Environmental Modelling & Software 79 (2016) 267–284

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Metrics to identify meaningful downscaling skill in WRF simulations of intense rainfall events

Marie Ekström

Land and Water CSIRO, Australia

ARTICLE INFO

Article history:
Received 28 April 2015
Received in revised form
26 November 2015
Accepted 30 January 2016
Available online xxx

Keywords:
Rainfall simulations
Fine resolution
Climate change
Weather research and forecasting model
Microphysics scheme

ABSTRACT

Dynamical downscaling attempts to provide regional detail to climate change projections that subsequently can be used as input to climate change impact models. However, unlike forecasts by numerical weather prediction models, downscaled projections cannot be tested for skill because the future of interest is decades away. Nevertheless, models can be tested in terms of how well they simulate current weather or climate, thus giving an indication of skill in representing the process of interest. Here, six configurations using different combinations of three microphysics and two planetary boundary layer schemes are assessed on their skill to simulate desired characteristics in daily rainfall fields from three two week simulations in southeast Australia; 'desired' meaning desirable in relation to the intended application. Of different metrics and analysis assessed, a metric based on variography analysis, summarising characteristics about spatial variability and dissimilarity, is shown to provide the most informative guidance relative to the desirable characteristics.

1. Introduction

Forecast verification is a core activity for numerical weather prediction, providing information on model skill when simulating weather into the future. As weather eventuates, forecasts can be compared against the recorded weather for a particular time and location. In a climate change context, the importance of accuracy in terms of timing and geographical precision is relaxed, as simulations are representations of plausible climates; simulating weather far beyond any predictive signal in the internal weather system. Skill in regional climate change projection is typically assessed by comparing simulated climatologies with observed ones for multi-decadal periods (Fowler et al., 2005). For methods such as dynamical downscaling (the use of a dynamical model to add regional detail to global climate model projections), long simulation times are associated with large computing costs. Hence, conducting multi-decadal model runs for the purpose of assessing skill of a particular model setup (e.g. the use of different parameter schemes) can be prohibitively expensive from a computing resource point a view.

Nevertheless, assessments of skill in methods used for deriving regional projections are desirable, as they can inform the level of confidence attributed to the simulated future climate. But what verification methods are most appropriate in a climate change context? In a NWP context, the ability to capture timing and extent of an event is central to a skilful simulation; hence metrics evaluated on co-locations are meaningful. This is not necessarily true for models used in a climate change context, where characteristics such as spatial dependence or full distributional representation might be more relevant. Further, the type or manifestation of model skill required by a researcher can differ depending on the intended application and should be reflected in the choice of model verification metrics. Here, different metrics and analyses are used to examine the performance of the more complex microphysics schemes available for the *Weather Research and Forecasting* (WRF) model (Skamarock and Klemp, 2008). The underlying motivation being an intent to identify the model configuration best suited for research on water resource planning under climate change for southeast Australia.

The meso-scale numerical model WRF hosted at the United State's National Centre for Atmospheric Research (NCAR) supports a wide range of modelling applications within the weather and climate community (Caldwell et al., 2009; Chotamonsak et al., 2011; Coniglio et al., 2013; Del Genio et al., 2012; Done et al., 2004; Heikkila et al., 2011; Kain et al., 2006; Leung et al., 2006; Ma et al., 2012). To accommodate its many uses it has a flexible structure that allows users to select physics and dynamics settings

E-mail address: Marie.Ekstrom@csiro.au.

http://dx.doi.org/10.1016/j.envsoft.2016.01.012
1364-8152/

that optimise the model for their particular needs. Selecting parameter schemes and other settings is not necessarily straightforward when multiple theoretically comparable schemes are available to the user. Though numerous assessments of schemes and settings exist in the literature (Evans et al., 2012; Jankov et al., 2005; Liu et al., 2012), some testing is sometimes required to assess performance in a particular geographical area for which there is limited advice to be drawn from the literature.

The complex microphysics schemes become particularly relevant when the researcher seeks to simulate rainfall at a fine resolution (~4 km), when the resolution is such that convection (at the grid scale resolution) is explicitly simulated by the model rather than parameterised (Kain et al., 2006). In a climate change context, recent research indicate that very fine (~1.5 km) convective-allowing (or permitting) simulations could be particularly important, as these models have the ability to provide a more realistic simulation of hazardous high intensity rainfall events (Kendon et al., 2012) and thus theoretically an improved understanding of plausible impacts to extreme rainfall events under changing greenhouse concentrations (Kendon et al., 2014; Westra et al., 2014).

However, very fine resolution regional climate models are computationally very expensive in terms of processing (increased iterations) and storage (volume of output), as climate simulations require significantly longer simulations than NWP models (that simulate for temporal domains bounded by days rather than decades) to allow detection of a change in the climate signal. Thus, with finite computing resources, researchers conducting downscaling with meso-scale dynamical models have to make important decisions around the extent and the resolution of the spatial and temporal model domain of their experiment. This is particularly relevant if output is intended to inform on policy, since the researchers should balance the potential added value of realism (by increasing the resolution of an experiment) against adequately representing the typically large uncertainty in projected rainfall stemming from lack of knowledge about future emissions, and inadequacies in simulating the global climate response to these emissions (see discussion in Ekström et al. (2015) in their appraisal of downscaling methods). In short, increasing the model resolution may hamper ability to sample the climate signal contained in an ensemble of global climate models (i.e. conducting downscaling only on a small sample of global models).

If limitations in computing resources exist, it is sensible to first test whether very fine convection-allowing configurations can add value relative to the intended application. Whilst recent experiments indicate that this is indeed the case for extreme rainfall impact assessments, it is not immediately clear that the finer resolution experiments add value for impact research in a water resource application; where the scale relevant to the topic is greater both in time (seasonal to annual rainfall) and scale (typically rainfall across multiple catchments) compared to the scale relevant to represent individual rainfall events generating extreme rainfall.

