Model evaluation of high-resolution urban climate simulations: using WRF/Noah LSM/SLUCM model (Version 3.7.1) as a case study

Zhiqiang Li[1], Yulun Zhou[2], Bincheng Wan[3], Hopun Chung[4], Bo Huang[1] and Biao Liu[5]

[1]Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, 999077, China
[2]Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, 999077, China
[3]Glarun Technology Co., Ltd., Nanjing, 211100, China
[4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, 999077, China
[5]Research Institute of Tsinghua University in Shenzhen, Shenzhen, 518057, China

Correspondence to: Bo Huang (bohuang@cuhk.edu.hk)

Zhiqiang Li and Yulun Zhou contributed equally to this work and should be considered co-first authors.

Abstract. The veracity of urban climate simulation models should be systematically evaluated to demonstrate the trustworthiness of these models against possible model uncertainties. However, existing studies paid insufficient attention to the model evaluation; most studies only provided some simple comparison lines between modelled variables and their corresponding observed ones on the temporal dimension. Challenges remain since such simple comparisons cannot concretely prove that the simulation of urban climate behaviours is reliable. Studies without systematic model evaluations, being ambiguous or arbitrary to some extent, may lead to some seemingly-new but scientifically-misleading findings. To tackle these challenges, this article proposes a methodological framework for the model evaluation of high-resolution urban climate simulations and demonstrates its effectiveness with a case study in the area of Shenzhen and Hong Kong, China. It is intended to remind (again) urban climate modellers of the necessity of conducting systematic model evaluations in urban-scale climatology modelling and reduce these ambiguous or arbitrary modelling practices.

1 Introduction

Recently, studies on urban climate have received growing attention. It is forecasted that there will be 66% of the world's population living in the urban area by 2050 (United Nations, 2014). The fundamental well-being of the urban population, such as their comfort and health, is directly and significantly affected by urban meteorological conditions, such as temperature, wind speed, and air pollution. Meanwhile, the ongoing global trend of climate change adds to the urgency and significance of achieving better understandings of urban climate and obtaining more precise predictions of future changes. In this vein, many tools have been developed, and the rapidly developing urban climate simulation models are among the most powerful ones. These simulation models have been widely applied in analyses and predictions of urban climate, as well as assessments of urban climate impacts brought by the dramatic human interferences in cities (Dale, 1997; Kalnay and Cai, 2003).

Model evaluation is necessary for urban climate simulations to make sure the results are reliable and trustworthy to some extent. Model evaluation refers to comparisons between the modelled variables and its corresponding observed ones. After modelling, a model evaluation should be conducted for establishing the trustworthiness to the modelling results because of the model incompleteness caused by the approximations and assumptions in the scientific mechanisms of the model even if the model was configured appropriately. Urban climate simulation is employed to obtain fine-scale details from the lateral boundary condition of coarse-scale meteorological data by using a limited area model. Moreover, in order to construct precisely the fine-scale details at utmost in the area of interest, the model takes the area of interest land surface forcing into account (Lo et al., 2008). The fine-scale details are constructed by a limited area model driven by the lateral boundary conditions of coarse-scale meteorological data and land surface forcing data but do not exist in the coarse-scale meteorological data. In its essence, the fine-scale details constructed by a limited area model have the possibility of deviating from their corresponding natural values. Urban climate simulation, with a higher requirement on its resolution (spatial and temporal) and modelling urban climatological phenomena (for example, urban heat island, temperature difference between urban and non-urban areas), is more sensitive to the inadequacies of the atmospheric model, the inappropriate configuration of the modelling system (Warner, 2011) and the quality of input data (Bruyère et al., 2014). Therefore, model evaluation is even more critical for urban climate simulation.

However, recent efforts understandably paid minimum attention to model evaluation in the community of urban climate modellers, which weakens the reliability of every conclusion based on the insufficiently justified model results. Among existing literature, researchers mostly conducted some simple comparisons between modelled variables and their corresponding observed ones by drawing their short-term time-history plots. For example, Jiang et al. (2008) made a bold prediction that the near-surface temperature in the Huston area will increase by 2°C in future years (2051–2053). However, the conclusion was only supported by a simple comparison between the observed and WRF-modelled diurnal 2-meter air temperature during August 2001–2003. Meng et al. (2011) modelled the 2-meter air temperature and heat island intensity by using three different modelling schemes, thus concluded which one is best in modelling performances. However, these seemingly robust conclusions are only based on a comparison of the observed temperatures with their corresponding modelled ones over a period of 3 days. With a simple model evaluation comparing diurnal patterns of 3-months-WRF-modelled 2-m surface temperature, special humidity, and relative humidity with its corresponding observed ones, Yang et al. (2012) asserted that the WRF model could reconstruct the urban climate features at high resolution of 1-km with a good performance in modelled surface air temperature and relative humidity in the Nanjing area. Although the afore-mentioned efforts partially addressed the evaluation issue, significant challenges remain in establishing the trustworthiness of the model: Even if an exact match between a modelled variable at some grids and its corresponding observed one in a period cannot conclude that the model simulates urban climate successfully, not to mention a non-exact match. These model evaluation methods are not convincing, or even reckless. That kind of modelling practices without a convincing model evaluation is still prevalent in climate modelling community even for the most recent literature, such as the papers of Gu and Yim (2016), Wang et al. (2016) and Bhati and Mohan (2016). To sum up, it is a blind point in the climate modelling community that the existing studies paid insufficient attention to the model evaluation.

In spite of some previous literatures adverted already the importance of model evaluation in interpreting the modelling results, such as Osborn and Hulme (1997), Caldwell et al. (2009), Gosling et al. (2009) and Sillmann et al. (2013), a systematic framework for model evaluation has not been provided in the previous literature. It is a research gap in urban climatology. Thus, in this paper, we dig deeply into the model evaluation to propose a systematic framework and methods for evaluating model results from multiple perspectives, to benefit future studies with more choices for model quality control and make urban scale simulation more robust. Moreover, we also provide a case analysis of the interval between the modelled atmospheric variable and its corresponding observed one.

The remainder of this paper is organized as follows. Section 2 introduces the proposed framework for model evaluation, experimental design, and data used for modelling and model evaluation. Section 3 introduces the technical preparation for urban climate simulation. Section 4 presents various results of the proposed model evaluation methods in our case study. Section 5 concludes the paper with discussions.

2 Methodology

2.1 Urban Climate Modelling

In an urban area, the natural texture of the land surface has remarkably changed to human-made, impervious land surface. The textural change of the land surface leads to modifications in the interchange of energy, momentum, and mass between the land surface and planetary boundary layer (PBL) (Wang et al., 2009). Moreover, in an urban area, the anthropogenic heat release caused by human activities increases the sensible and latent heat emission. Furthermore, the urban building morphology also has an impact on the radiation exchange and the airflow. Tewari et al. (2007) developed the Urban Canopy Model (UCM) to couple with the Advanced Research WRF (ARW) model via Noah Land Surface Model (Noah-LSM) to improve the simulation accuracy of urban processes by integrating these physical characters below the urban canopy.

We took Shenzhen and Hong Kong, a region in China that had gone through intensive urbanization process, as the study area. We took the year of 2010 as the study period because both of the land surface data and observation data were obtainable in 2010. WRF ARW model coupled with Noah LSM/SLUCM (WRF ARW/Noah LSM/SLUCM v3.7.1) was used for modelling urban climate in 2010 at 1-km^2 grid spacing. Through comparison, we found that some of the terrestrial input data provided by NCAR were out-of-date, especially for data describing the fast-developing area. To reflect more precisely the artificial changes on the physical environment brought by the urbanization, we developed four sets of high-resolution urban data, including the vegetation coverage, building morphology, land cover, and anthropogenic heat, by using them as inputs for the follow-up urban climate simulation, the simulated urbanization impacts on urban climate would be more accurate.

