layer depths lasting for more than one day. PCAPs have been observed in many mountain areas, such as the Salt Lake Valley (SLV) in Utah (Lareau et al. 2013), the Columbia basin in Washington (Whiteman et al. 2001), the Arve River valley in the French Alps (Chemel et al. 2016), and the Grünloch basin in the eastern Alps (Dorninger et al. 2011). Air quality models are useful tools to predict pollutant transport and air pollution concentrations. Simulation of the elevated air pollution concentrations during PCAP events using currently available atmospheric models is still challenging (Baker et al. 2011), where meteorological conditions, emissions, and atmospheric chemistry all impact the air quality modeling uncertainties (Gilliam et al. 2015).

Accurate meteorological simulations cannot be achieved without correct representation of the land–atmosphere interactions, which are the major driver in the surface hydrological and energy cycles. Because of this fact, the land–atmosphere interactions impact the near-surface meteorological conditions and modulate clouds and precipitation (Santanello et al. 2018). Simulations using the Weather Research and Forecasting (WRF) Model with the Noah land surface model (LSM) focusing on the southern Great Plains by Qian et al. (2013) found that the surface sensible heat flux decrease or latent heat
flux increase lead to cooling of the surface, an increase in surface specific humidity, decrease of the lifting condensation level, and decrease of the mixing-layer height. While these results are not specific to stable conditions, they illustrate the importance of the land–atmosphere coupling in numerical weather prediction. In addition to sensible and latent heat fluxes, surface radiative fluxes are major terms in the surface energy budget. Cloud cover impacts the surface radiative fluxes and, therefore, the surface energy budget, and simulating the correct cloud cover is important to model cold-air pool (CAP) formation and evolution (Hughes et al. 2015). Biases in simulated radiative fluxes can lead to modeling deficiencies in surface heating and surface melting when snow is present (King et al. 2015). Modeling uncertainties in the simulation of surface turbulent and radiative fluxes still exist when compared with observations (Massey et al. 2017; Raja et al. 2016; Ruiz-Arias et al. 2016), which highlights the requirement of ongoing efforts to identify deficiencies of numerical models in representing the land–atmosphere exchange.

Surface turbulent heat fluxes are parameterized in LSMs using Monin–Obukhov (M–O) similarity theory (Monin and Obukhov 1954). The theory is based on assumptions that the atmospheric surface layer is steady state and homogenous, which can be problematic under stable conditions and over complex terrain (Babić et al. 2016; Sfyri et al. 2018; Stiperski and Rotach 2016). Several stability functions for momentum and heat in the surface layer have been proposed in the literature (Byun 1991; Chenge and Brutsaert 2005; Dyer 1974; Holtslag and Boville 1993; Lettau 1979), but they still remain an issue for models that rely on them for calculations of surface turbulent fluxes. For example, Vignon et al. (2017) evaluated four typical stability functions and found the nondimensional temperature gradient was underestimated under stable conditions based on observations collected in the East Antarctic Plateau. This highlights the need for more empirical datasets to evaluate the existing formulations of the M–O similarity functions to model the land–atmosphere exchange, particularly for cases where the steady state and homogeneous assumptions are no longer valid.

In addition to similarity functions, the surface energy partitioning depends on land-use types with different physical properties, which suggests the importance of realistic and dynamic land-use datasets in numerical models. Ran et al. (2016) demonstrated that the meteorological fields and air quality simulated by WRF/Community Multiscale Air Quality model (CMAQ) were improved with updated land-use information, in terms of leaf area index, surface albedo, and vegetation fraction data. Another parameter that directly benefits from realistic land-use information is the surface exchange coefficient, which governs the calculation of surface turbulent fluxes using the bulk transfer method (Högström 1988). The surface exchange coefficient is formulated using roughness lengths, which depends on canopy height. Chen and Zhang (2009) found that the Noah LSM overestimated the surface exchange coefficient for short vegetation and that modifying the model with realistic canopy heights reproduced the observed land–atmosphere coupling strength due to improved thermal roughness lengths.

To improve weather forecasting and eventually air quality prediction during PCAPs over complex terrain, it is necessary to investigate how well the surface energy fluxes are represented by numerical models, especially when surface fluxes of chemical species have similar exchange pathways as the moisture fluxes (Pleim and Ran 2011). The objective of this study is to examine land–atmosphere interactions simulated by the WRF Model during PCAPs to identify the model strengths and deficiencies. The performance of the WRF Model using different LSMs and PBL schemes, in simulating surface turbulent fluxes during PCAPs, as well as surface radiation components, are assessed using observations from the Persistent Cold-Air Pool Study (PCAPS) in the SLV (Sun and Holmes 2019). Section 2 presents the observational datasets, the WRF Model configurations, and a brief overview of the M–O similarity theory used to model the land–atmosphere exchange in WRF. Section 3 shows the results, a discussion focusing on WRF Model uncertainties in simulating turbulent fluxes over complex terrain during wintertime is given in section 4, and the conclusions are presented in section 5. This study is novel in highlighting the importance of the surface exchange coefficient in simulating the surface turbulent fluxes during stable boundary layer events based on data collected during PCAPs.

2. Data and methods

a. PCAPS observational datasets

The PCAPS field campaign took place in the Salt Lake Valley from December 2010 to February 2011 (Lareau et al. 2013). Our focus is on the land–atmosphere exchange process during PCAPs. Note that in this paper, PCAPS will be used to designate the field campaign, and PCAP/PCAPS (with a lowercase s) is used to denote the atmospheric phenomenon. Thorough analysis of this PCAPS observational dataset is given in the first part of this research series (Sun and Holmes 2019). Surface turbulent fluxes and four-component radiation fluxes in 30-min intervals were measured at seven sites with different land-use types based on the 2011 National
Table 1. Surface energy balance observation sites and their modeled land-use type and soil type information.

<table>
<thead>
<tr>
<th>Code</th>
<th>Lat</th>
<th>Lon</th>
<th>Grid land-use type</th>
<th>Grid soil type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>40.779°N</td>
<td>112.103°W</td>
<td>Barren land</td>
<td>Sandy clay loam</td>
</tr>
<tr>
<td>DH</td>
<td>40.738°N</td>
<td>111.913°W</td>
<td>Developed high intensity</td>
<td>Silt</td>
</tr>
<tr>
<td>DM</td>
<td>40.697°N</td>
<td>111.839°W</td>
<td>Developed medium intensity</td>
<td>Silt</td>
</tr>
<tr>
<td>DL1</td>
<td>40.681°N</td>
<td>111.995°W</td>
<td>Developed low intensity</td>
<td>Sandy clay loam</td>
</tr>
<tr>
<td>DL2</td>
<td>40.614°N</td>
<td>111.804°W</td>
<td>Developed low intensity</td>
<td>Sandy loam</td>
</tr>
<tr>
<td>PH</td>
<td>40.594°N</td>
<td>112.062°W</td>
<td>Pasture/hay</td>
<td>Silt</td>
</tr>
<tr>
<td>CR</td>
<td>40.514°N</td>
<td>111.974°W</td>
<td>Cultivated crops</td>
<td>Silt</td>
</tr>
</tbody>
</table>

Land Cover Dataset (NLCD), including barren land (BL), developed high intensity (DH), developed medium intensity (DM), developed low intensity (DL), pasture/hay (PH), and cultivated crops (CR). The observation sites are coded as abbreviations based on their modeled land-use types in WRF, based on the NLCD (Table 1). The NLCD-modeled land-use types were consistent with observations based on photographs taken during the experiment and information recorded in log notebooks (UCAR/NCAR Earth Observing Laboratory 1990). Measurements made at the seven surface energy balance sites also included near-surface meteorological parameters, that is, 2-m temperature (T), 10-m wind speed (WS), 10-m wind direction (WDIR), and 2-m specific humidity (r). In addition, soil temperatures were measured at four levels (0.6, 1.9, 3.1, and 4.4 cm) and soil moisture was measured at 5 cm underground. Surface turbulence measurements were conducted, quality controlled, and postprocessed by the Earth Observing Laboratory (EOL) at the National Center for Atmospheric Research (NCAR) using their protocol for eddy covariance data processing with a 30-min Reynolds averaging time. Descriptions of the detailed sensor information, the correction methods, and the quality control procedures are given in Sun and Holmes (2019).