To assess whether fine resolution experiments are application appropriate for water resource impact research in southeast Australia, there is an interest in using WRF to conduct a multi-year simulation to assess relative differences in parameterised versus explicitly resolved convection. Given the multiple configuration options available, an assessment of physics parameter scheme options is desirable to ensure that the most appropriate configuration is used to conduct a multi-year simulation for current climate (a long simulation period being required to derive robust estimates about average climate conditions, so called climatologies). Of course, crucial to the assessment is the definition of 'appropriate'. What skills or characteristics are desired of the model and what metrics can be used to quantify these skills to enable a performance ranking of differently configured models.

This paper demonstrates learnings from a case study in southeast Australia, where application relevant combinations of WRF physics scheme combinations are assessed on their ability to capture gross spatial, temporal and distributional characteristics desired from rainfall fields intended for impact work in the water resource domain.

2. Methods and data

2.1. WRF setup

The simulations presented here are generated using WRF version 3.6.1 with the Advanced Research WRF (ARW) dynamical core. A one-way telescopic nest with 3 spatial domains using a Lambert conformal projection is used. The outer nest (D01) cover the Australian continent into the Southern Ocean with a 50 km resolution, the intermediate domain (D02) focus on southeast Australia and coastal waters with a 10 km resolution, and the innermost domain (D03) include the southernmost part of the Great Dividing Range and its western slopes at a spatial resolution of 2 km (Fig. 1a and b). The model has 50 vertical levels in the atmosphere and tops out at 10 mb. Spectral nudging is applied to larger scale features in the wind and geopotential fields in the upper atmospheric layers of the outer model domain (D01) only.

As noted earlier, the WRF model is highly configurable and allows the user to select schemes appropriate to their intended application. Here, physics schemes were selected based on their appropriateness for simulations in a climate change context and for representation rainfall on a 2 km resolution (allowing for explicit simulation of convective rainfall). Guidance on parameter selection are given in the ARW user's guide (NCAR, 2013) and from peer review literature. For this application a pertinent study is Evans et al. (2012), who tested 36 different physics schemes combinations for the purpose of dynamical downscaling. A detailed motivation for the selection of physics parameter schemes used here is available in Ekstrom (2014), hence only a brief summary of options and key motivation for their selection is given below.

The rapid radiative transfer model for global applications (RRTMG) (Iacono et al., 2008) was used for the radiation schemes of the long and short wave spectra; this scheme is recommended for a 1–4 km resolution case in the ARW user's guide and allows for temporally varying greenhouse gas concentrations relevant for a future climate change implementation. The mixing of surface heat and moisture fluxes into and onwards within the boundary layer is governed by the land surface model, the surface physics scheme and the planetary boundary layer (PBL) scheme. With respect to convective permitting experiments Coniglio et al. (2013) compared three 'local' schemes and two 'non-local' PBL schemes, where local refers to closure schemes that consider only adjacent fields when solving equations for unknown variables in estimating the vertical mixing. Low biases are reported for the local Mellor-Yamada Nakanishi and Niino (MYNN) (Nakanishi and Niino, 2006) Level 2.5 scheme, and hence this scheme was selected for the WRF setup. The PBL scheme has strong influence on rainfall simulations in this region (Evans et al., 2012). For this reason, the non-local Yonsei University (YSU) scheme (Hong et al., 2006) was included to represent uncertainty in using conceptually different methods; a scheme that has been applied with success in a southeast Australian context (Evans et al., 2012). The PBL schemes were used in combination with the MM5 surface physics scheme and an intermediate complexity land surface model (Noah LSM).

Even though convection is explicitly resolved in D03, a convective parameter scheme is required to represent rainfall generated by convection on sub-grid cell scales in D01 and D02. In

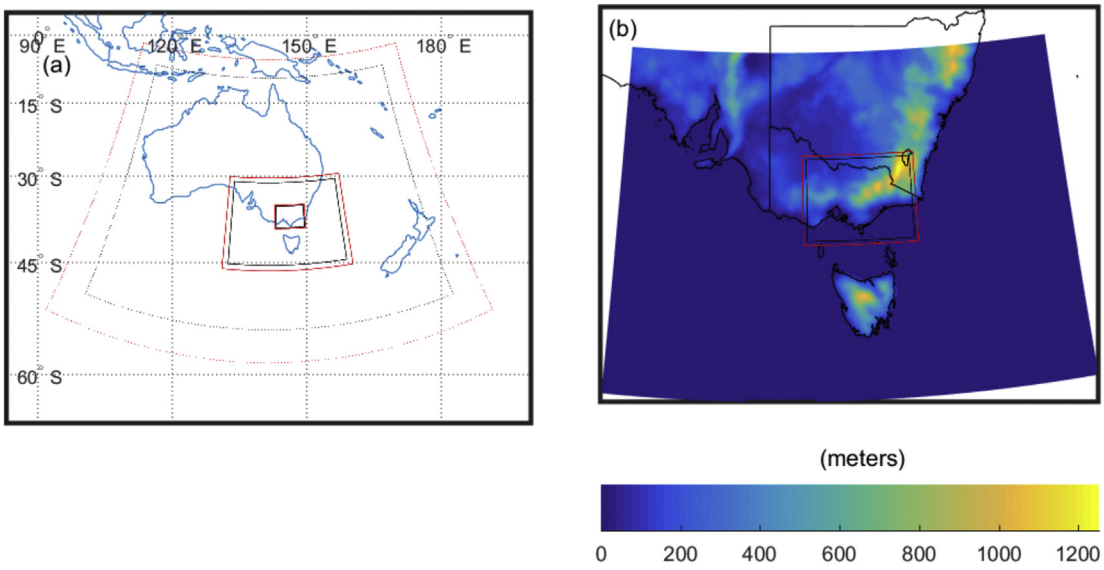

Fig. 1. In panel a): the spatial dimensions of the outer domain D01 at 50 km resolution, the intermediate domain at 10 km resolution and the innermost convective permitting resolution domain (at 2 km resolution). The red markers denote the native model domain and the black markers indicate the model domain after the relaxation zone of 10 grid cells is removed. In panel b): the topography (metres) of D02 with state boundaries and boundaries of D03 overlayed (red and black lines denote the native and native minus relaxation zones). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Evans et al. (2012) two cumulus schemes were tested, the Kain-Fritch and the Betts-Miller-Janjic (BMJ) scheme. The former gave consistently poor results when combined with PBL scheme YSU and radiation scheme RRTMG. For this reason, the Betts-Miller-Janjic (BMJ) scheme was used. The BMJ scheme is an 'adjustment type' whereby values are relaxed towards a post-convective (mixed) sounding (Janjic, 2000, 1994).