Since running an atmospheric model consumes a considerable amount of computational resources, especially for simulating long-term climate, we divided the urban climate simulation case into sequenced four-days simulation segments due to limitations in computational resources. For each segment, the first day overlaps with the last day of its previous simulation segment, which was used for model spin-up. For more details, please refer to Section S3 of Supplementary Material.

2.2 The Methodological Framework for Urban Climate Model Evaluation

For urban climate model evaluation, comparing modelled meteorological attributes with their corresponding observed ones is the most widely-accepted way of model evaluation in literature. Given a certain study area and period, such comparisons are carried out respectively for each meteorological variables of interest.

Different views on your data are vital for urban climate model evaluation since meteorological processes contain substantial spatial-temporal patterns and variances. Most existing literature conducted comparisons simply including all observations within their spatial-temporal coverage. Despite that comparing all observations provides an aggregated evaluation of model performance, such a comparison is conducted under the assumption that urban climate behaviours are similar across space and time, which is usually not true. Therefore, we included three different temporal resolutions in our model evaluation framework (Table 1): annual, monthly, and daily. In doing so, our instinct can decide whether the modelled results could replicate the temporal and spatial patterns in the observations or not.

Table 1. An Evaluation Framework for Urban Climate Modelling

<table>
<tr><th colspan="2">Metrics</th><th colspan="3">Temporal Resolution</th></tr>
<tr><th>Statistical Perspectives</th><th>Method</th><th>Annual</th><th>Monthly</th><th>Daily</th></tr>
<tr><td rowspan="2">Descriptive Statistics</td><td rowspan="2">Temporal Comparison of Spatial Variation (TCSV)</td><td>Annual Variation Pattern</td><td>Monthly Variation Pattern</td><td>Diurnal pattern</td></tr>
<tr><td colspan="2">Urban Climatological Spatial Pattern</td><td></td></tr>
</table>

Statistical Distributions	Perkins Skill Score (PSS)	Annual mean PSS	Monthly PSS	
	PDF of the difference between modelled and observed data	Annual mean score	Monthly score	

For each perspective, existing literature commonly compares the descriptive statistics, that is, the range, mean, and variance, between the modelled and observed attributes. The importance of examining climate statistics other than climate means is not new (Katz and Brown 1992; Boer and Lambert 2001). The descriptive statistics are useful in providing aggregated information on the distribution of the attributes, but they can be very misleading since the diverse distributions can lead to similar descriptive statistics, and these aggregated metrics can be sensitive to outliers. Therefore, we compared not only the descriptive statistics but also the statistical distributions of modelled and observed meteorological variables. The probability density function (PDF) was used to calculate the statistical distribution of modelled and observed meteorological variables or the differences between pairs of them. The overlap of two distributions was quantified using the Perkins Skill Score (PSS). The PSS ranges from 0 to 1, while 1 indicates perfect modelling and 0 indicates the worst modelling. The advantages of using PDFs and PSS for climate statistics have been discussed in Perkins et al. (2007).

In urban climatology, the urban-rural difference is among the most important spatial patterns to investigate. Therefore, at the spatial dimension, we evaluated the model by comparing the temporal evolutions of the observed and simulated meteorological characteristics in urban and non-urban areas.

Following the proposed framework, we designed a guideline (Section S4 of Supplementary Material) and a workflow (Figure 1) in the practice of model evaluation.

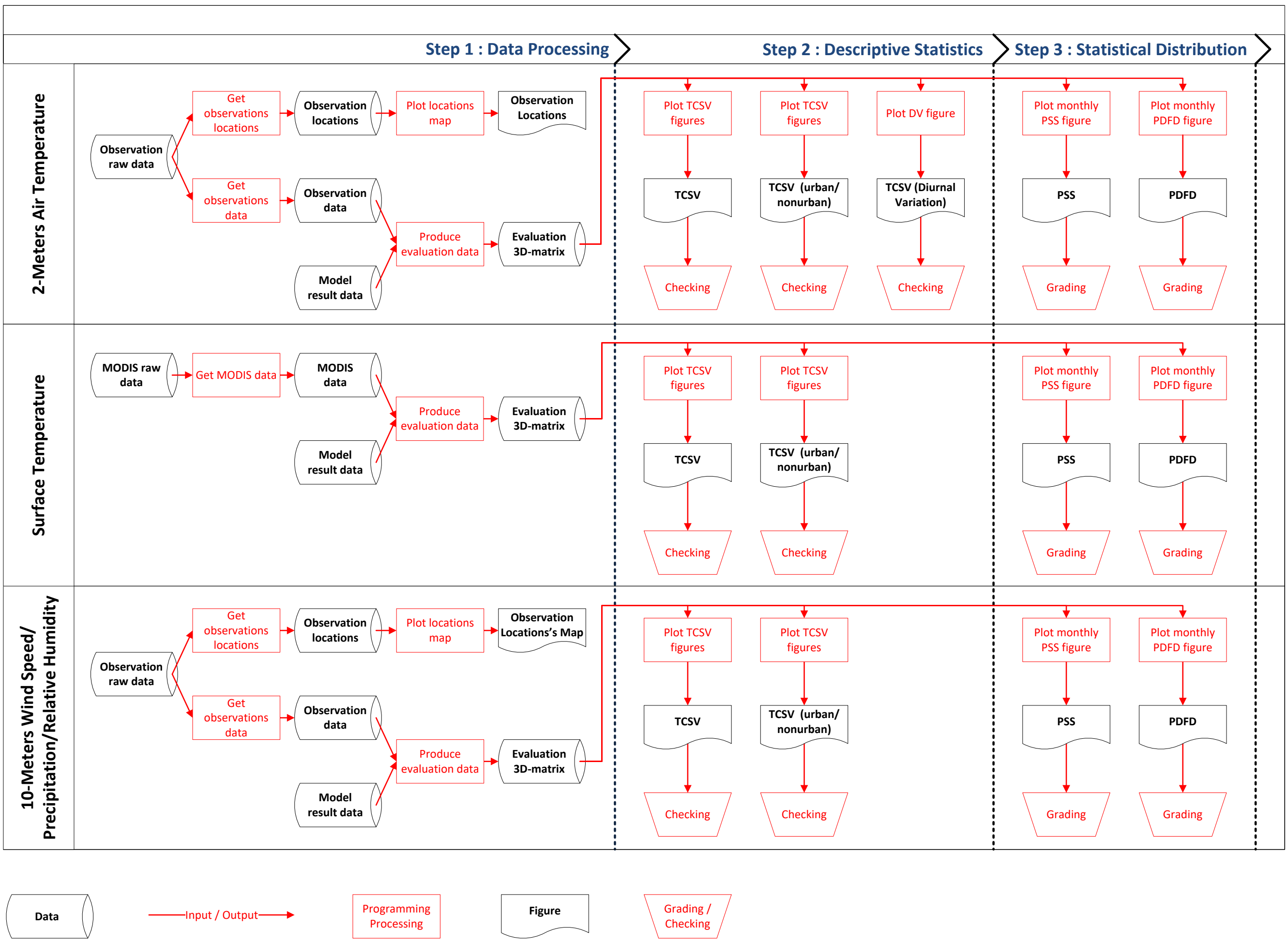

Figure 1: The Workflow for Model Evaluation.

2.3 Observation Datasets and Modelled Variables for Model Evaluation

In existing literature, Numerical Weather Prediction (NWP) models are typically evaluated by comparing the spatial-temporal patterns of the modelled variables with those of its corresponding near-surface observations. Moreover, we selected 7 meteorological variables for the comparison, including 2-meter air temperature, surface temperature, 10-meter wind at u direction, 10-meters wind at v direction, accumulated total cumulus precipitation, accumulated total grid precipitation and 2-meter relative humidity, because these variables are the critical variables in the prognostic and diagnostic equations in the NWP model.