There were 10 PCAP episodes during the field study, corresponding to 10 intensive observation periods (IOPs), determined by using a forecast potential temperature threshold of 8 K difference from valley floor to ridge height during the observation period (Lareau et al. 2013). The average ridge height for the Oquirrh Mountains on the west side of the SLV is around 2200 MSL. Analysis after the experiment confirmed the PCAP strengths using the valley heat deficit method (Sun and Holmes 2019). We focus on two IOPs, IOP3 (lasting from 0500 MST 12 December 2010 to 1400 MST 14 December 2010) and IOP5 (lasting from 1700 MST 31 December 2010 to 0500 MST 9 January 2011), to evaluate the WRF Model performance on the simulated surface energy balance for PCAP events with varied duration and strength. IOP3 (IOP5) featured weak (strong) PCAP due to its weak (strong) atmospheric stratification and short (long) duration. Formation of the brief PCAP during IOP3 was associated with subsidence accompanied by a surface high pressure system in the SLV on 12 December. The IOP5 PCAP broke down on 15 December with cold-air advection aloft bringing into the valley by a trough (Lu and Zhong 2014). During IOP5, the dominant anticyclone system covered most areas in the western United States from 1 to 8 January. Snow cover, which was deposited in SLV on 30–31 December as a result of a cold-frontal passage, prolonged the IOP5 PCAP with dampened surface turbulent mixing together with increased surface albedo (Foster et al. 2017). The cold-frontal passage on 9 January brought in cold air aloft and ended the strongest PCAP during the field campaign.

b. WRF Model configurations and simulations

The WRF Model (version 3.7.1) configuration consists of three two-way nested domains with horizontal resolutions of 12 (d01), 2.4 (d02), and 0.48 km (d03) (Fig. 1a). The most up-to-date NLCD 2011 was used to provide the land-use information for WRF. Table 1 lists the modeled land-use types and soil types of the seven surface energy balance observation sites during PCAPS. We used 30 sigma levels in the vertical domain, extending from near surface to 50 hPa. There were 10 sigma levels below 1 km with the first half sigma level at around 19 m above ground level. While increased vertical resolution might impact the model performance in simulating cold-air pools (Hughes et al. 2015; Vosper et al. 2013), sensitivity tests of model performance associated with horizontal and vertical grid resolutions are beyond the scope of this study. Meteorological initial and boundary conditions were taken from the National Centers for Environmental Prediction (NCEP) North American Mesoscale Forecast System analysis (NAM-ANL at 0000, 0600, 1200, and 1800 UTC) dataset and 3-h NAM forecast (NAM-NMM at 0300, 0900, 1500, and 2100 UTC) dataset with 12 km horizontal resolution. The NAM reanalysis dataset was found to be able to synoptically capture PCAP events by Lu and Zhong (2014).
To examine the WRF Model performance on the simulation of the surface energy balance terms during PCAPs, four simulations—NAM-asymmetrical convective model, version 2 (NAM_ACM2); NAM_YSU; NAM_MYJ; and NAM_MYNN—using different PBL schemes together with their recommended LSMs and surface-layer schemes (Table 2), were conducted per IOP (IOP3 and IOP5). The physics configurations utilized in all simulations included Thompson microphysics (Thompson et al. 2008), Rapid Radiative Transfer Model Global (RRTMG) longwave radiation (Iacono et al. 2008), Dudhia shortwave radiation (Dudhia 1989), and Kain–Fritsch cumulus parameterization (Kain 2004).

Four-dimensional data assimilation (FDDA) has been demonstrated to improve meteorological simulations (Gilliam et al. 2012; Stauffer and Seaman 1990; Sun et al. 2017) and was employed in our coarsest domain (d01) in all simulations. The NCEP Automated Data Processing (ADP) Global Surface and Upper Air Observational Weather Data were blended with NAM data using the “OBSGRID” tool (NCAR 2009) for FDDA. In addition, the Pleim–Xiu LSM uses the 2-m temperature and humidity analysis fields generated by OBSGRID in its own indirect soil nudging scheme (Pleim and Xiu 2003). Following Gilliam and Pleim (2010), the nudging of wind was applied in all layers, while the temperature and moisture nudging was not applied within the model-diagnosed PBL. The nudging strength for wind, temperature, and moisture was set to $3 \times 10^{-4}$, $3 \times 10^{-4}$, and $1 \times 10^{-5}$ s$^{-1}$, respectively (Otte 2008).

c. Monin–Obukhov similarity theory

The surface-layer schemes, which provide surface exchange coefficients for the calculation of surface turbulent fluxes in LSMs (Skamarock et al. 2008), are based on M–O similarity theory (Monin and Obukhov 1954). The flux–profile functions for the gradients of mean wind speed $U$ (m s$^{-1}$) and potential temperature $\theta$ (K), $\phi_m$ and $\phi_h$, respectively, in the surface layer are given as (Stull 2012)

\[
\phi_m = \frac{k_z \partial U}{u* \partial z} \quad \text{and} \quad \phi_h = \frac{k_z \partial \theta}{\theta_* \partial z},
\]

where the dimensionless length scale $z/L$ represents the M–O stability parameter $\zeta$, defined as

<table>
<thead>
<tr>
<th>Expt</th>
<th>Surface-layer scheme</th>
<th>Land surface model</th>
<th>Planetary boundary layer scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAM_YSU</td>
<td>Revised MM5 (Jiménez et al. 2012)</td>
<td>Noah (Ek et al. 2003)</td>
<td>YSU (Hong et al. 2006)</td>
</tr>
<tr>
<td>NAM_MYJ</td>
<td>Eta similarity (Janjic 2001)</td>
<td>Noah</td>
<td>MYJ (Janjic 2001)</td>
</tr>
<tr>
<td>NAM_MYNN</td>
<td>MYNN (Nakanishi 2001)</td>
<td>Noah</td>
<td>MYNN (Nakanishi and Niino 2004)</td>
</tr>
</tbody>
</table>

FIG. 1. (a) The WRF Model domains and (b) orography of the innermost domain with observation sites during PCAPS. The orography map was plotted using the ETOPO1 data from the National Geophysical Data Center.
\[
\zeta = \frac{z}{L} = k \frac{g}{\theta} \frac{\theta_0}{u^*},
\]

where \(k\) is the von Kármán constant (0.40), \(g\) is the gravitational acceleration (m s\(^{-1}\)), \(z\) is the height above the surface (m), \(\theta_0\) is the surface-layer turbulent temperature scale (K), and \(u^*\) is friction velocity (m s\(^{-1}\)). The observed \(\phi_m\) and \(\phi_h\) can be estimated from in situ measurements using Eqs. (1) and (2). The vertical gradient of wind speed was calculated based on the 10-m wind speed and surface wind speed (assumed to be zero). Potential temperature gradient was based on the surface skin temperature and 2-m air temperature. Rearranging Eqs. (1) and (2) and including terms to account for roughness height and atmospheric stability, one can obtain equations for the vertical profiles of wind speed and temperature:

\[
U(z) = \frac{u^*}{k} \left[ \ln \left( \frac{z}{z_{0m}} \right) - \psi_m \left( \frac{z}{L} \right) \right] \quad \text{and} \quad (4)
\]

\[
\theta(z) - \theta_0 = \frac{\theta_0}{k} \left[ \ln \left( \frac{z}{z_{0h}} \right) - \psi_h \left( \frac{z}{L} \right) \right], \quad \text{and} \quad (5)
\]

where \(z_{0m}\) and \(z_{0h}\) are aerodynamic roughness height and thermal roughness height, respectively and \(\theta_0\) is the potential temperature at the thermal roughness height. Functions \(\psi_m\) and \(\psi_h\) are integrated similarity functions, given by

\[
\psi_m \left( \frac{z}{L} \right) = \int_{z_{0m}/L}^{z/L} \frac{1 - \phi_m(\zeta)}{\zeta} d\zeta \quad \text{and} \quad (6)
\]

\[
\psi_h \left( \frac{z}{L} \right) = \int_{z_{0h}/L}^{z/L} \frac{1 - \phi_h(\zeta)}{\zeta} d\zeta, \quad \text{and} \quad (7)
\]

where these similarity functions differ based on the atmospheric stability \((z/L)\) and relate the surface turbulent fluxes to mean profile gradients, specifically for momentum and heat [Eqs. (6) and (7), respectively]. In numerical weather prediction models the formulation of the mean profile gradients \((\phi_m\) and \(\phi_h\)) differs based on the selected surface-layer scheme. However, they are all based on M–O similarity theory and rely on empirical data to formulate the flux–profile relationship.

The flux–profile relationship for heat is used in the surface-layer model to determine the bulk transfer coefficient that the land surface model uses to estimate the surface sensible heat flux. The surface sensible heat flux \(H\) is given as

\[
H = -\rho c_p u^* \theta_0 = -\rho c_p CH(\theta_0 - \theta_0), \quad \text{and} \quad (8)
\]

where \(\rho\) is the air density (kg m\(^{-3}\)); \(c_p\) is the specific heat capacity (J kg\(^{-1}\) K\(^{-1}\)); \(\theta_0\) and \(\theta_0\) are the potential temperature in the air and at the ground (K), respectively; and \(CH\) is the surface exchange coefficient for heat (m s\(^{-1}\)) and has been multiplied by wind speed, which is parameterized in the WRF Model based on M–O similarity theory as

\[
CH = \frac{k u^* \ln \left( \frac{z}{z_{0h}} \right) - \psi_h \left( \frac{z}{L} \right)}{L}. \quad \text{and} \quad (9)
\]

The CH can be estimated from observations based on Eq. (8) using \(H\) measured by a sonic anemometer and the potential temperature gradient between the air and ground [see the methods section in Sun and Holmes (2019)]. The ground temperature (sometimes referred to as a skin temperature) is calculated from the measured outgoing longwave radiation, which is proportional to the fourth power of ground temperature, and the 2-m temperature is used as approximation of the air temperature. This enables a comparison between the WRF simulated bulk transfer coefficient and the bulk transfer coefficient calculated based on observations.