The microphysics (MP) scheme is responsible for heat and moisture flux within the atmosphere and gives the surface resolved rainfall. Initially all five double moment schemes available for WRV 3.6.1 were considered. However, two schemes repeatedly failed when run in combination with physics options above, hence only three schemes remained for further analysis, these are: the WRF double moment 6-class scheme (WDM6) (Lim and Hong, 2010), the Thompson scheme (Thompson et al., 2008), and the Milbrandt scheme (Milbrandt and Yau, 2005). With two PBL schemes in combination with 3 microphysics schemes, a total of six physics scheme combinations are compared here (Table 1).

2.2. Boundary and verification data

The WRF simulations were run using six hourly inputs from the re-analysis data set ERA Interim (Dee et al., 2011) using climate information from: the surface, 37 pressure levels, and 4 sub-surface levels. All fields having a spatial resolution of approximately 80 km. ERA Interim data are assimilated into WRF simulations along the lateral and lower boundaries of the outer domain (D01) with a lateral relaxation zone of 10 grid cells. This re-analysis product has previously been used in this region in a downscaling context to investigate performance of WRF to simulate the broader climate (Evans et al., 2012) and rainfall characteristics in conjunction with the East Coast Lows (Gilmore et al., 2015; Ji et al., 2014). Digital elevation data from the 9 s digital elevation model of the Geoscience Australia and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) of Australia was used instead of the topography information provided with WRF for domain D02 and D03. The 9 s data set was smoothed to reduce impact by steep relief on the simulations using WRF's pre-processing system (WPS).

Due to the particular interest in the influence of the microphysics scheme on fine resolution rainfall, all verification analysis of simulated data are conducted on the inner domain (D03, 2 km resolution). The skill of the two coarser domains (D01 and D02, 50 and 10 km resolution respectively) are to some extent implicitly accounted for in this analysis through the nested model structure, i.e. poor skill in D01/D02 would likely lead to poor skill in D03.

Model simulations for D03 are assessed against 5 km gridded observed daily rainfall totals from the Australian Water Availability Project (AWAP) (Jones et al., 2009) produced by the Australian Bureau of Meteorology, and atmospheric sounding data of water vapour mixing ratio (WVMR, g/kg) and temperature (°C) from the Wyoming Weather Web station 94866 YMM (37.66°S, 144.85°E) at Melbourne Airport at 00 UTC (corresponding to 10 am Eastern Standard Time (EST) in winter, and 11 am Eastern Daylight Time (EDT) in summer).

For comparison with the coarser resolution AWAP data, the output fields for D03 were re-interpolated onto the regular latitude longitude coordinates of AWAP using routines from the Earth System Modelling Framework (ESMF)[1]; the re-interpolated area framed by coordinates 35.55°–39° S and 143.20°–149.30° E at 0.05° resolution. Note that whilst AWAP is used as a representation of reality, it is a smoothed gridded product with its own associated uncertainty. For the state of Victoria, error estimates based on cross-validation of daily/monthly rainfall are typically less than 5/

Table 1
List of WRF physics scheme options for each ensemble member (N1–N6).

Nb	PBL	MP	Surf_phys	RA sw/lw	LSM	CU D01/D02
1	MYNN	WDM6	MM5	RRTMG	Noah	BMJ
2	MYNN	Thompson				
3	MYNN	Milbrandt				
4	YSU	WDM6				
5	YSU	Thompson				
6	YSU	Milbrandt				

[1] https://www.earthsystemcog.org/projects/esmf/.

25 mm, as estimated over the period 2001 to 2007 (Jones et al., 2009). Hence biases between simulated and observed data will comprise elements of uncertainty of both data sets.

2.3. Case study periods

Three 15 day case study periods were selected for testing the selected WRF configurations. These represent different synoptic conditions associated with events of intense rainfall during the 2010–11 period of extended flooding in the state of Victoria, and seasonal variability: cool season (April–October), warm season (November to March), and the shoulder period between the two seasons. For all simulated periods, the first day was discarded as spin-up.

During the cold season case study (8th to 21st of August 2010), rainfall occurs in conjunction with an upper level trough and low level cold front associated with a low pressure system. This system develops on the 10th of August over Victoria and subsequently moves westward over the next few days. Further passages of cold fronts occur during the period 15th–17th and again on the 19th–20th of August. The shoulder season case study (6th to 19th October 2010) include rainfall from passages of cold fronts on the 7th and an upper level trough on the 13th followed by intense rainfall associated with a deep low centred over Victoria on the 15th and 16th. The warm season case study (31st January to 13th February 2011) include the passage of tropical cyclone Yasi far north of the study area; its northern passage enabling advection of a moist tropical air mass ahead of a westerly approaching prefrontal trough.

2.3.1. Selected case study days

For all but the temporal assessment of daily rainfall amounts, two non-consecutive days per case study period simulation were selected for the verification exercise (see section 3.1); the selected days being those with the largest recorded rainfall totals. The sub-selection of days was made to maximise opportunity to discriminate between simulations, as the influence of using different physics options tend to be greater for heavy rainfall events (Evans et al., 2012).