Table 2 lists the modelled variables and their corresponding observations in the model evaluation. The observation datasets are the point data except for the MODIS dataset which is the grid data. All the modelled variables are grid data. The comparisons between modelled variables with its corresponding observed ones are comparisons between the grid value of the modelled variable with its point value matched by to geographical locations, except for the comparison between the modelled surface temperature (TSK) and its corresponding observation retrieved from MODIS imagery. Moreover, a MODIS land surface temperature is a result of the inverse calculation based on the longwave radiation through the atmosphere received by satellite according to the theory of blackbody. A MODIS land surface temperature is a manifestation of the surface synthetic radiation brightness temperature. Furthermore, in the land surface process, TSK is calculated iteratively according to the energy balance which involves longwave radiation, shortwave radiation, sensible heat and latent heat, and accordingly, the final TSK value is also a manifestation of the surface synthetic radiation brightness temperature. Although there are some differences between TSK and the brightness temperatures observed by satellites, they describe relatively similar physical quantities. Therefore, we use TSK to compare with the MODIS land surface temperature.

Table 2: Modelled Variables for Model Evaluation.

Modelled Variables for Model Evaluation		**Corresponding Observation Datasets**	
Name	**Description**	**Datasets**	**Sources**
T2	2-meter air temperature	2010 PRD 2-meter Air Temperature	Meteorological Bureau of Shenzhen Municipality
U10	10-meters wind at U direction	2010 PRD 10-Meters Wind Speed	
V10	10-meters wind at V direction		
RAINC	Accumulated total cumulus precipitation	2010 PRD Precipitation	
RAINNC	Accumulated total grid scale precipitation		
RH2	2-meter relative humidity	2010 PRD Relative Humidity	
TSK	Surface temperature	2010 MODIS/Aqua Land Surface Temperature and Emissivity (LST/E) product	NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center

3 Technical Preparation

3.1 Model Setup

A telescoping nests' structure with four nested domains which are centred at 22˚39′30″ N, 114˚11′30″, was set up as the horizontal domain baseline configuration in this study. Moreover, the same set of eta levels with 51 members was used in each horizontal domain. Furthermore, there were some physics components in the model, and each component had some different schemes for choosing. Table 3 shows the scheme chosen for each component. For more details, please refer to Section S4 of Supplementary Material.

Table 3: Physics Components' Schemes.

Component	Scheme
Cumulus	New Simplified Arakawa-Schubert
Microphysics	WDM5
Radiation	RRTMG
Planetary Boundary Layer	Bougeault–Lacarrere
Surface Layer	Revised MM5
Land Surface Model	Noah LSM
Urban Canopy Model	Single-layer

3.2 Data Preparation

Firstly, the 2010 NCEP FNL (Final) Operational Global Analysis Dataset (1-degree grid spatial resolution and 6-hourly temporal resolution) was used as the Gridded Data in this study. Secondly, the Completed Dataset of WRF Preprocessing System (WPS) Geographical Input Data was used as the Static Geographical Dataset in this study. Thirdly, the 2010 PRD Urban Land Surface Dataset, whose major sets of data include the land cover, vegetation coverage, urban morphology, and anthropogenic heat, which was specially developed for refining the WRF primary data.

3.3 Primary Data Processing

Firstly, the primary data included the interpolated geo-data files, the intermediate format meteorological data files, the horizontally interpolated meteorological data files, the initial condition data files, and the lateral boundary condition data files. Secondly, two primary data processing software packages (geo_data_refinement processing package and wrf_input_refinement processing package) were developed for extracting the urban land surface attributes from the 2010 PRD Urban Land Surface Dataset and revising the corresponding fields of the related primary data files with these attributes.

4 Model Evaluation

4.1 Evaluation of the 2-Meter Air Temperature

Using descriptive statistics, Figure 2 compares the range and median values of the observed and the modelled 2-meter air temperature at 2:00, 8:00, 14:00, and 20:00 in each month of the year. It is evident that the modelled air temperatures always have similar temporal-spatial behaviour with the observed ones. Moreover, Figure 3 compares the diurnal range and median of the observed and modelled air temperature each month of the year. Both the range and median of the 2-meter modelled air temperature have the same diurnal variation with its corresponding observed ones in each month, although there are differences between the modelled ones and the corresponding observed ones. Furthermore, as shown in Figures 4 and 5, the modelled air temperatures have the same urban climatological spatial pattern as the observed ones in which the air temperature is higher in the urban areas than in non-urban areas irrespective of the time at which it is measured.

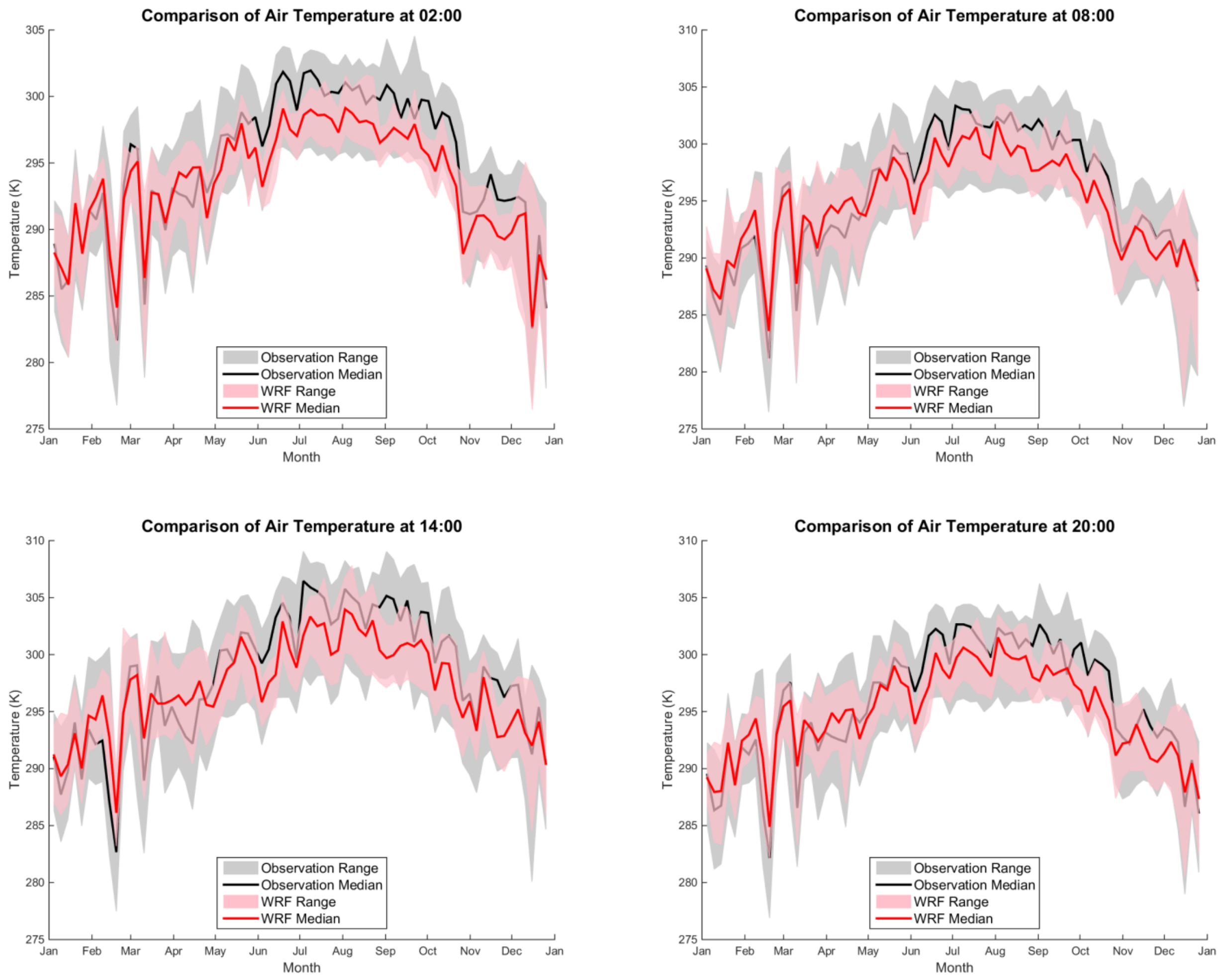

Figure 2: Comparison of Modelled and Observed 2-meter Air Temperature at 2:00, 8:00, 14:00, and 20:00.