3. Results

To assess the WRF Model performance for PCAP events, we compared the model outputs with observations for both the surface meteorological variables and surface energy fluxes. Considering the observational data coverage in both IOPs (IOP3 and IOP5), a case study analysis was conducted based on data collected at site DL2 (i.e., site with the most complete data record for both IOPs) to investigate the temporal evolution of the WRF simulations and compare the strong and weak PCAPs (Table 3 for weak PCAP and Table 4 for strong PCAP). The model and observation comparisons for site DL2 are presented below in section 3a for the surface meteorological conditions, section 3b for radiation, and section 3c for surface turbulent fluxes. While the data coverage at all of the monitoring sites would not allow for stratification by each IOP to investigate the atmospheric stability impacts, a spatial analysis was done to compare the results by land-use type. The spatial statistical model evaluations, including mean bias (MB), root-mean-square error (RMSE), and mean absolute error (MAE), are presented in section 3d to illustrate the dependence of model performance on simulating the surface energy balance terms for different land-use types using data from all sites. Readers are referred to Sun et al. (2017) for detailed equations used to calculate these evaluation statistics.
a. Surface meteorological conditions

1) WEAK PCAP CASE STUDY (IOP3)

The observed and simulated $T$, WS, WDIR, and $r$ at site DL2 during IOP3 are presented in Fig. 2. The observed average $T$ during IOP3 was 281.14 K. The WRF simulations captured the diurnal variations of $T$ with an underestimation of $-2.39$ K on average. The cold bias of $T$ occurred both during daytime and nighttime, with an average bias of $-2.22$ and $-2.48$ K, respectively. Two-meter temperature $T$ is diagnosed using simulated surface skin temperature $T_s$ and $H$ in the WRF Model (Grossman-Clarke et al. 2010). Therefore, the underestimation of $T$ in WRF is related to an underestimation of $T_s$ and/or overestimation (underestimation) of $H$ transferred from the surface (atmosphere) to the atmosphere (surface). The model spread in $T$ simulation was large during nighttime, indicating a large simulation difference of the nighttime $T_s$ and/or $H$ among the four sensitivity experiments. The WRF scenario of NAM_MYJ compared best with observations, with an MAE of 2.00 K and RMSE of 2.57. The WRF scenario of NAM_MYNN compared poorest with observations, with an MAE of 3.11 K and RMSE of 3.73 K. This shows that although $T$ is parameterized in the LSMs, different PBL schemes using the same LSM (i.e., NAM_MYJ and NAM_MYNN both use the Noah LSM) can still generate different temperatures due to feedbacks of the land–atmosphere interactions to the PBL.

The near-surface wind speed is also a crucial parameter in numerical models, especially during PCAPs, when air pollutants tend to accumulate due to decreased wind speeds and therefore dampened mixing. The observed average WS during IOP3 was 1.65 m s$^{-1}$, with a maximum value of 4.92 m s$^{-1}$ on 14 December, when the CAP began to break up. The model did not capture some sporadically elevated WS, for example, 1800 MST 12 December and 0000 MST 14 December, which are suspected to be associated with lake breezes based on WDIR in Fig. 2c. The simulation ensemble successfully reproduced the increased WS near the end of IOP3. This implies that the WRF Model is capable of modeling the life cycle of weak PCAPs in the aspect of air pollution.

Table 3. Statistical metrics of surface meteorological conditions, net radiation, and surface turbulent fluxes during the weak PCAP case of IOP3. The lowest RMSE for each simulated variable is marked in boldface type.

<table>
<thead>
<tr>
<th>Runs</th>
<th>$T$</th>
<th>WS</th>
<th>$R$</th>
<th>$R_s$</th>
<th>$H$</th>
<th>LE</th>
<th>$G$</th>
<th>$u_*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAM_ACM2</td>
<td>MB</td>
<td>$-2.93$</td>
<td>$-0.28$</td>
<td>$-0.40$</td>
<td>$21.27$</td>
<td>$17.90$</td>
<td>$7.62$</td>
<td>$-24.74$</td>
</tr>
<tr>
<td>MAE</td>
<td>2.98</td>
<td>1.06</td>
<td>0.64</td>
<td>31.58</td>
<td>24.04</td>
<td>15.10</td>
<td>30.33</td>
<td>0.13</td>
</tr>
<tr>
<td>RMSE</td>
<td>3.64</td>
<td>1.37</td>
<td>0.82</td>
<td>$50.52$</td>
<td>35.52</td>
<td>23.79</td>
<td>34.83</td>
<td>0.18</td>
</tr>
<tr>
<td>NAM_YSU</td>
<td>MB</td>
<td>$-1.90$</td>
<td>$-0.03$</td>
<td>$-0.58$</td>
<td>$39.34$</td>
<td>$12.64$</td>
<td>$5.22$</td>
<td>$-14.23$</td>
</tr>
<tr>
<td>MAE</td>
<td>2.01</td>
<td>0.93</td>
<td>0.70</td>
<td>43.99</td>
<td>16.73</td>
<td>11.53</td>
<td>22.59</td>
<td>0.12</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.57</td>
<td>$1.24$</td>
<td>0.83</td>
<td>$55.38$</td>
<td>25.25</td>
<td>16.97</td>
<td>29.19</td>
<td>0.17</td>
</tr>
<tr>
<td>NAM_MYJ</td>
<td>MB</td>
<td>$-1.61$</td>
<td>$-0.44$</td>
<td>$-0.50$</td>
<td>$40.57$</td>
<td>$14.11$</td>
<td>$5.93$</td>
<td>$-15.24$</td>
</tr>
<tr>
<td>MAE</td>
<td>2.00</td>
<td>1.00</td>
<td>0.73</td>
<td>43.63</td>
<td>18.99</td>
<td>12.69</td>
<td>22.33</td>
<td>0.11</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.57</td>
<td>1.29</td>
<td>0.89</td>
<td>$55.00$</td>
<td>27.47</td>
<td>19.25</td>
<td>28.35</td>
<td>$0.16$</td>
</tr>
<tr>
<td>NAM_MYNN</td>
<td>MB</td>
<td>$-3.11$</td>
<td>$-0.60$</td>
<td>$-0.41$</td>
<td>$41.71$</td>
<td>$10.95$</td>
<td>$1.42$</td>
<td>$-8.75$</td>
</tr>
<tr>
<td>MAE</td>
<td>3.11</td>
<td>1.01</td>
<td>0.63</td>
<td>44.99</td>
<td>16.18</td>
<td>8.34</td>
<td>17.10</td>
<td>0.11</td>
</tr>
<tr>
<td>RMSE</td>
<td>3.73</td>
<td>1.33</td>
<td>$0.74$</td>
<td>$55.66$</td>
<td>$24.74$</td>
<td>$11.73$</td>
<td>$23.24$</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 4. As in Table 3, but for the strong PCAP case of IOP5.
dissipation, which partially relies on realistic surface wind fields. Specifically, the PBL schemes that employ nonlocal mixing approaches, ACM2 and YSU, generated stronger surface WS than the local PBL schemes (MYJ and MYNN) at 1200 MST 14 December. Overall, the NAM_YSU case outperformed the other three cases in the simulation of WS with an MAE of 0.93 m s$^{-1}$ and RMSE of 1.24 m s$^{-1}$. For reference, simulated and observed WDIR are presented in Fig. 2 and will be used in the discussion of model discrepancies associated with terrain flows in section 4 below.

The observed average $T$ at site DL2 during IOP5 in January was only 265.72 K (Fig. 3), about 15 K lower when compared with the weak PCAP case (IOP3) in December (Fig. 2). The NAM_ACM2 case generated an overall warm-biased $T$ with an MB of 0.15 K, while the other three cases underestimated $T$ with an MB of $-1.19$ K. The model spread, as well as the model deviation from observations, was still wider at nighttime compared to daytime. The cold bias during nighttime was as high as $-9.16$ K, which occurred around midnight on 2 January on the NAM_MYNN case. Daytime $T$ was underestimated in the first half of IOP5 and then was overestimated after 5 January, when the surface stability returned to strong stable conditions after some perturbations caused by cold-air advection aloft on 4 January accompanied with a upper-level trough passage (Foster et al. 2017). The cold air aloft may have mixed down to the surface on 5 January and was not captured by the model, thus resulting in an overall warm bias afterward. The NAM_ACM2 case simulated the most realistic $T$ with an MAE of 1.63 K and RMSE of 2.42 K.