These days are referred to as Day 1–6, where Day 1 and 2 are day 4 and 12 of C1 (11th and 19th of August 2010); Day 3 and 4 are day 2 and 11 of C2 (7th and 16th of October 2010); and Day 5 and 6 are day 6 and 12 of C3 (5th and 11th of February 2011). Their synoptic characteristics are displayed in Fig. 2 showing a low pressure centre located immediately south of domain D03 for Day 1; a cold front passage on Day 2; Day 3 and 4 experience strong westerly and southwesterly winds following the passage of a cold front; and an upper level trough with southerly winds to the west and northerly winds to the east on Day 5 and 6 (noting the weakened Tropical storm Yasi over the centre of Australia on Day 5).

2.4. Skill metrics and analyses

The measure of skill needs to be relevant to those characteristic that are desired of the simulated rainfall field, bearing in mind what can be expected reasonably in terms of performance (given model structure and quality and frequency of input data). The intended application for the WRF set-ups presented here is that of dynamical downscaling, where rainfall fields are of particular interest.

Because the model simulates its own climate guided by 'reality' only at its outermost domain (and the initial starting conditions in all three domains), high accuracy in the spatial distribution of rainfall is perhaps only reasonable to expect when caused by large scale, well defined, synoptic features such as frontal passages, rather than convective rainfall (noting that in a climate change simulation, the information by a global climate model is likely to be provided on a coarser spatial and temporal resolution compared to that of ERA Interim). For this reason, the metrics relevant to this assessment should attempt to quantify skill that relate to general characteristics of the rainfall event, such as: the total magnitude of rainfall associated with the rainfall event in the simulate region, ability to capture the full spectra of observed rainfall intensities at the simulated scale, realistic simulation of the typical scale of rainfall (the spatial extent of rainfall events). Of lesser importance is skill measured by metrics based on continuous and categorical statistics, as these are derived from space and time corresponding coordinates. In this aspect we are accepting less accuracy than verification exercises in a NWP forecasting context.

Unlike spatially continuous fields such as temperature, rainfall is event based; its characteristic different in time and space depending on the process responsible for genesis. For this reason standard methods based on continuous (e.g. root-mean-square error (RMSE), spatial correlation) and categorical statistics (metrics based on scores of hits/misses/false alarm/correct rejects in contingency tables), described in depth in by Wilks (2006) are accompanied by more elaborate methods that consider other aspects, such as spatial displacement, magnitude and orientation errors, and examining relative skill at different scales. A comprehensive review and examples of these methods are found in Gilleland et al. (2009, 2010), who suggest a broad categorisation of these methods into four types: 1. Scale separation/decomposition, 2. Neighbourhood/fuzzy, 3. Features/object based and 4. Field deformation/morphing.

Scale separation or decomposition approaches typically involve some form of filtering to enable assessment on different spectral scales (Casati, 2010), neighbourhood or fuzzy methods apply smoothing filters to fields, comparing the 'footprint' of a rainfall event rather than grid-cell specific comparison (Roberts and Lean, 2008). Feature or object based methods focus on particular attributes of rainfall fields (e.g. in displacement, orientation or size) (Li et al., 2015; Ebert and Gallus, 2009), and field deformation or morphing methods derive distortion vectors that allow field wide manipulation of the simulated field to resemble (morph into) observed characteristics (Gilleland et al., 2009). All approaches provide information on skill, though the aspect of skill varies amongst methods.

For this verification exercise, a neighbourhood method is selected as a meaningful metric (see section 2.4.1), noting that it provides information on models that "… may not pinpoint the exact location of each storm cell, but they can correctly place the overall envelope of precipitation". (Gilleland et al., 2010: p.1369). It is however noted that this type of approach is not able to give information on skill about structural errors in the rainfall field. For this reason a variography analysis is conducted, by which spatial characteristics of rainfall fields in terms of dissimilarity scale and variance are captured in the parameters of the variogram model (Lepioufle et al., 2012; Emmanuel et al., 2012) (see section 4.2.2). Other approaches to assess spatial characteristics are possible, such as assessment of the spatial cross-correlation relationship (Burton et al., 2013). The techniques of geostatistics used for the variography analysis are however attractive in that the theory is well established and multiple software and open source code offer robust model fitting (Deutsch and Journel, 1998; Chiles and Delfiner, 2012; Isaaks and Srivastava, 1989).

Temporal agreement in daily rainfall occurrence and amount is further used as an indication of skill in the overall simulation of the synoptic circulation, whilst quantile–quantile plots give an indication of ability to represent observed daily rainfall totals. Further, different characteristics of WRF ensemble members are gleaned from comparing vertical profiles of simulated water vapour mixed