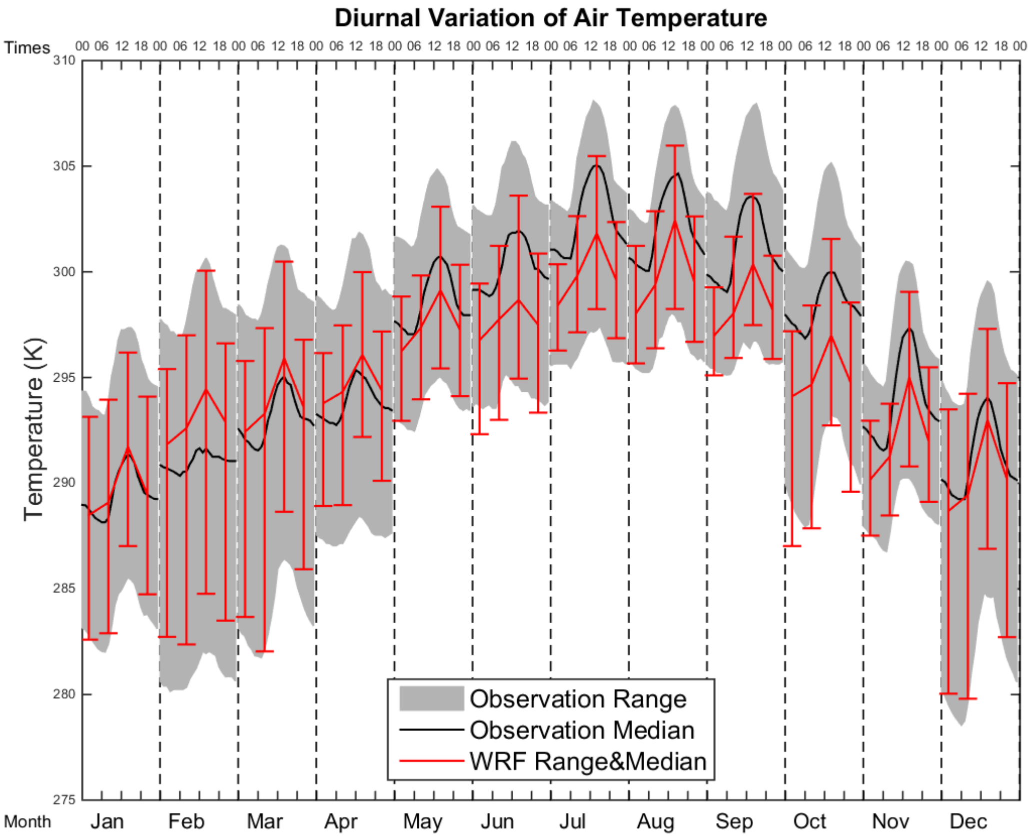

Figure 3: Diurnal Variation of 2-meter Air Temperature.

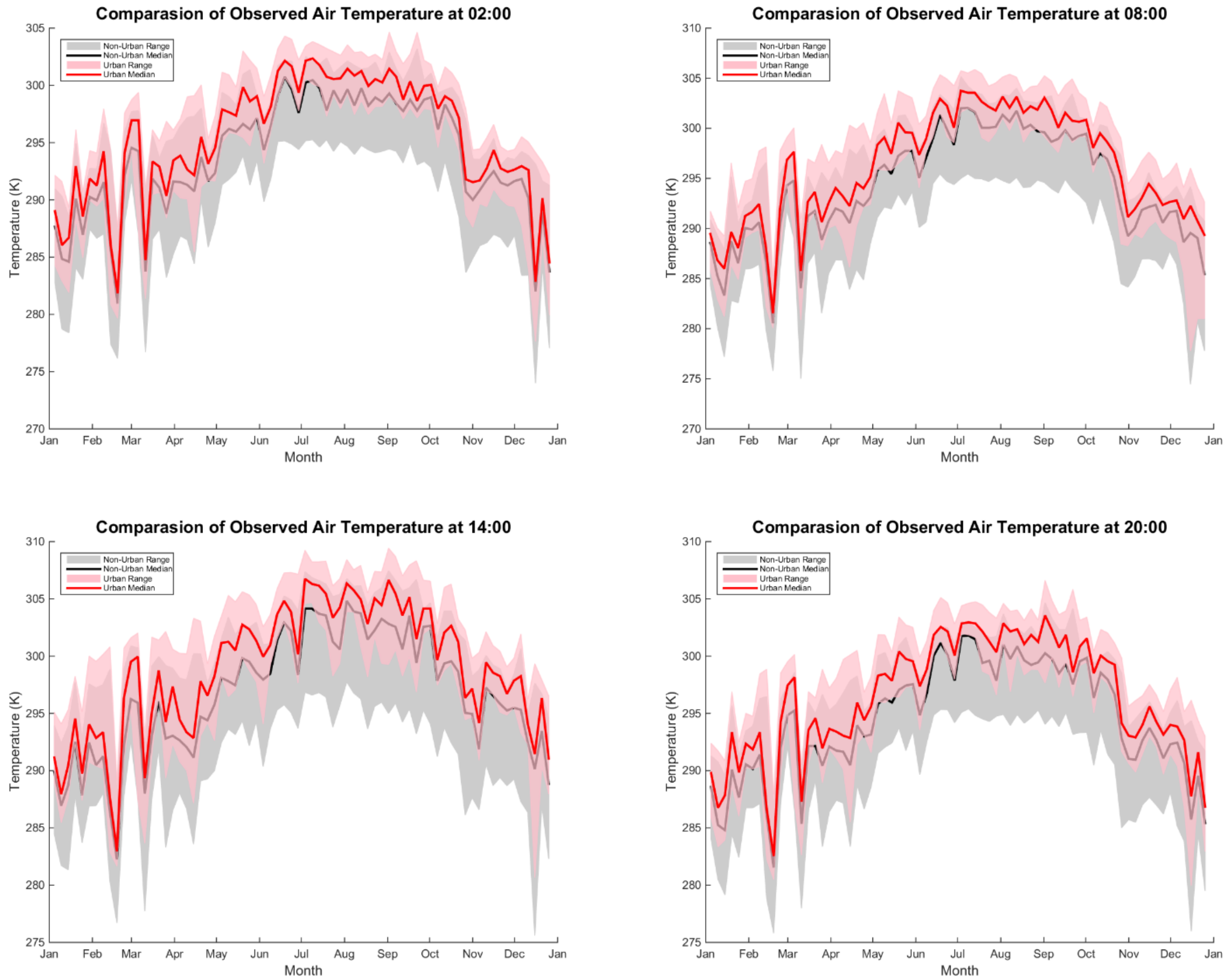

Figure 4: Comparison of Observed Air Temperatures (at 2:00, 8:00, 14:00 and 20:00) in Urban and Non-Urban Areas.