The mean observed WS during IOP5 (1.19 m s$^{-1}$) was slightly weaker than IOP3 (1.65 m s$^{-1}$). The model captured the magnitudes of WS well with an MB of $-0.32$ m s$^{-1}$ but did not capture the temporal variation. It has been reported that the WRF Model systematically overestimates WS under stable conditions (Gilliam et al. 2012). Unlike IOP3, for which the NAM_YSU run performed best, the NAM_MYNN case produced the most realistic WS in IOP5 with an MAE of 0.61 m s$^{-1}$ and RMSE of 0.78 m s$^{-1}$.

The average $r$ during IOP5 was 2.03 g kg$^{-1}$. All four simulations generated a wet $r$ bias during daytime and
nighttime, with an average MB of 0.15 g kg\(^{-1}\). This implies an excessively simulated surface evapotranspiration, possibly accompanied by an overestimated latent heat flux. The NAM_MYNN case had the least wet bias of \(r\) with an MAE of 0.32 g kg\(^{-1}\) and RMSE of 0.45 g kg\(^{-1}\). Unlike the weak PCAP, the NAM_ACM2 run generated the highest correlation coefficients in all three surface meteorological parameters, which were 0.60, 0.26, and 0.61, for \(T\), \(WS\), and \(r\), respectively.

b. Radiation components

1) Weak PCAP Case Study (IOP3)

Reliable simulations of radiation components are of primary importance because they provide the atmospheric heating rate in numerical models (Ruiz-Arias et al. 2013) and are major terms in the surface energy budget. Figure 4 shows the simulated and observed four-component radiative fluxes and net all-wave radiation (i.e., shortwave and longwave) at site DL2 during IOP3. While the radiation models were the same in our sensitivity experiments, it is expected the land–atmosphere and boundary layer mixing feedbacks on cloud formation will impact the surface radiation components. Therefore, we evaluate the differences in the surface radiation components here because the net radiation is a major factor in the magnitude of the surface turbulent fluxes.
The mean observed downward shortwave radiation (DSR) was 63.92 W m$^{-2}$. All four simulations exhibited positive mean biases (average 24.78 W m$^{-2}$) in DSR, with the NAM_MYJ showing a slightly lower MAE (28.17 W m$^{-2}$) and RMSE (62.38 W m$^{-2}$) than the other three. The greatest overestimation occurred around 1200 MST 14 December when the observed daily maximum DSR dropped from 331.26 W m$^{-2}$ (on 13 December) to 87.33 W m$^{-2}$. The narrower bell shape of the modeled DSR diurnal distribution on 14 December indicated that the model captured the cloud enhancement on that day more abruptly and slightly later compared to observations and exhibited positive biases in DSR. The DSR overestimation was probably caused by underestimation of the cloud coverage and/or a deficient representation of the cloud radiative impact (Ruiz-Arias et al. 2016), such as cloud albedo and absorption.

The spread of model results for upward shortwave radiation (USR) was larger relative to DSR. The overall average observed USR was 11.18 W m$^{-2}$. The NAM_ACM2 run exhibited a negative mean bias (−1.39 W m$^{-2}$) in USR, whereas USR was positively biased in the other three runs (average 7.51 W m$^{-2}$). Similar to the observations, USR is determined by surface albedo \(\alpha\) and DSR in numerical models. Surface \(\alpha\) in numerical models is parameterized based on zenith angle and surface conditions. Thus, biases in the modeled USR can be attributed to biases in simulated DSR and surface albedo (King et al. 2015). The observed \(\alpha\) was calculated based on the ratio of observed USR over DSR. The mean observed daytime \(\alpha\) during IOP3 was 0.17. The negative bias of \(\alpha\) in NAM_ACM2 (mean daytime \(\alpha\) was 0.11) offset its overestimation of DSR, leading to its best performance in the simulation of USR. NAM_YSU showed the largest bias in USR, resulting from the overestimated DSR as well as the positive \(\alpha\) bias (mean daytime \(\alpha\) was 0.23). The lower \(\alpha\) from the NAM_ACM2 case was mainly attributed to the lower background albedo (0.12) from the Pleim–Xiu LSM as compared with Noah (0.2). The Pleim–Xiu LSM uses the land-use fraction to calculate the background albedo, while Noah uses the dominant land use (Byun et al. 2011). The background albedo difference overpowered the soil moisture discrepancies, which is one important parameter impacting the time-varying \(\alpha\), since Noah overpredicted soil moisture [discussed in section 3c(1)] and should theoretically have a lower \(\alpha\).

Unlike DSR and USR, there was no discernable diurnal variation for the observed or simulated downward longwave radiation (DLR). This is expected since in addition to the atmospheric temperature, DLR is impacted by the sky emissivity, which depends on both air temperature and humidity (Iacono et al. 2000). The overall mean DLR during IOP3 was 289.48 W m$^{-2}$. DLR was negatively biased in all four WRF runs with an average MB of −19.23 W m$^{-2}$. The underestimation of DLR was partly attributed to the negative bias of simulated \(T\) since DLR is parameterized proportionally to the fourth power of atmospheric temperature. Similar to DSR, there was no discernable difference in the modeled DLR among the four WRF simulations except in the morning of 14 December, implying simulation discrepancies in cloud properties during that time. The NAM_ACM2 case showed a slightly lower MAE (20.89 W m$^{-2}$) and RMSE (32.13 W m$^{-2}$) than the other three cases.

Bias in upward longwave radiation (ULR) was also lowest in the NAM_ACM2 case, which had an MB of −14.59 W m$^{-2}$, MAE of 14.70 W m$^{-2}$, and RMSE of 17.60 W m$^{-2}$. This might be attributed to the indirect soil nudging scheme in the Pleim–Xiu LSM. Simulated ULR by WRF runs using the same Noah LSM but with different PBL schemes (NAM_YSU, NAM_MYJ, and NAM_MYNN) varied little during IOP3, suggesting its higher sensitivity to LSMs than to PBL schemes. The negative bias of ULR in the WRF Model with an MB of −36.73 W m$^{-2}$ indicated an unrealistic weak radiative cooling, which can be ascribed to a low \(T_s\). This supports the discussion of \(T\) in section 3a for IOP3, in which we hypothesize that the cold bias of \(T\) is associated with a negative bias in \(T_s\).

The net all-wave radiation \(R_n\) was calculated based on the four-component radiative fluxes analyzed above. The WRF Model exhibited an overall positive bias of 35.72 W m$^{-2}$ in \(R_n\), with excessive surface heating during daytime and insufficient surface cooling during nighttime (Fig. 4e). The NAM_ACM2 case performed best in simulating \(R_n\) with an MAE of 31.58 W m$^{-2}$ and RMSE of 50.52 W m$^{-2}$.

2) STRONG PCAP CASE STUDY (IOP5)

The maximum DSR during IOP5 reached 597.87 W m$^{-2}$ (Fig. 5), about 75 W m$^{-2}$ higher compared to the peak DSR during IOP3. Meanwhile, higher average DSR was observed in this strong PCAP case (87.45 W m$^{-2}$) as compared with the weak PCAP case (63.92 W m$^{-2}$). All four WRF runs produced a negative bias in simulating DSR with an average MB of −14.10 W m$^{-2}$. The NAM_ACM2 case simulated unrealistically low DSR at 1200 MST 4 January. This lower DSR from NAM_ACM2 relative to observations and the other three WRF runs was associated with its higher simulated cloud amount, also supported by its higher DLR shown in Fig. 5c. The NAM_MYJ case again generated the most realistic
DSR with an MAE of 26.69 W m$^{-2}$ and RMSE of 62.30 W m$^{-2}$.

Recall that USR is determined by both DSR and surface $\alpha$. The average daytime $\alpha$ was 0.80 due to snow coverage. With the enhanced DSR and higher surface $\alpha$, the average observed USR during IOP5 (66.85 W m$^{-2}$) was 55.67 W m$^{-2}$ larger than the average during IOP3. The simulated daytime $\alpha$ over the snow-covered surface was underestimated in all four WRF runs, indicating that WRF did not adequately simulate the snow (snow depth measurements were unavailable). The Noah LSM simulated a higher $\alpha$ (average 0.59) than the Pleim–Xiu LSM (average 0.45). With the negative bias of USR and surface $\alpha$, the WRF Model underpredicted USR with an average MB of $-26.40$ W m$^{-2}$. Different from the mixed performances of the four WRF runs during IOP3, the consistent underestimation of USR during IOP5 indicates the dominant role of $\alpha$ in simulating USR during snow-covered periods. Foster et al. (2017) demonstrated that modifying snow cover based on observations can improve the simulated radiative fluxes. The NAM_MYJ case exhibited the smallest bias in USR with an MAE of 24.91 W m$^{-2}$ and RMSE of 53.64 W m$^{-2}$.