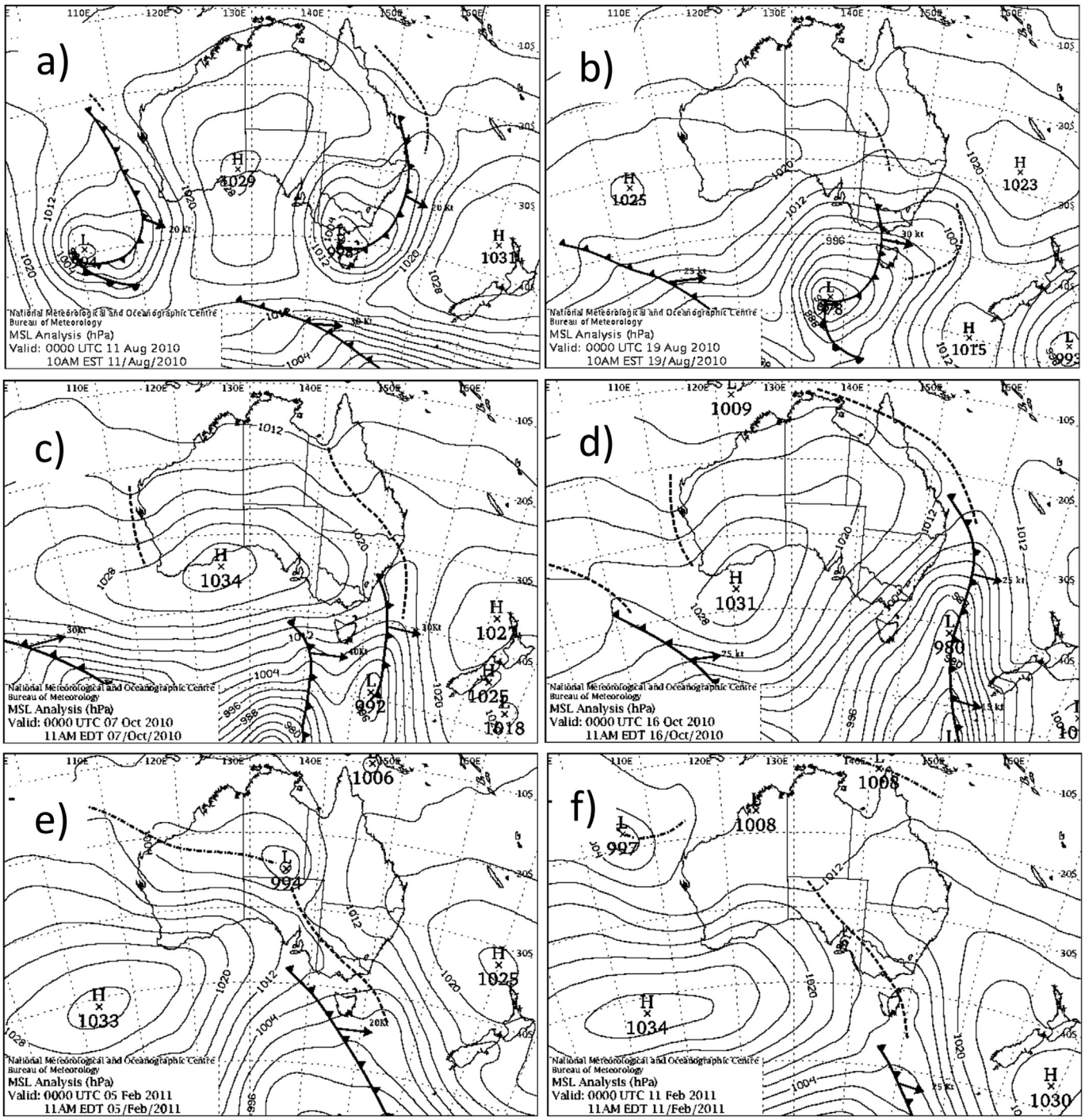

Fig. 2. Mean sea level pressure charts for the 11th and 19th of August 2010 (a and b), 7th and 16th of October 2010 (c and d), and the 5th and 11th of February 2011. The maps are sourced from the online Weather Map archive of the Australian Bureau of Meteorology.[21]

ratio and temperature to sounding measurements from Melbourne airport and horizontal plots of temperature and moisture fields at near surface pressure levels.

2.4.1. Neighbourhood analysis

Neighbourhood (or fuzzy analysis) lends itself well to verification of fine resolution experiments by reducing the impact of double penalty in the analysis, i.e. the instances where a forecast is penalised because it forecasted rain where none occurred and further did not forecast rain where it occurred and thus penalised twice (Ebert, 2008). Rather than expecting skill on grid cell level, the neighbourhood method expects skill across a predefined 'neighbourhood' or window. The metric used here is that of the Fractions Skill Score (FSS) (Roberts and Lean, 2008; Mittermaier et al., 2013), where forecasted and observed rainfall fields are first translated to binary fields with values of 1 denoting grid cells with

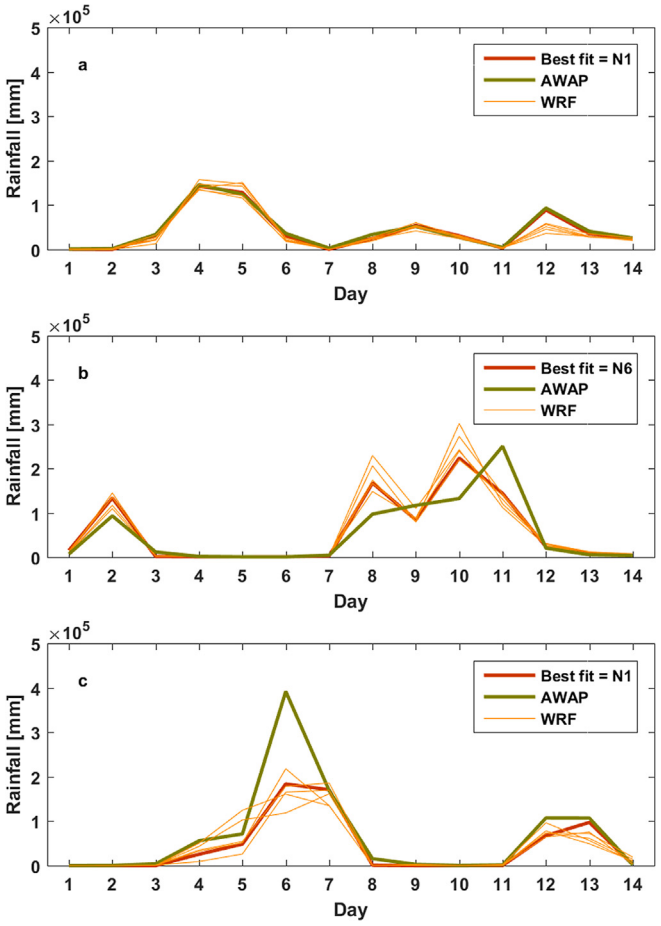

Fig. 3. Rainfall totals (mm/day) within domain D03 for AWAP (green) and WRF simulations (orange). The best fit (MAE) is shown in red and identified in top right legend. The panels show case study 1 (a), case study 2 (b) and case study 3 (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

estimated from the mean square error (MSE) of the observed (O) and modelled (M) fractions from a particular neighbourhood (n) (eq. (1)), a reference value $MSE_{(n)ref}$ (eq. (2), best possible MSE) and the MSE for a perfect forecast $MSE_{(n)prefect} = 0$:

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[O_{(n)i,j} - M_{(n)i,j} \right]^2 \quad (1)$$

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O_{(n)i,j}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} M_{(n)i,j}^2 \right] \quad (2)$$

$$FSS_{(n)} = \frac{MSE_{(n)} - MSE_{(n)ref}}{MSE_{(n)perfect} - MSE_{(n)ref}} = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}} \quad (3)$$

2.4.2. Variography analysis

The experimental variogram describes the dissimilarity between data points as a function of distance. Or more specifically, the expected squared difference between values of the intrinsic random function $Z(x)$ at separation distance (or lag) h:

$$2\gamma(h) = E\left\{ [Z(x) - Z(x+h)]^2 \right\} \quad (4)$$

The semi-variogram, is half that of the expected squared difference and provides a measure of variability that increases as data points become more dissimilar.

The semi-variogram construct holds true if the random function $Z(x)$ is intrinsically stationary, i.e. that the expectation of $[Z(x)-Z(x+h)]^2$ is constant across the assessed domain. Here, the random function is that of daily rainfall totals and the spatial domain the geographical extent of D03. It is reasonable to expect that condition of intrinsic stationarity is violated as one might expect the dissimilarity in daily rainfall totals across D03 to depend not only on distance but also on geographical location (e.g. across the Great Dividing range relative to other areas) and rainfall type (e.g. convective versus stratiform rainfall). However, whilst intrinsic stationarity is crucial if the semi-variogram is to inform further analysis across the entire domain (such as mapping), it is less relevant here as the semi-variogram is used to capture average spatial dissimilarity or simply put, give a 'spatial signature' encapsulating the characteristics of the considered data set.

Well-established geostatistical techniques exist for calculating

Table 2
For each case study and domain, the mean absolute error (MAE) and root square mean error (RMSE) of daily rainfall total per simulation (N1–N6). Metrics are based on daily rainfall totals accumulated for the common spatial domain (35.55°–39° S and 143.20°–149.30° E). Yellow highlight indicate the lowest value, and red highlight indicate second lowest value.

Case	Error metric	N1 (mm)	N2 (mm)	N3 (mm)	N4 (mm)	N5 (mm)	N6 (mm)
1	MAE	3570	7939	8157	9659	9087	11220
1	RSME	4277	12297	13767	13748	15255	18204
2	MAE	28835	35568	28769	27769	30945	26334
2	RSME	49007	65207	51079	46401	59493	44340
3	MAE	24045	30790	25495	30264	27541	29680
3	RSME	57779	65033	50185	65759	60459	75117

recorded rainfall over a particular threshold of interest, absolute or percentile value (in this instance, exceedances of the 90th percentile), then assessed according the fractional rainfall occurring within the neighbourhood. The FSS metric (eq. (3)) is then

[2] http://www.bom.gov.au/australia/charts/archive/index.shtml.