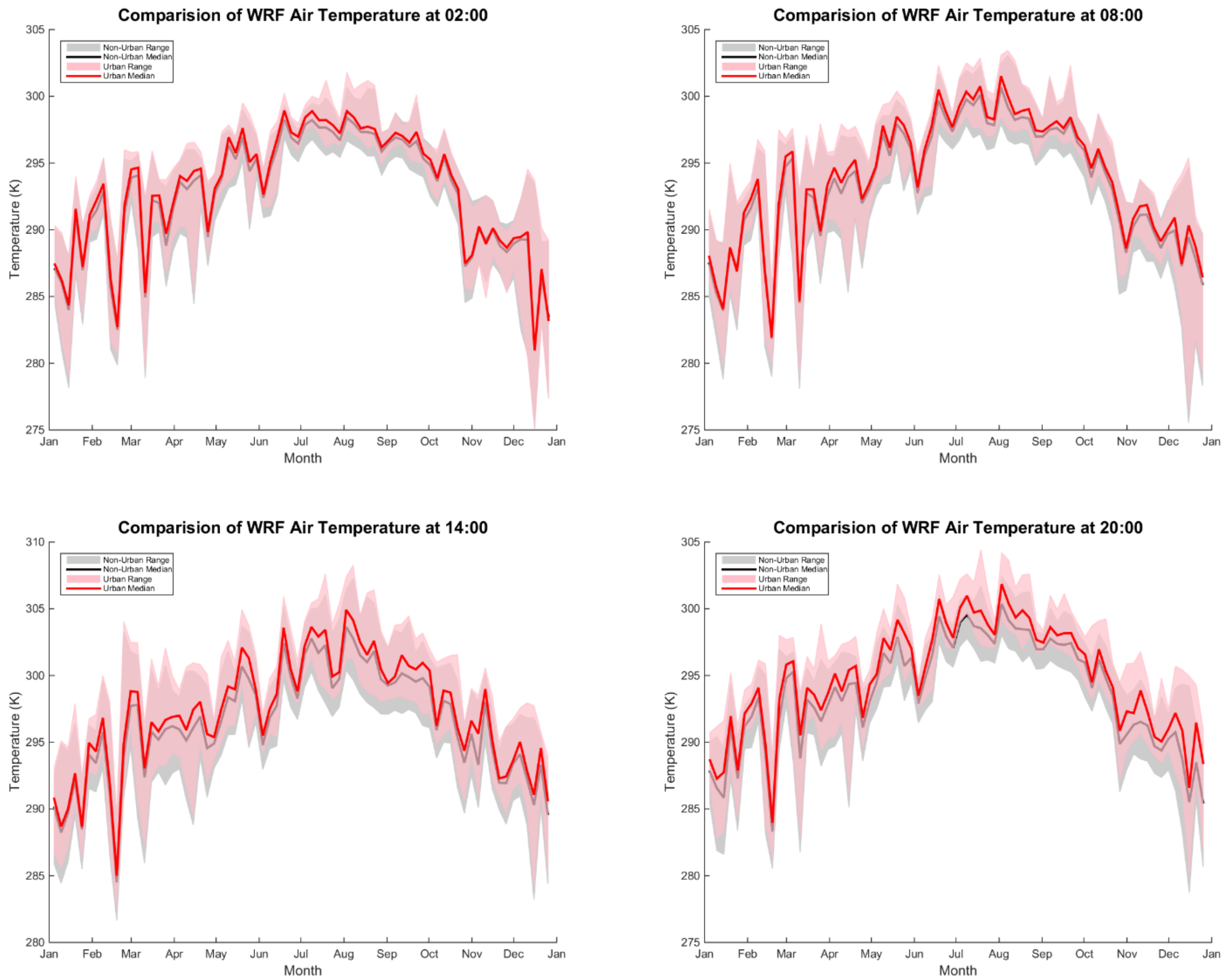

Figure 5: Comparison of Modelled Air Temperatures (at 2:00, 8:00, 14:00 and 20:00) in the Urban and Non-urban Areas.

Using PSS to compare the statistical distribution of the observed and modelled air temperature, the model produces quite a good simulation of 2-meter air temperature with annual mean PSS of 0.724. It also captures the behaviours of monthly and diurnal variation of observed 2-meter air temperatures. Figure 6 shows that the monthly PSS of 2-meter air temperature ranges from a minimum of 0.595 in July to a maximum of 0.886 in January and has an annual mean value of 0.724. This demonstrates that the model captured the PDF for the observed air temperature at least about 60% in a month and over 72% in a year. Figure 7 shows the PDF of differences between each value of each month's time series of modelled grid air temperatures and its corresponding observed ones. The probability of 3-degrees bias interval (the absolute value of the difference between modelled surface temperature and its corresponding observed one is 3 degrees) in a month varies from 64% to 91% and has an annual mean probability of this interval of 78%.

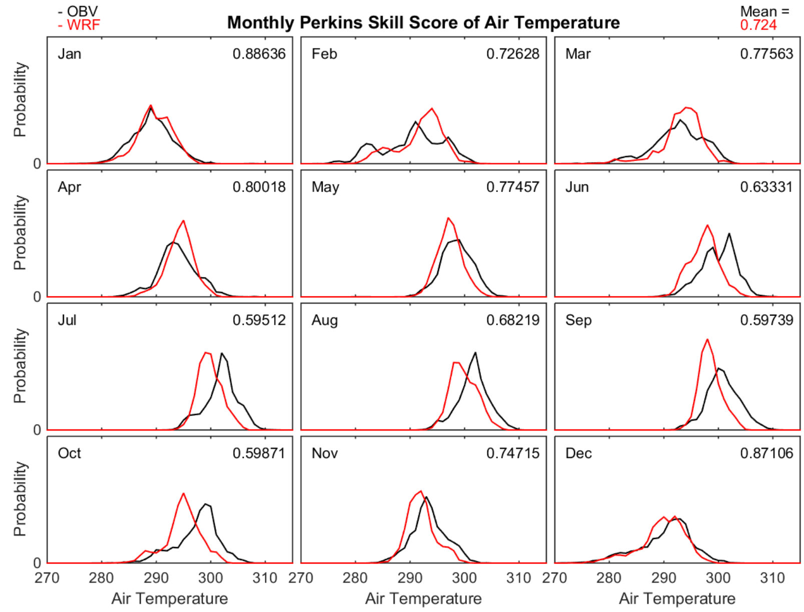

Figure 6: Monthly PSS of 2-meter Air Temperature.

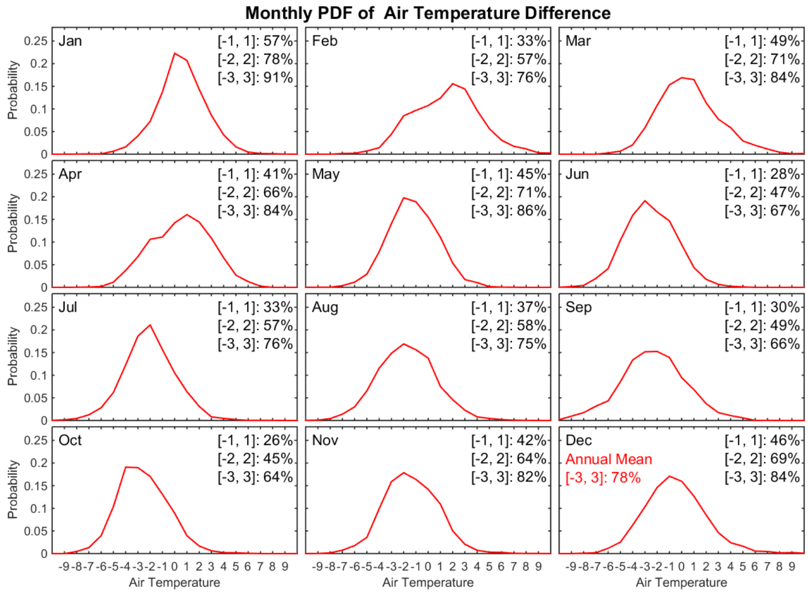

Figure 7: Monthly PDF of 2-meter Air Temperature Difference.