The average observed DLR during IOP5 was 236.18 W m$^{-2}$, which was 53.3 W m$^{-2}$ lower than the average value observed during IOP3. The smaller DLR was partly due to the lower air temperature during IOP5. The observed maximum (295.69 W m$^{-2}$) and minimum (168.43 W m$^{-2}$) values of DLR were well captured by the model, but there was greater variability in the simulated DLR from the four sensitivity experiments. It is well acknowledged that the modeled cloud property discrepancies are the main factors that influence DLR (Caron et al. 2013). The wide model spread in the simulation of $r$ (Fig. 3d) also contributed to the variability of modeled DLR, considering that water vapor is one of the major components that impacts DLR. Similar to IOP3, the WRF Model exhibited negative biases in simulation of DLR, with an average MB of $-11.02$ W m$^{-2}$. The modeled DLR from the NAM_MYNN case compared best with observations with an MAE of 26.19 W m$^{-2}$ and RMSE of 35.37 W m$^{-2}$.

As expected, the average observed ULR during IOP5 (266.82 W m$^{-2}$) was smaller than IOP3 (340.60 W m$^{-2}$) because of the low temperatures in this strong PCAP case. The NAM_ACM2 case using the Pleim–Xiu LSM exhibited positive biases in ULR with an average MB of 8.04 W m$^{-2}$, while the other three cases using the Noah LSM showed negative biases, with an average MB of $-12.15$ W m$^{-2}$. The overestimations of ULR during both daytime and nighttime from the NAM_ACM2 case indicate a warm bias of $T_s$. Generally, a higher modeled $T_s$ tends to produce excessive surface heat transfer and thus leads to a higher $H$. Indeed, the NAM_ACM2 case generated the largest $H$ among the four WRF cases (further discussed in section 3c below). Still, the NAM_ACM2 run had the smallest bias in simulating ULR with an MAE of 13.54 W m$^{-2}$ and RMSE of 16.73 W m$^{-2}$.

As a result of the high daytime surface $\alpha$ observed during IOP5 because of snow cover, more solar radiation was reflected back to the atmosphere. The low
daytime $R_n$ (average 41.89 W m$^{-2}$) results in the negative daily (daytime and nighttime) mean $R_n$ during IOP5, which was $-9.18$ W m$^{-2}$. Biases of the four radiation components discussed above accumulated and lead to the highest positive $R_n$ bias during daytime ($36.43$ W m$^{-2}$) and highest negative $R_n$ bias during nighttime ($-19.99$ W m$^{-2}$) in the NAM_ACM2 case. Overestimation of $R_n$ in numerical models has the potential to weaken the CAP strength, especially at night with weak surface radiative cooling being simulated. Zamora et al. (2003) found that an 80 W m$^{-2}$ decrease in the forecast radiation flux divergence could inhibit the formation of nighttime inversions using the fifth-generation Pennsylvania State University–NCAR Mesoscale Model (MM5). The WRF results are similar, where the bias in $R_n$ is associated with a weaker simulated CAP strength quantified by valley heat deficit (average 4.95 MJ m$^{-2}$) when compared with sounding observations (average 6.25 MJ m$^{-2}$) during IOP5. Slightly excessive surface radiative cooling during nighttime occurred in the Noah LSM with an MB of $-1.22$ W m$^{-2}$. The NAM_MYNN case simulated the most realistic $R_n$ with an MAE of 21.61 W m$^{-2}$ and RMSE of 27.73 W m$^{-2}$.

c. Surface turbulent fluxes

1) WEAK PCAP CASE STUDY (IOP3)

Figure 6 compares the modeled and measured surface turbulent heat fluxes and friction velocity during IOP3. The model captured the observed time variation of $H$ well except for the significant positive bias on 13 December at night. The high magnitude of observed $H$ on 13 December at night was related to strong mechanical turbulent mixing (a high observed $u^{*}$ shown in Fig. 6d). However, the model did not reproduce the high $u^{*}$ at that time and underestimated the amplitude of $|H|$. The occurrence of downward $H$ during daytime on 14 December can be explained by warm air advection near the surface as evidenced by an elevated $T$ shown in Fig. 2. Overall, the model displayed a mean positive bias (16.64 W m$^{-2}$) in $H$, both during daytime (14.30 W m$^{-2}$) and nighttime (17.92 W m$^{-2}$). The overestimated $R_n$ discussed above partly contributes to the positive bias of $H$. The NAM_YSU reproduced the measured $H$ most closely with an MAE of 22.10 W m$^{-2}$ and RMSE of 31.07 W m$^{-2}$.

The model performance on simulating latent heat flux (LE) was less straightforward than $H$, where some cases have a mean positive bias (NAM_YSU) and others negative (NAM_ACM2, NAM_MYJ, and NAM_MYNN). LE is closely related to soil moisture. The observed average soil moisture ($Q_{soil}$) was 0.13 m$^3$ m$^{-2}$ measured at 5 cm underground (Fig. 7). The simulated average $Q_{soil}$ at the first model soil layer from Pleim–Xiu LSM (1 cm) and Noah LSM (10 cm) was 0.09 and 0.24 m$^3$ m$^{-2}$, respectively. In general, the $Q_{soil}$ vertical gradient has been found to be around +0.02 m$^3$ m$^{-3}$ per 5 cm from 0 to 10 cm (Kumar et al. 2014; Sadeghi et al. 2016). This means the near-surface $Q_{soil}$ was overestimated in the Noah LSM and underestimated in the Pleim–Xiu LSM. It is true that LE had a positive bias in the Noah LSM, with an average MB of 3.60 W m$^{-2}$, and a negative bias in the Pleim–Xiu LSM, with an average MB of $-2.57$ W m$^{-2}$, during daytime. However, all simulations underestimated LE at night, with an average MB of $-6.55$ W m$^{-2}$. The average measured LE at night was 6.94 W m$^{-2}$, suggesting a high surface $Q_{soil}$. This implies that the simulated soil moisture deficiency plays a less important role in contributing to LE biases at night.
Other possible contributors to the LE biases include the near-surface humidity biases and unrealistic parameterization of stomatal resistance employed in the evapotranspiration parameterization module in LSMs during nighttime.

Observations of ground heat flux $G$ take more effort to obtain compared with radiation and surface turbulent heat fluxes since the measurement sensors for $G$ need to be installed in the soil. The average measured $G$ (measured at 5 cm underground) was 17.54 W m$^{-2}$ during daytime and $-16.45$ W m$^{-2}$ during nighttime in IOP3. All four WRF runs generated a positive bias with an average MB of 12.18 W m$^{-2}$ during daytime and a negative bias with an average MB of $-20.37$ W m$^{-2}$ during nighttime. $G$ is determined by the vertical gradient of soil temperature $T_{\text{soil}}$ between the surface and underground in both observations and simulations (Liebenthal and Foken 2007). The surface skin temperature was underestimated all the time in all four WRF runs (i.e., negative ULR biases shown in Fig. 4). The observed $T_{\text{soil}}$ at the first model soil-layer depth, which is 1 cm in the Pleim–Xiu LSM and 10 cm in the Noah LSM, was not available in our case. We use the $T_{\text{soil}}$ measured at 0.6 and 4.4 cm to interpret the results qualitatively. $T_{\text{soil}}$ at 0.6 cm was warmer during daytime and cooler during nighttime than 4.4 cm (Fig. 7). This is expected because the shallower soil is more sensitive to the surface variations, which are subject to daytime solar heating and nighttime radiative cooling. With the assumption of a linear relationship of soil temperature and soil depth, we can infer that $T_{\text{soil}}$ was overestimated during daytime and underestimated during nighttime in both LSMs. The daytime overestimation of $G$ was due, in part, to the warm bias of $T_{\text{soil}}$ and cold bias of surface skin temperature. The nighttime underestimation of $G$ was mainly attributed to the cold bias of $T_{\text{soil}}$. Also, the modeled peak of $G$ occurred earlier in the Noah LSM and late in the (Pleim–Xiu LSM when compared with measurements.

There was no general diurnal variation of friction velocity $u^*$ observed during IOP3 (Fig. 6). The average $u^*$ during daytime and nighttime was 0.22 and 0.21 m s$^{-1}$, respectively. Recall that the high WS caused by drainage flows starting on the afternoon of 12 December and lasting through the night (Lu and Zhong 2014) was not reproduced by the model (Fig. 2). This contributed to the overall negative bias of $u^*$ with an average MB of $-0.10$ m s$^{-1}$, which relies on realistically simulated wind fields. The NAM_MYJ case performed best in simulating $u^*$ with an MAE of 0.17 m s$^{-1}$ and RMSE of 0.23 m s$^{-1}$.