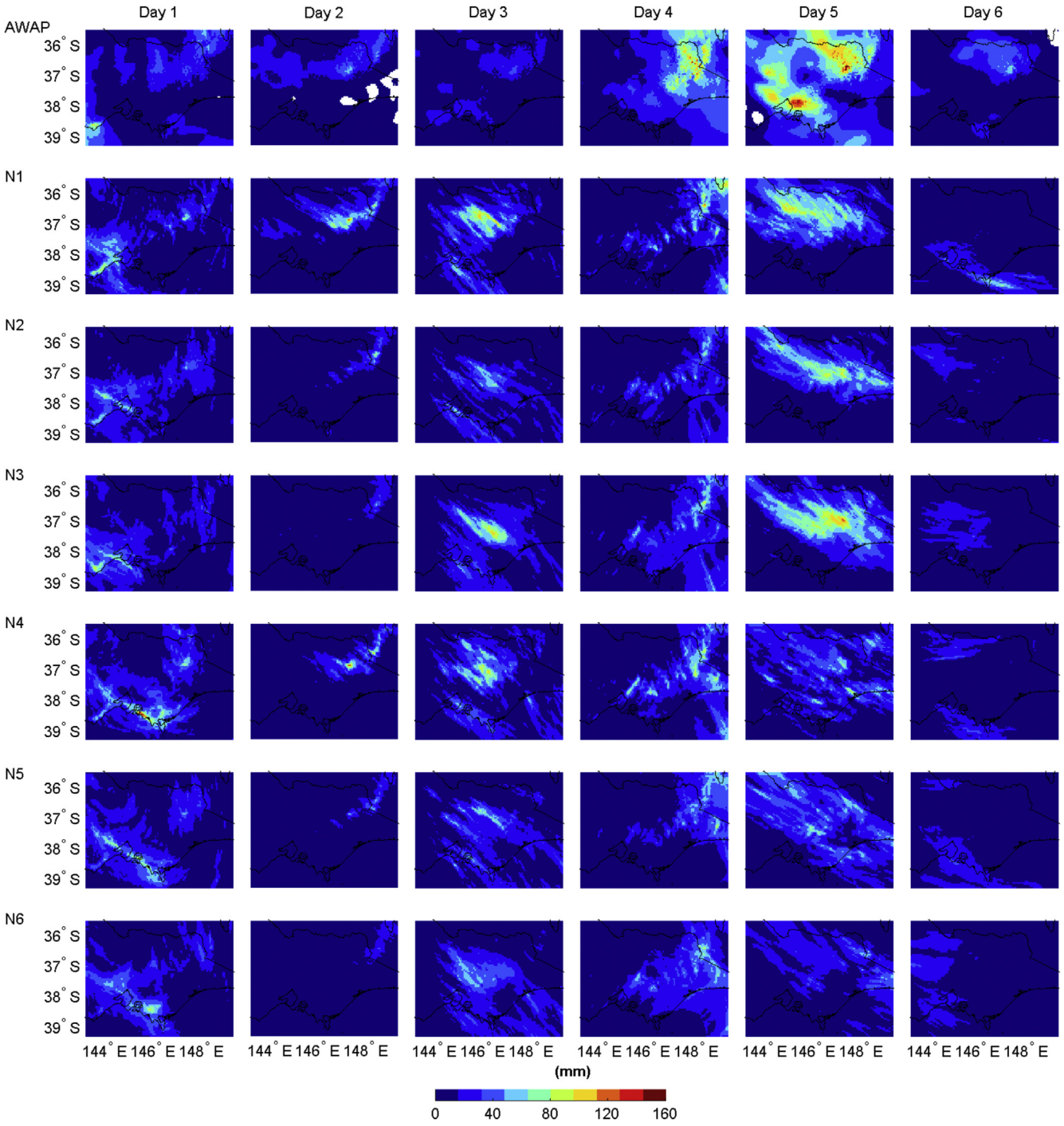

Fig. 4. Daily rainfall totals (mm) for selected Days 1–6 over domain D03. Top panel shows observed (AWAP) and following rows of panels show daily totals of WRF simulations N1–N6. All maps have resolution 0.05° (~5 km).

the empirical variogram (visualised by plotting the semi-variance (y-axis) with associated lag (x-axis)) and characterise its shape. The latter is done by fitting a suite of possible theoretical variogram models to the empirical semi-variogram. These models typically have two parameters that determine their two-dimensional shape, the sill and range. The former representing the zero-correlation semi-variance and the latter the spatial distance (lag) associated with the sill. Some empirical semi-variograms do not exhibit a well defined plateau for zero-correlation, in these instances the range is taken as an arbitrary 95% of the distance at which the sill parameter is estimated (Isaaks and Srivastava, 1989). A fitted model that does not have zero semi-variance at zero lag is said to have a nugget effect; typically interpreted as variance due to measurement error and variability in the data set on shorter range than the minimum sampled data spacing. Here, empirical semi-variograms were calculated and models (exponential, spherical, Gaussian and circular) fitted using geoR (Diggle and Ribeiro Jr., 2007). All models were fitted using ordinary least squares; the automated fitting procedure making use of about 60 different initial starting values of the sill and range parameters. For each empirical variogram, the

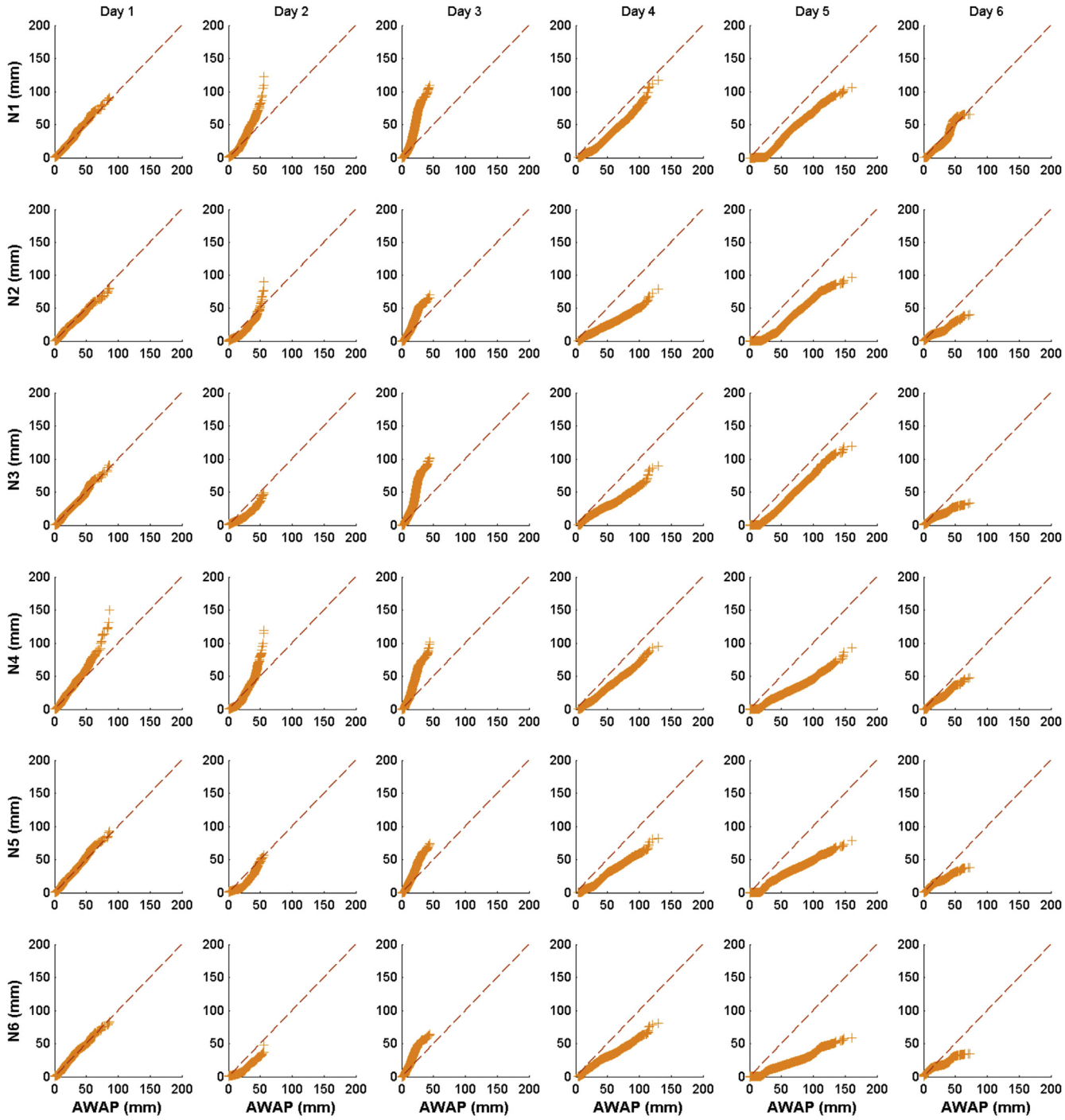

Fig. 5. Quantile–quantile plots of daily rainfall totals (mm) for selected days Day 1–6 over domain D03. Observed (AWAP) percentiles follow the x-axis and WRF simulation percentiles follow the y-axis. Columns separate simulation days and rows correspond to different WRF simulations (N1–N6).

model with best fit was kept for the computation of the variography metric.