In Figure 6, the modelled distribution shifts to low temperature in the period of June to October (summertime in the research area). Figure 7 shows that the differences between the modelled 2-meter air temperatures and their corresponding observed ones exist in the whole year. In fact, the difference includes not only the modelling bias but also an essential difference between a 1-km grid spatial average value and a value of a point located in this grid. Moreover, the observation always locates in an open area, and thus, the observed 2-meter air temperature is the temperature of a point in the open area. The modelled 2-meter air temperature is a mean temperature of a 1-km grid which always includes some vegetation covered areas. In the summertime, the point air temperature in the open area without coverage of trees is always higher than its corresponding mean air temperature of a 1-km grid with some vegetation coverage.

To sum up, the model produces quite a good simulation of 2-meter air temperature with annual mean PSS of 0.724. It also captures the behaviours of monthly and diurnal variation of observed 2-meter air temperatures. Moreover, the modelled air temperatures have the same urban climatological patterns as that of the observed ones.

4.2 Evaluation of Surface Temperature

From descriptive statistics, the modelled surface temperatures have the same annual variations and the same urban climatological patterns as that of the MODIS ones. Moreover, both the modelled surface temperatures and their corresponding observations from MODIS also have the same pattern, that is, urban areas having higher surface temperatures than non-urban areas in the entire day. For more details, please refer to Figures S6, S7, and Figure S8 in Supplementary Material.

From the statistical distribution, the modelled 2:00 and 14:00 surface temperatures represent the corresponding MODIS ones with an acceptable PSS. The monthly PSS of modelled surface temperatures ranges from 0.629 to 0.794 at 2:00 and from 0.479 to 0.777 for modelled at 14:00 respectively. The annual mean PSS of modelled surface temperatures at 2:00 and 14:00 is 0.702 and 0.623 respectively. Accordingly, both modelled surface temperatures at 2:00 and 14:00 are quite a good expounding in MODIS surface temperature with a PSS of over 0.6. Moreover, the monthly probabilities of 3-degrees bias interval (the absolute value of the difference between modelled surface temperature and its corresponding MODIS one is 3 degrees) at 2:00 ranges between 69% and 98% and has quite a high annual mean value of 87%. The probabilities of a 3-degree bias interval at 14:00 ranges from 54% to 84% and has a high annual mean value of 73%. For more details, please refer to Figures S9, S10, S11 and S12 of Supplementary Material.

However, we also observed noticeable differences between the modelled surface temperature and its corresponding MODIS one is noticeable in some grids. An analysis which was conducted to the MYD11A1 dataset finds that there are many grids whose quality was not evaluated in the MYD11A1 dataset and accordingly, it is highly possible that this difference includes an observation bias. Moreover, due to the difference between the temporal coverages of the model outcome and its corresponding observation from MODIS, the observed difference also includes a bias introduced by the difference in measured time. Furthermore, the resampling operation on the MODIS dataset also causes a technical bias in some grids.

To sum up, the modelled 2:00 and 14:00 surface temperatures represent the corresponding MODIS ones with an acceptable PSS. Moreover, the modelled surface temperatures also have the same annual variations and the same urban climatological spatial patterns as that of the MODIS ones.

4.3 Evaluation of the 10-Meters Wind Speed

Using descriptive statistics, we compared the range and median values of the observed and the modelled 10-meters wind speeds at 8:00, 14:00, 20:00, and 2:00 in each month of the year. It is evident that the modelled 10-meters wind speed always has a similar behaviour of temporal-spatial variation with the observed ones. Moreover, the modelled 10-meters wind speed also has the same urban climatological spatial pattern as the observed ones in which the 10-meters wind speed is lower in the urban areas than in non-urban areas irrespective of the time at which it is measured. For more details, please refer to Figures S13, S14, and S15 of Supplementary Material.

Form the point of view of the statistical distribution, the monthly PSS of the modelled 10-meter wind speed ranges between 0.482 and 0.802 and has an annual mean value of 0.660. Moreover, the monthly probabilities of 3 m/s bias interval (the absolute value of the difference between modelled wind speed and its corresponding observed one is 3 m/s) range between 61% and 83%. For more details, please refer to Figures S16 and S17 of Supplementary Material.

We also observed the deviation which the modelled distribution shifts to high speed. The difference in the speed of modelled 10-meters wind and its corresponding observed one is not entirely caused by the model bias. The observation altitude of the modelled 10-meters wind is different from its corresponding observed one. The modelled outcomes measure the upper air movement of the urban canopy, but the observations measure the air movement inside the canopy. The locations of modelled and observed air movements concerning the canopy would cause an essential difference between the modelled and observed values. Moreover, this difference also includes an essential difference between a 1km-grid spatial average value and a value of a point located in this grid.

Demonstrated by comparisons, the modelled ones of 10-meters wind speed also have the same annual variation and the same urban climatological spatial pattern as that of the observed ones. The model also simulates 10-meters wind speed with acceptable PSS and accuracy.

4.4 Evaluation of Precipitation

From descriptive statistics, the modelled precipitations always have similar behaviour of spatial-temporal variation with the observed ones. For more details, please refer to Figures S18, S19, and S20 of Supplementary Material.

Demonstrated by the statistical distribution, the monthly PSS of modelled precipitation ranges between 0.444 and 0.747 and has an annual mean value of 0.579. Moreover, the model simulated precipitation with an accuracy in which the monthly probabilities of the 3-mm bias interval (the absolute value of the difference between modelled precipitation and its observed one is 3 mm) range between 39% and 89% and have an acceptable annual mean value of 67%. For more details, please refer to Figures S21 and S22 of Supplementary Material.

However, the probability of 3-mm bias intervals is quite low in some months; for example, the one was 39%, 50%, and 53% in June, September, and May respectively. The modelled precipitations deviated from its corresponding observed ones in these three months.

To sum up, the modelled precipitations also have the same annual variation as that of the observed ones. Moreover, the comparison of experiments and observations concerning the modelled and observed measurements of precipitation provide evidence that the model simulates precipitation with an acceptable PSS.

4.5 Evaluation of Relative Humidity

We used the descriptive statistics to compare the range and median values of the observed and modelled air relative humidity values' spatial variation range and median values at 8:00, 14:00, 20:00, and 2:00 in each month. It is apparent that the modelled values always have similar behaviour in spatial-temporal variation with the observed ones, although all modelled median values are higher than the corresponding observed ones. Moreover, the modelled relative humidity has the same urban climatological spatial pattern as the observed one in which the relative humidity is lower in the urban areas than in non-urban areas irrespective of the time at which it is measured. For more details, please refer to Figures S23, S24, and S25 of Supplementary Material.

Demonstrated by the statistical distribution, the monthly PSS of the modelled relative humidity ranges between 0.525 and 0.786 and has an annual mean value of 0.673. Moreover, the model simulates the relative humidity with quite a good accuracy in which the monthly probabilities of the 20% bias interval (the absolute value of the difference between modelled precipitation and its observed one is 20%) range between 77% and 96% and have a high annual mean value of 91%. For more details, please refer to Figures S26 and S27 of Supplementary Material.

To sum up, the model simulates the relative humidity with acceptable PSS and accuracy. Moreover, it also simulates the monthly variation and urban climatological spatial patterns of relative humidity appropriately.

5 Discussions and Conclusions

5.1 Model evaluation using observations

We need more practical model evaluation methods for better comparisons between model outcomes and observations to serve as partial support for the reliability of urban climate simulations and any conclusions based on the simulation results. A model can simulate a resonance with the natural system (Oreskes et al. 1994), and accordingly, a climate simulation should aim at modelling the temporal and spatial meteorological features of climate. Therefore, a model evaluation should aim at assessing the similarity of temporal and spatial features between the modelled results and observations. In this study, the PSS was used for assessing the similarity quantitatively, and the graphic of temporal comparison of spatial variation was used for assessing the similarity qualitatively. The satisfied quality of simulation was evaluated by the acceptable annual values of PSS of modelled variables and the similar behaviours of modelled variables with their corresponding observed ones shown in the graphic of temporal comparison of spatial variation figures of modelled variables.