2) STRONG PCAP CASE STUDY (IOP5)

The peak value of the observed $H$ during IOP5 was 38.27 W m$^{-2}$, which was 61.37 W m$^{-2}$ smaller than the maximum $H$ observed during IOP3. The smaller $H$ during IOP5 was partially due to a decreased $R_{\text{n}}$ (Fig. 5e), as well as its stronger PCAP strength. The daytime maximum $H$ was overpredicted by the WRF Model, especially in the NAM_ACM2 run, during this strong PCAP case (Fig. 8), implying that the atmospheric stability impacts the PBL scheme and LSM performance, where strongly stable atmospheric boundary layers (ABLs) cannot be well simulated. The average MB in the simulation of $H$ during daytime was 13.28 W m$^{-2}$, which was largely contributed by the positive bias from NAM_ACM2 (27.14 W m$^{-2}$). The large overestimation of the NAM_ACM2 case was related to its high $T_s$ together with a large ULR (Fig. 5d). The overestimations of $H$ during nighttime from the four WRF runs were similar, with an average MB of 5.73 W m$^{-2}$.

Excessive high biases in modeled $H$ can lead to an unrealistically warm boundary layer and thus deteriorate the model performance in the prediction of PCAP life cycle. The NAM_MYNN case produced the smallest bias in the simulation of $H$ with an MAE of 10.89 W m$^{-2}$ and RMSE of 17.11 W m$^{-2}$.

Similar to $H$, the observed LE was more dampened during IOP5 than IOP3, with a daytime maximum value of 19.34 W m$^{-2}$. The largest positive bias in simulating daytime LE still came from the NAM_ACM2 case with an average MB of 14.09 W m$^{-2}$. Discrepancies among the four runs in the simulation of LE were not as large as $H$. The overall positive bias of LE (8.95 W m$^{-2}$) by
The model was possibly related to an overestimated surface canopy moisture (Fig. 9a). A slightly drier \( Q_s \) was measured at 5 cm depth during IOP5 (average 0.11 m\(^3\) m\(^{-2}\)) relative to IOP3 (average 0.13 m\(^3\) m\(^{-2}\)). However, both the Pleim–Xiu LSM and the Noah LSM generated enhanced \( Q_s \) at the first soil layer with an average value of 0.44 and 0.29 m\(^3\) m\(^{-2}\), respectively, as compared with IOP3 for which \( Q_s \) from the two LSMs was 0.09 and 0.24 m\(^3\) m\(^{-2}\). It is possible that the warm \( T_s \) in NAM_ACM2 case led to a faster snow melting rate in the model and thus a significantly enhanced \( Q_s \) (+0.35 m\(^3\) m\(^{-2}\)). The slight increase of \( Q_s \) (0.05 m\(^3\) m\(^{-2}\)) from the Noah LSM (at 10 cm) also indicated signs of snow melting in the WRF simulations. Note that soil moisture in deep layers is less sensitive to surface changes compared with shallow layers, which might contribute to the small increase of \( Q_s \) in the Noah LSM. The NAM_MYNN case compared best with observed LE with an average MAE of 7.71 W m\(^{-2}\) and an RMSE of 10.75 W m\(^{-2}\).

The surface snow coverage substantially reduced the soil heat transfer with an average observed \( G \) of \(-6.40\) W m\(^{-2}\) and no discernable diurnal variations (Fig. 8c). The model simulated an average \( G \) of \(-27.92\) W m\(^{-2}\) and largely overpredicted the magnitude of \( |G| \). The negative values of \( G \) during the majority of IOP5 indicated a lower surface skin temperature compared with the deeper soil. The overestimation in the magnitude of \( G \) indicates excessive heat being transferred to the surface layer, and likely resulted in the overestimation of snowmelt. Because of the “insulation” impact of snow, the observed soil temperature difference between 0.6 and 4.4 cm during IOP5 was much smaller than IOP3 (Fig. 9b). The NAM_MYNN case exhibited the smallest bias in \( G \) with an MAE of 8.48 W m\(^{-2}\) and RMSE of 13.57 W m\(^{-2}\).

The mechanical induced turbulence was further dampened during IOP5 (when compared with IOP3), with an average \( u_* \) of 0.12 m s\(^{-1}\) (Fig. 8d), because of the more stable atmospheric boundary layer accompanied by limited radiation. The daytime and nighttime \( u_* \) were respectively overestimated and underestimated by the model with an average MB of 0.05 m s\(^{-1}\) and \(-0.06\) m s\(^{-1}\), respectively. This positive bias of \( u_* \) during daytime contributed to the overestimation of surface turbulent fluxes. The nighttime underestimation of \( u_* \) was possibly related to the model deficiency in capturing the mountain flow over complex terrain, such as katabatic flows at nighttime. The NAM_ACM2 case produced the largest positive
bias in $u_\infty$, with an MAE of 0.11 m s$^{-1}$, whereas the NAM_MYNN compared best, with an MAE of 0.08 m s$^{-1}$.

d. Statistical evaluation metrics over different land-use types

Land–atmosphere interactions depend on land-use characteristics, such as surface emissivity, albedo, and canopy resistance (Tao et al. 2013). Observations over the seven measurement sites during PCAPS allow for additional model evaluations to investigate the spatial variability of surface turbulent heat fluxes, as well as the model performance over different land-use types. Due to limitations related to the availability of data at each site, the results were aggregated over the two IOPs presented above, versus stratifying by each IOP. The statistics, including the RMSE and MAE as well as the averaged observed and simulated values during IOP3 and IOP5, are shown in Fig. 10.

The DH site observed the highest average $H$ (average 20.71 W m$^{-2}$) compared with other sites, which was associated with additional anthropogenic heat sources in an urban environment. Wang et al. (2015) also observed higher $H$ in urban areas compared with rural areas in Beijing, China. However, the model did not replicate the highest $H$ over the DH site. The WRF runs exhibited significant negative biases at the DH site and the DM site with an average MB of $-12.26$ and $-6.42$ W m$^{-2}$, respectively, with positive biases at the other sites with an average MB of 6.31 W m$^{-2}$. Recall that overestimation (underestimation) of $H$ will generate a warm and dry (cold and wet) boundary layer. This implies that the modeled boundary layer structure discrepancies compared with observations are also likely to depend on land-use type. The NAM_ACM2 case produced the largest overprediction compared with other cases, except for the CR site. The DM site had the smallest bias, with an average MAE of 14.01 W m$^{-2}$ and RMSE of 18.13 W m$^{-2}$. The DL1 site had the largest bias, with an average MAE of 28.30 W m$^{-2}$ and RMSE of 40.59 W m$^{-2}$.

The largest observed average LE occurred at the CR site (15.10 W m$^{-2}$). The model produced positive biases in simulating LE most of the time with an average MB of 6.37 W m$^{-2}$, and the PH site exhibited the largest average MB (9.62 W m$^{-2}$). The overestimation of LE was partially due to the positively biased soil moisture, possibly from initial and boundary conditions, as well as unrealistically fast snow melting simulated in the models (as discussed in section 3c). The four WRF runs showed mixed model performance in simulating LE across all sites. The DL2 site exhibited the least overestimation, with an average MAE of 11.91 W m$^{-2}$ and RMSE of 17.94 W m$^{-2}$. The DH site exhibited the most overestimation, with an average MAE of 15.72 W m$^{-2}$ and RMSE of 22.71 W m$^{-2}$.

The model showed a uniform underestimation of $G$ at the observation sites except for site DM, where three of
the four simulation runs show an overestimated $G$. The NAM_ACM2 case exhibited the largest negative MB ($-24.19$ W m$^{-2}$). Overall, the Pleim–Xiu LSM had higher MAE and RMSE compared with the Noah LSM. The three cases using the same Noah LSM (NAM_YSU, NAM_MYJ, and NAM_MYNN) simulated similar magnitudes of $G$, which indicates that there is less sensitivity of $G$ to PBL schemes than to LSMs. The DM site had the lowest bias, with an MAE of $11.37$ W m$^{-2}$ and RMSE of $19.86$ W m$^{-2}$. The DL1 site had the highest bias, with an MAE of $35.78$ W m$^{-2}$ and RMSE of $41.98$ W m$^{-2}$. One can see that the MAE and RMSE of simulated $G$ had a similar magnitude to the simulated $H$ and LE, indicating that the soil temperature profiles in numerical models need to be improved to improve simulations of the surface energy budget.

The average observed $u_*$ ranged from $0.10$ ms$^{-1}$ at the BL site to $0.25$ ms$^{-1}$ at the CR site. The model displayed mixed performances in the simulation of $u_*$ with positive biases at some sites (BL, DH, DL1, and PH) and negative biases at others (DM, DL2, and CR). Surface friction velocity in numerical models depends on the simulated surface roughness lengths. The significant overestimation of $u_*$ at the DL1 site might be attributed to the underestimated roughness lengths, which are related to the subgrid drag effect (Jiménez and Dudhia 2012). The roughness length uncertainties derive from misclassification of the land-use types or unresolved topographical features. The DL1 site exhibited the largest bias, with an MAE of $0.20$ ms$^{-1}$ and RMSE of $0.29$ ms$^{-1}$. The DM site exhibited the smallest bias, with an MAE of $0.08$ ms$^{-1}$ and RMSE of $0.10$ ms$^{-1}$.