Utilizing the proposed model evaluation methods, evaluation results in this case study indicate that this atmospheric model appropriately portrayed the annual variations in the climatological patterns of air temperature, surface temperature, 10-meters wind speed, and air relative humidity. We observe that the simulation model captured similar temporal and spatial meteorological features of urban climate. From a quantitative perspective, the model achieved at least an acceptable PSS and accuracy in the simulations of 2-meter air temperature, surface temperature, 10-meters wind speed, precipitation, and air relative humidity, which means that the simulation results are acceptable approximations of the observations. Apparently, according to the above evaluations, the proposed simulation model in our case study is sufficiently reliable in reproducing meteorological features of urban climate at the 1-km spatial resolution.

The good match in our study or any other study, between the model outcomes and observations can only support that the simulation results are acceptable approximations of the observations in the specific spatial-temporal coverages in respective studies. These comparisons are inadequate for model 'verification' or 'validation'. Returning to the philosophical basis, terminologies "verification" and "validation" imply the confirmation of truth and legitimacy respectively (Oreskes et al., 1994). We get observations of meteorological characters from monitoring stations, and that is why the observations come in points and suffer from frequent missing data. Therefore, it is common that the spatial-temporal coverage of the observations can only partially match that of the modelling outcomes, which can be proved by the model evaluation process regarding air temperature, surface temperature and other factors mentioned above. A good match between a model outcome and its corresponding observation at specific locations is no guarantee of a good match at other locations. Similarly, a good match between the model outcomes and the corresponding observations for a historical period is no guarantee of a good match in the future. Moreover, a good match between the model outcomes and corresponding observations for a limited spatial-temporal range does not guarantee that the model is free from initial and uncertainties. Consequently, even a complete match between the observations and model outcomes does not ensure a successful verification and validation of the modelling system, let alone an incomplete match in practice (Oreskes et al., 1994). Theoretically, verifying or validating a model is impossible. However, a model can represent a natural system accurately to some extent (Oreskes et al. 1994), so it is feasible to evaluate model outcomes with the observation data using practical spatial and temporal comparisons, which seems to be the best way possible to evaluate the performance of a model. Nevertheless, we should always be aware of the imperfectness.

5.2 The Essential Difference, Observation bias, and Model bias

The model evaluation is not complete even a complete match between the observations and model outcomes does not ensure a successful verification and validation of the modelling system, let alone an incomplete match in practice (Oreskes et al., 1994). Observations are probably the best reference we get to evaluate the simulation results, but that does not mean observations are perfect for such an evaluation. The comparison between the model outcome and observations alone cannot make a complete model evaluation since it does not rule out the essential difference, observation bias, and model bias.

The essential difference refers to the fact that model outcomes from the simulation models are average values of a grid, while the observations are point-based which only measures the meteorological conditions around the location of the monitoring station. Comparing the average value within a spatial area, the size of which ranges from 0.25 km^2 to over 100 km^2, with point-based observations is problematic for two main reasons. 1) The average value in a grid is calculated under the assumption that the grid is homogeneous, which is usually not true especially when detailed urban morphology is considered, and so the average value is usually lower than that of point-based observations; 2) the point-based observations are likely to be significantly affected by the surrounding environment of the monitoring station, which does not reflect the meteorological condition in the same area. Therefore, the comparison between modelled outcomes and observations has biases, although it is usually the only model evaluation approach we get so far. The only exception is using the observations from remotely sensed imagery; for example, we used the land surface temperature product from MODIS/Aqua to evaluate the modelled temperature of the surface skin. However, there are many grids whose quality have not been evaluated in MODIS/Aqua Land Surface Temperature Product, and accordingly, the difference between the modelled temperature of the surface skin of a grid and its corresponding one in MODIS/Aqua Land Surface Temperature Product includes an observation bias highly possible.

The model bias refers to the uncertainty caused by differences between the actual atmospheric physical processes and the approximations in the model (Skamarock et al., 2005, 2008). The fine-scale details are constructed by a limited area atmospheric model which consists of physical components driven by the lateral boundary conditions of coarse-scale meteorological data and land surface forcing data (Lo et al., 2008; Hong et al., 2014). However, these details do not exist in the coarse-scale meteorological data (Hong et al., 2014). A limited area atmospheric model can represent a natural atmospheric system accurately to some extent rather than entirely. The simulation models are supposed to include many more complex atmospheric physical processes to explain meteorological states with high spatial and temporal resolutions, but many of them have to be omitted or empirically approximated due to limitations in knowledge and computational efficiency. This situation is unusually severe for high-resolution climate simulations. Given the complexity of simulation models, estimating error propagation in these models are difficult, and thus model evaluation becomes the only quality control of simulation results, especially for high-resolution urban climate simulations which are more sensitive to the inadequacies of the atmospheric model, inappropriate configuration of the modelling system (Warner, 2011), and the quality of input data.

5.4 Conclusions

Following the proposed framework, we first measured both the descriptive statistics of the modelled and observed variable and the difference between them at each spatial-temporal epoch for each meteorological variable. Secondly, we respectively analyzed the probability density function (PDF) of modelled and observed variables, the temporal-spatial variation of the modelled and observed variable, and the probability of the difference values between modelled variable and its corresponding observed one. With visualized PDFs, we can understand the empirical distribution of the simulation bias and notice outliers directly if any, which may shed light for further model results' calibrations. Thirdly, we apply the analysis using descriptive statistics and statistical distributions to the other temporal scales: monthly and time-of-day. By doing so, we further investigate temporal variations in different months of the year and times of the day.

In conclusion, we emphasize in this paper that model evaluation is necessary and usually the only process that guarantees the reliability of simulation outcomes, and so utilizing a practical model evaluation process to reach an acceptable agreement between the simulated and observed meteorological variables should be the premise of any conclusion drawn from the modelling results. The emerging high-resolution urban climate simulation models are especially sensitive to possible initial and model uncertainties. In this vein, we proposed a practical methodological framework for urban climate model evaluation that examines not only the matches between the spatial-temporal patterns of the modelled and observed variables but also the statistical distribution of the difference between the modelled variables and their corresponding observations. Moreover, the proposed method utilized PSS to statistically quantifies the extent of overlap between the PDFs of modelled variables and their corresponding observations, which, we argue, was a more informative and useful indicator for the quality of modelling outcomes compared to existing metrics such as residuals and correlations. By doing so, we hope to provide more capable tools that improve the quality control in future researches using numerical meteorological simulations, especially high-resolution urban climate simulations.

We also intend to raise the awareness and attention over model evaluation methods within the modelling community, since new findings without sophisticated understanding, control of model uncertainties and systematic assessments of model outcomes may be scientifically misleading. Moreover, we reminded that the modeller should be cautious about concluding a quantitative finding because it is impossible to identify the essential difference, observation bias, and model bias in the difference between observations and its corresponding modelled results. Furthermore, although this methodological framework of the model evaluation was designed for urban climate simulation, it can also be applied in the local scale climate simulation wherever in urban or non-urban regions.

At the spatial dimension, the areas of climate were classified into the diffident scales, such as local (less than 10^4 km^2), regional (from 10^4 to 10^7 km^2) and global (greater than 10^7 km^2) scales (Intergovernmental Panel on Climate Change, 2012). The similarity of the modelled and observed variables' spatial patterns is indeed a major content in model evaluation of regional and global climate, especially, the spatial difference of precipitation belt and atmospheric circulation. In the previous literature, there were not many papers on the methods of spatial pattern comparison for local climate simulation, which is a research gap for future exploration.

Finally, some future research ideas were inspired. The effects of the selected physical components on the evaluated modelling accuracy is not clear, which requires further control experiments. Also, the effects of the refined urban land surface datasets on the evaluated modelling accuracy also requires further discussions.

Code availability. Information on the availability of source codes used in this study is tabulated below.

<table>
<tr><th>Source codes</th><th>Availability</th></tr>
<tr><td>WRF Model 3.7.1</td><td rowspan="2">These source codes are publicly available at
http://www2.mmm.ucar.edu/wrf/users/download/get_source.html</td></tr>
<tr><td>WRF Pre-Processing System (WPS) 3.7.1</td></tr>
<tr><td>namelist.wps</td><td rowspan="6">These source codes are available upon request from the corresponding author.</td></tr>
<tr><td>namelist.input</td></tr>
<tr><td>Changes in the programs of WRF for inputting
the 2D anthropogenic sensible and latent heat data</td></tr>
<tr><td>geo_data_refinement processing package</td></tr>
<tr><td>wrf_input_refinement processing package</td></tr>
<tr><td>model_evaluation package</td></tr>
</table>

Data availability. Information on the availability of data used in this paper is tabulated below.

Data	Availability

2010 NCEP FNL (Final) Operational Global Analysis Dataset	This dataset is publicly available at https://rda.ucar.edu/datasets/ds083.2/
Completed Dataset of WRF Preprocessing System (WPS) Geographical Input Data	This dataset is publicly available at http://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html
2010 PRD Observation Locations	These datasets are available upon request from the corresponding author.
2010 PRD Urban Land Surface Dataset	
2010 PRD 2-meter Air Temperature	
2010 PRD 10-Meters Wind Speed	
2010 PRD Precipitation	
2010 PRD Relative Humidity	
2010 MODIS/Aqua Land Surface Temperature and Emissivity (LST/E) product	This dataset is publicly available at https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/myd11a1_v006
Modelling Variables for Model Evaluation (T2, TSK, U10, V10, RAINC, RAINNC, RH2, and SWDOWN)	This dataset is available upon request from the corresponding author.

Competing interests. The authors declare that they have no conflict of interest.

References

[1] Bhati, S., & Mohan, M.: WRF model evaluation for the urban heat island assessment under varying land use/land cover and reference site conditions. J. Thero. Appl. Climatol., 126(1-2), 385-400, 2016.

[2] Bruyère, C. L., Done, J. M., Holland, G. J., & Fredrick, S.: Bias corrections of global models for regional climate simulations of high-impact weather. Climate Dynamics, 43(7-8), 1847-1856, 2014.

[3] Caldwell, P., Chin, H. N. S., Bader, D. C., & Bala, G.: Evaluation of a WRF dynamical downscaling simulation over California. Climatic change, 95(3-4), 499-521, 2009.

[4] Dale, V. H.: The relationship between land-use change and climate change. Ecological applications, 7(3), 753-769, 1997.

[5] Department of economic and social affairs, United Nations: World urbanization prospects: The 2014 revision, highlights, United Nations, New York, 32 pp., available at https://esa.un.org/unpd/wup/publications/files/wup2014-highlights.pdf (last access: 14 April 2018), 2014.

[6] Gosling, S. N., McGregor, G. R., & Lowe, J. A.: Climate change and heat-related mortality in six cities Part 2: climate model evaluation and projected impacts from changes in the mean and variability of temperature with climate change. International journal of biometeorology, 53(1), 31-51, 2009.

[7] Gu, Y., & Yim, S. H. L.: The air quality and health impacts of domestic trans-boundary pollution in various regions of China. J. Environ. Int., 97, 117-124, 2016

[8] Hong, S. Y., & Kanamitsu, M.: Dynamical downscaling: fundamental issues from an NWP point of view and recommendations. Asia-Pacific Journal of Atmospheric Sciences, 50(1), 83-104, 2014.

[9] Intergovernmental Panel on Climate Change: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation - Glossary of terms, 2012. Retrieved April, 9, 2015, from http://www.ipcc.ch/pdf/special-reports/srex/SREX-Annex_Glossary.pdf.

[10] Jiang, X., Wiedinmyer, C., Chen, F., Yang, Z. L., & Lo, J. C. F.: Predicted impacts of climate and land use change on surface ozone in the Houston, Texas, area, J. Geophys. Res., 113, D20312, doi:10.1029/2008JD009820, 2008.

[11] Kalnay, E., & Cai, M.: Impact of urbanization and land-use change on climate. Nature, 423(6939), 528, 2003.

[12] Lo, J. C.-F., Z.-L. Yang, and R. A. Pielke Sr.: Assessment of three dynamical climate downscaling methods using the Weather Research and Forecasting (WRF) model, J. Geophys. Res., 113, D09112, doi:10.1029/2007JD009216, 2008.

[13] Meng, W. G., Zhang, Y. X., Li, J. N., Lin, W. S., Dai, G. F., & Li, H. R.: Application of WRF/UCM in the simulation of a heat wave event and urban heat island around Guangzhou city, J. Trop. Meteorol., 17(3), 1006-8775(2011)03-0257-11, doi:10.3969/j.issn.1006-8775.2011.03.007, 2011.

[14] National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce: NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999, Data file, available at https://rda.ucar.edu/datasets/ds083.2/ (last access: 22 March 2016), 2016.

[15] National Center for Atmospheric Research: Completed Dataset and the New Static Data Released with v3.7 of WRF Preprocessing System (WPS) Geographical Input Data, Data file, available at http://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html (last access: 22 March 22, 2016), 2016.

[16] NASA EOSDIS Land Processes DAAC/USGS Earth Resources Observation and Science (EROS) Center: MODIS/Aqua Land Surface Temperature and Emissivity Daily L3 Global 1 km Grid SIN, Data file, available at https://e5eil01.cr.usgs.gov/ops/esir/10960.zip (last access: 13 January 2017), 2012.

[17] Osborn, T. J., & Hulme, M.: Development of a relationship between station and grid-box rainday frequencies for climate model evaluation. Journal of Climate, 10(8), 1885-1908, 1997.

[18] Oreskes, N., Shrader-Frechette, K., & Belitz, K: Verification, validation, and confirmation of numerical models in the earth sciences, Science, 263(5147), 641-646, doi: 10.1126/science.263.5147.641, 1994.

[19] Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. Journal of Geophysical Research: Atmospheres, 118(4), 1716-1733, 2013.

[20] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G.: A description of the advanced research WRF version 2, Mesoscale and Microscale Meteorology Div., National Center for Atmospheric Research, Boulder, Co., United State, 2005.

[21] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., & Barker, D. M.: A description of the Advanced Research WRF version 3, Mesoscale and Microscale Meteorology Div., National Center for Atmospheric Research, Boulder, Co., United State, 2008.

[22] Tewari, M., Chen, F., Kusaka, H., & Miao, S.: Coupled WRF/Unified Noah/urban-canopy modelling system, NCAR, Boulder, 22 pp., 2007.

[23] Wang, J., Huang, B., Fu, D., Atkinson, P. M., & Zhang, X.: Response of urban heat island to future urban expansion over the Beijing–Tianjin–Hebei metropolitan area, J. Appl. Geogr., 70, 26-36, 2016.

[24] Wang, X., Chen, F., Wu, Z., Zhang, M., Tewari, M., Guenther, A., & Wiedinmyer, C.: Impacts of weather conditions modified by urban expansion on surface ozone: comparison between the Pearl River Delta and Yangtze River Delta regions, J. Adv. Atmos. Sci., 26(5), 962-972, 2009.

[25] Warner, T. T.: Quality assurance in atmospheric modelling. B. Am. Meteorol. Soc., 92(12), 1601-1610, 2011.

[26] Yang, B., Zhang, Y., & Qian, Y.: Simulation of urban climate with high-resolution WRF model: A case study in Nanjing, China, Asia-Pacific J. Atmos. Sci., 48(3), 227-241, doi:10.1007/s13143-012-0023-5, 2012.