4. Discussion

The above results show that during wintertime in the SLV, the WRF Model does not simulate the correct surface turbulent fluxes, especially for strong PCAP conditions. To aid in guiding the direction of future WRF Model improvements, a discussion is given here to investigate the model parameterizations associated with the uncertainties in simulating the turbulent fluxes. The focus here is primarily on the surface turbulent heat flux because the data from the field experiment are the most complete for comparison with the empirical formulations in WRF (i.e., there are measurements to quantify the surface skin temperature but no measurements for the moisture right at Earth’s surface). Brief discussions related to the empirical function used to model the momentum flux and regarding the ability of WRF to simulate terrain flows are also given.

The surface sensible heat flux is determined by the product of surface exchange coefficients ($CH$; m s$^{-1}$) and near-surface gradients of temperature in LSMs, and for latent heat flux the near-surface gradients of moisture. The surface exchange coefficients are calculated in surface-layer schemes based on M–O similarity theory (Monin and Obukhov 1954), which describes the mean wind speed and mean temperature as a function of the atmospheric stability within the surface layer. We compared CH calculated from the sonic anemometer observations (Sun and Holmes 2019) and outputs from the four WRF cases grouped by stability (M–O stability $\zeta$) to investigate the performance of surface-layer schemes under different stability regimes. Besides being a function of atmospheric stability, CH varies over land-use types with different roughness lengths, which is one crucial parameter used in M–O similarity theory. Representation of the land–atmosphere coupling strength over different land-use types in different surface-layer schemes is also examined by comparing the spatial variability in CH from the observations to the WRF results.

Given the stably stratified atmosphere during the wintertime in Utah, especially over snow-covered surfaces, the simulated and observationally estimated CH variations with atmospheric stability under stable conditions at the DL2 site (Fig. 11) are investigated here. The overall decreasing tendency in CH values with increasing stability calculated from observations was well captured by the model. The average median observed CH during slightly stable conditions ($0.025 < \zeta \leq 1.0$) and very stable conditions ($\zeta > 1.0$) was $1.34 \times 10^{-2}$ and $0.54 \times 10^{-2}$ m s$^{-1}$, respectively. All of the WRF scenarios generated an overestimation of CH, except for the NAM_MYJ case. The NAM_ACM2 case produced the highest CH in almost all stability ranges under stable conditions ($\zeta > 0.25$) with an average median value of

![Fig. 11. Median values of surface exchange coefficients (CH) simulated from the four WRF runs (with an average standard deviation of 0.007 m s$^{-1}$) and estimated from sonic anemometer measurements (with a standard deviation of 0.05 m s$^{-1}$) using the bulk transfer equations [Eq. (8)] at the DL2 site during IOP3 and IOP5 vs Monin–Obukhov stability parameter $\zeta$ bins. The x-axis tick label represents the upper limit of the corresponding stability bin. The lower stability bin limit is the preceding tick label.](http://journals.ametsoc.org/jamc/article-pdf/59/6/1029/4946672/jamcd190250.pdf)
1.1 \times 10^{-2} \text{ m s}^{-1}, which contributed to its highest positive bias in the simulation of $H$ for the strong PCAP case (Fig. 8a). CH from the NAM_MYJ case exhibited a concave upward distribution at the stability range of $0.025 < \zeta \leq 0.125$, which should lead to lower $H$ values being simulated. However, comparing the time series of $H$ in Fig. 8a it can be seen that the underestimation of CH did not lead to a negative bias of $H$ in the NAM_MYJ case. At first, we thought this was due to the lower surface stability simulated in the NAM_MYJ case. However, the NAM_MYJ case actually generated higher surface stability compared with the other WRF simulations and higher stability than the observations (figure not shown). This suggests that the lack of a negative bias (associated with the underestimated CH) in the surface turbulent fluxes in the NAM_MYJ case could be related to the resolution of the stability classifications used in the eta similarity surface-layer scheme compared with the other WRF surface-layer schemes. The eta similarity scheme calculates the similarity profile functions for only two stability classifications, where it defines stable (unstable) conditions when Richardson number is larger (smaller) than 0 (Janjič 2001). This simple stability criterion with only two classifications, versus three or four stability classifications in the other surface-layer models, may blur the exchange coefficient dependence on surface stability and lead to unrealistic simulations of surface turbulent fluxes. The general overestimation in CH from WRF brings into question the applicability of M–O similarity in the surface layer over complex terrain under stable conditions, where the assumptions, such as horizontally homogeneous and quasistationary, may not apply (Stiperski and Rotach 2016; Stiperski et al. 2019).

To calculate the bulk transfer coefficients using the bulk transfer method [e.g., Eq. (8) for $H$], suitable functions for both dimensionless wind and temperature vertical profiles ($\phi_m$ and $\phi_h$) are needed. Figure 12 presents the scatterplots of $\phi_m$ and $\phi_h$ as a function of M–O stability $\zeta$ under stable conditions calculated from observations, as well as two stability function curves proposed by Dyer (1974) and Chenge and Brutsaert (2005), which are adopted in the WRF surface-layer schemes used in our case study. The NAM-YSU uses Chenge and Brutsaert (2005), and NAM_ACM2, NAM_MYJ, and NAM_MYNN use Dyer (1974), with minor modifications to these curves in each surface-layer scheme, specific details for each surface-layer scheme can be found in the corresponding reference listed in Table 2. Figure 12 shows that the dimensionless gradient functions for $\phi_m$ performed better under slightly stable conditions ($0.025 < \zeta \leq 1.0$) and tends to overestimate $\phi_m$ under very stable conditions ($\zeta > 1.0$). The Chenge and Brutsaert (2005) function gave a better estimated $\phi_m$ based on $\zeta$ under very stable conditions.

The dimensionless temperature gradients calculated from the two stability functions underestimated $\phi_h$, especially for the slightly stable conditions ($0.025 < \zeta \leq 1.0$). The major underestimation of $\phi_h$ in the stability range of $0.025 < \zeta \leq 1.0$ was also reported by Vignon et al. (2017) based on measurements conducted in the East Antarctic Plateau. One possible explanation for the scatter in $\phi_h$ is neglecting the radiation divergence impacts on temperature profiles close to the surface under stable conditions. The radiative divergence is mostly located near the surface, where steep vertical gradients of temperature and moisture usually exist (Gentine et al. 2018). It could amplify the temperature profile curvature and lead to elevated $\phi_h$ (Garratt and Brost 1981).
Gentine et al. (2018) found that the radiation flux vertical divergence was coupled to surface turbulent fluxes by showing a strong negative correlation at a grassland site in the Netherlands. This suggests the potential failure of M–O similarity theory for stable conditions, where the radiation divergence impact is not included.

Another experimental study also found the near inverse behavior of the local $\phi_u$ versus $\zeta$ over steep terrain in the Swiss Alps at nighttime (Oldroyd et al. 2016a). Their hypothesis is that the elevated turbulent kinematic energy (TKE) is associated with the katabatic flow in mountain areas and yields a stronger mixing with increased stability (H. J. Oldroyd 2018, personal communication). Katabatic jets are characterized by decreased buoyant suppression and in some cases, production of TKE due to the impacts of the along-slope vector component of buoyancy flux on the net vertical buoyancy flux in the TKE budget equation (Oldroyd et al. 2016b,c). These findings highlight the requirement for improving flux–profile relationships over complex terrain by considering the impacts of the diurnal mountain wind system.

Realistic simulation of $u^*$ over complex terrain also relies on a well-represented diurnal mountain wind system (Zardi and Whiteman 2013) in models. The negative bias of WS in our case might be attributed to some terrain-forced flows that were not dynamically resolved by WRF, where the WDIR also deviated from observations. Underestimation of $u^*$, for example, at the DL2 site located near a steep slope, was likely related to the unresolved downslope flow in WRF, which is related to the surface radiative cooling at nighttime (as discussed in section 3c(2)). Similar terrain impacts on surface winds have been found in previous simulations of the PCAPS time period for IOP1 (Wei et al. 2013) and IOP3 (Lu and Zhong 2014). The focus of both papers was on the mesoscale conditions and not the surface turbulent fluxes; therefore, they provide a comparison of the simulated and observed surface winds in Fig. 7 of Wei et al. (2013) and Fig. 4 of Lu and Zhong (2014). The general finding is that sites closer to the mountain slopes had larger discrepancies in the simulated surface winds due to the unresolved surface and terrain characteristics in WRF (Lu and Zhong 2014).

We do not have surface flux monitors located on sloped terrain to provide a detailed analysis on the terrain impacts of the flow, but we have shown that the flux–profile behavior on the valley floor also needs to be investigated. Our results, the scatterplots shown in Fig. 12, provide a prototype for the potential flux–profile behavior that could better simulate the land–atmosphere exchange during stable boundary layer events. However, additional measurements are needed to investigate the uncertainties associated with the eddy covariance data processing steps required to calculate the turbulent fluxes before new flux–profile relationships can be formulated. We suggest that more experimental data be collected in future field campaigns during wintertime in mountainous regions to obtain the turbulence and mean profile data required to develop new empirical flux–profile relationships.

In addition to the flux–profile functions for mean temperature and wind speed, the simulation of surface turbulent fluxes relies on realistic parameterizations of the momentum roughness length $z_{0m}$ and thermal roughness length $z_{0t}$ (used for CH calculation in Eq. (9)), which depend on land-use type. Thus, variations in CH are expected at different sites (Fig. 13). The average observed median value of CH during the two PCAPs ranged from $1.19 \times 10^{-3}$ m s$^{-1}$ at the CR site to $7.25 \times 10^{-3}$ m s$^{-1}$ at the DH site. A higher CH is associated with rougher surfaces and indicates a stronger coupling between the surface and atmosphere (Chen and Zhang 2009). Simulated CH from the WRF runs showed an overall overestimation across all sites. Real-time or well-represented land-use information in LSMs would improve simulation results with more realistic parameters, such as canopy height, canopy stomatal conductance, and surface albedo. The CH bias varied over different land-use types, with the least bias at the DH site (Fig. 13). The varied model performance and model spread of the four WRF runs in simulating CH at different sites might be related to the different treatments of $z_{0t}$ in the Pleim–Xiu and Noah LSMs. In the Noah LSM, $z_{0t}$ is related to $z_{0m}$ as a
function of roughness Reynolds number and one empirical coefficient, which depends on different canopy heights (Chen and Zhang 2009). While the Pleim–Xiu LSM adds an extra quasi-laminar boundary layer resistance $R_a$, which is a function of friction velocity, to the aerodynamic resistance to correct for the inconsistencies of $z_0$ and $z_{om}$ (Pleim 2006). It is possible that the approach adopted by Pleim–Xiu LSM somehow generates a higher $z_0$ and leads to the overall higher CH compared to the Noah LSM. Again, we suggest that more data over many land-use types be collected in future experiments to better quantify the impact of empirical formulations for the surface roughness lengths on the simulated fluxes.

In addition to improved land–atmosphere exchange simulations, there are other factors that could contribute to the ability for numerical models to simulate cold-air pool events, such as increased vertical resolution, coupled meteorological and air quality modeling to consider aerosol impacts, and to adopt urban canopy schemes. For example, Vosper et al. (2013) found that using higher vertical resolution lead to a lower temperature bias during CAPs. However, Wei et al. (2013) found that increased vertical resolution did not have a significant impact on their model performance for CAP simulations in the Salt Lake Valley. Air pollutants tend to accumulate during CAPs due to suppressed vertical mixing and lower boundary layer heights, and the enhanced aerosol loadings can reduce surface net radiation through aerosol scattering and absorption. Including the aerosol impacts in numerical models, using coupled meteorological and air quality models, can potentially reduce radiation bias. Urban canopy schemes include land-use changes due to urbanization and could lead to improved surface energy exchange simulations due to the more realistic surface characteristics (e.g., surface roughness). While these urban canopy schemes would likely improve the WRF simulation results, these PBL schemes are currently not available in the chemical transport models used to simulate air quality.

5. Conclusions

An evaluation of four PBL schemes (ACM2, YSU, MYJ, and MYNN) with their recommended LSMs and surface-layer schemes in WRF was performed using measurements collected from seven sites with different land-use types during PCAPS in Salt Lake Valley, Utah. The model sensitivity experiments were run for two IOPs during the wintertime of 2010/11, IOP3 and IOP5, featuring weak PCAP and strong PCAP conditions, respectively.

The model–observation comparison for the weak PCAP case study shows that, overall, the daily evolution of surface meteorological fields ($T$, $WS$, and $r$) at the DL2 site during IOP3 were well captured by the model. The model exhibited a cold bias in simulating $T$ with a wide model spread at nighttime, which was attributed to underestimation of $T_a$ (represented by ULR, which is proportional to $T_a$ to the fourth power) and overestimation of $H$. PBL schemes with nonlocal mixing approaches (ACM2 and YSU) simulated the elevated WS near the end of IOP3, suggesting their potential advantage in predicting PCAP break up over the local closure schemes (MYJ and MYNN). Underestimation of WS at night might be related to the model deficiencies in simulating nighttime katabatic flows in mountain areas. For the strong PCAP case (IOP5), which lasted longer than the weak one, the model did not capture the $T$ perturbations caused by cold-air advection aloft.

The variability in the model performance for the four WRF runs was larger in the simulation of the upward radiative fluxes, both shortwave and longwave, than the downward radiation during the weak PCAP case. This indicates the importance of well represented surface $a$ and $T_a$. Underestimation of surface $a$ in the Pleim–Xiu LSM was mainly ascribed to its lower background albedo compared with the Noah LSM. The NAM_ACM2 run outperformed the others in simulating $R_n$. Overall, $R_n$ at site DL2 during IOP3 was overpredicted, both during daytime and nighttime, which weakened the simulated PCAP strength. In the strong PCAP case with snow coverage (IOP5), surface $a$ was underestimated in all WRF runs and lead to the underestimation of USR. The negative bias of surface $a$ from all four WRF runs implied an underprediction of snow-cover fraction and would leave excess energy in the surface layer. This suggests that the model uncertainties and simulation differences among the schemes during snow-covered periods, which usually strengthen the PCAP with decreased $T$, are due to enhanced surface albedo. Unlike IOP3, the NAM_MYNN run produced the least bias in $R_n$ for the strong PCAP case.

Modeled surface turbulent fluxes, including $H$ and LE, compared reasonably well with measurements during the weak PCAP case (IOP3) at the DL2 site. The four WRF runs exhibited mixed model performance for the simulation of LE. In daytime, positive biases of LE during in the Pleim–Xiu LSM and negative biases of LE in the Noah LSM were related to the overestimation and underestimation of $Q_{soil}$, respectively. Daytime warm biases of $T_{soil}$ contributed to the overprediction of $G$, and nighttime cold biases of $T_{soil}$ contributed to the underprediction of $G$. Misrepresentation of drainage flows could lead to the underestimation of $u_*$ at nighttime, which contributed to $H$ and LE discrepancies. The model performance deteriorated for the strong PCAP
case (IOP5) with significant overpredictions of $H$ and LE compared with the weak PCAP case, indicating that the model performance for the variables related to the land–atmosphere exchange depends on the atmospheric stability (i.e., stable ABL conditions are not well simulated). The associated elevated simulated soil moisture might be related to faster snow melting being simulated in the models. Overprediction of $R_s$ during the daytime may also contribute to the positive biases of the $H$ and LE by providing more energy to the surface. The NAM_MYNN run performed best in the simulation of surface turbulent fluxes for the strong PCAP case, which is expected due to its local, higher-order turbulence closure scheme.

In addition to the near-surface temperature and moisture gradient, CH is the other major term in the calculation of $H$ and LE in LSMs. Overall, WRF generated a higher CH compared with observations at the DL2 site under stable conditions. Two typical surface-layer similarity functions used in the WRF Model were evaluated using all of the in situ data collected during PCAPS. The Chenge and Brutsaert (2005) function performed better in estimating $\phi_h$ compared to the Dyer (1974) function. However, neither of the stability functions provided accurate estimates for $\phi_w$ in the SLV during the wintertime for the PCAPS field campaign, with major underestimation for the slightly stable conditions. Neglecting the near-surface radiation divergence impacts, the M–O similarity theory is suspected to account, in part, for the biases in $\phi_h$. One hypothesis for the increased $\phi_h$ with decreased stability is that the positive buoyancy flux associated with katabatic flows in mountain areas enhances the TKE at nighttime.

Our study also highlights the need for well-represented, dynamic land-use information for numerical models, which determines the surface albedo and roughness lengths. For example, snow-covered surfaces are conducing to the formation of strong PCAPs. Under this condition, improvement of the simulated land–atmosphere interactions during strong PCAP events requires realistic snow-related fields, such as snow-cover fraction and snow melting. These fields are closely related to surface albedo and soil moisture, which further impact surface radiation fluxes and turbulent fluxes. The behavior of gradient similarity functions under stable conditions in mountainous areas needs further investigation by including more measurement sites. A more appropriate flux–profile parameterization is needed in mesoscale numerical models to incorporate the impacts of radiative divergence and mountain flow under stable conditions.

Acknowledgments. We acknowledge the Persistent Cold-Air Pool Study field campaign research group led by C. David Whiteman at the University of Utah for providing the observation data and the NCAR/EOL team who made these measurements and processed the ISFS data.

REFERENCES