A metric is estimated by considering the location in parameter space defined by coordinates given by the sill (including nugget) and range (both parameters normalised by their respective mean and standard deviation). The metric is simply the inverse Euclidean distance between the location of the observed (AWAP) parameters and those of the WRF simulation, so that simulations further away in parameter space has a smaller metric (see demonstration in Fig. 8). This metric was previously used as weights to combine projections of extreme rainfall fields for the UK (Fowler and Ekström, 2009).

3. Results

3.1. Daily rainfall characteristics

For comparison with observed daily rainfall fields WRF

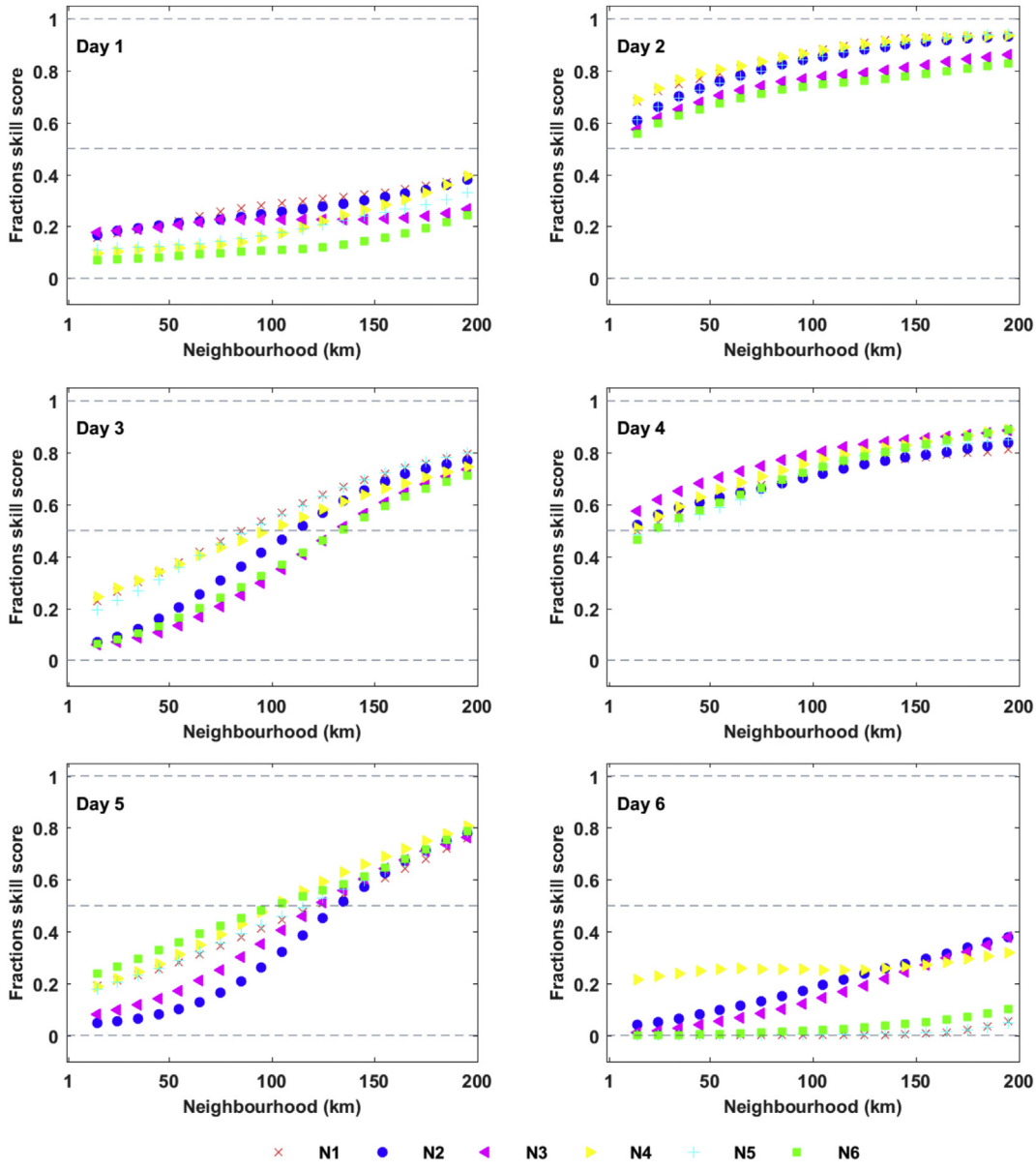

Fig. 6. FSS metric as calculated on the 90th percentile of daily rainfall (mm) across domain D03 for each of the selected days (Day 1–6) and WRF configuration (N1–N6, see legend). Grey lines denote lower and upper limits of the FSS metric (0 and 1), and the lower limit of skilful forecast (0.5).

Table 3
The FSS for selected days (Day 1–6) for a neighbourhood of 105 km. Yellow marking note the highest scoring model simulation, whilst red marks the second highest scoring simulation.

Case	Day	N1	N2	N3	N4	N5	N6
1	1	0.29	0.26	0.23	0.17	0.18	0.11
1	2	0.88	0.86	0.78	0.88	0.86	0.75
2	3	0.57	0.47	0.35	0.52	0.56	0.37
2	4	0.73	0.72	0.81	0.78	0.74	0.75
3	5	0.45	0.32	0.41	0.52	0.46	0.51
3	6	0.00	0.19	0.14	0.25	0.00	0.02

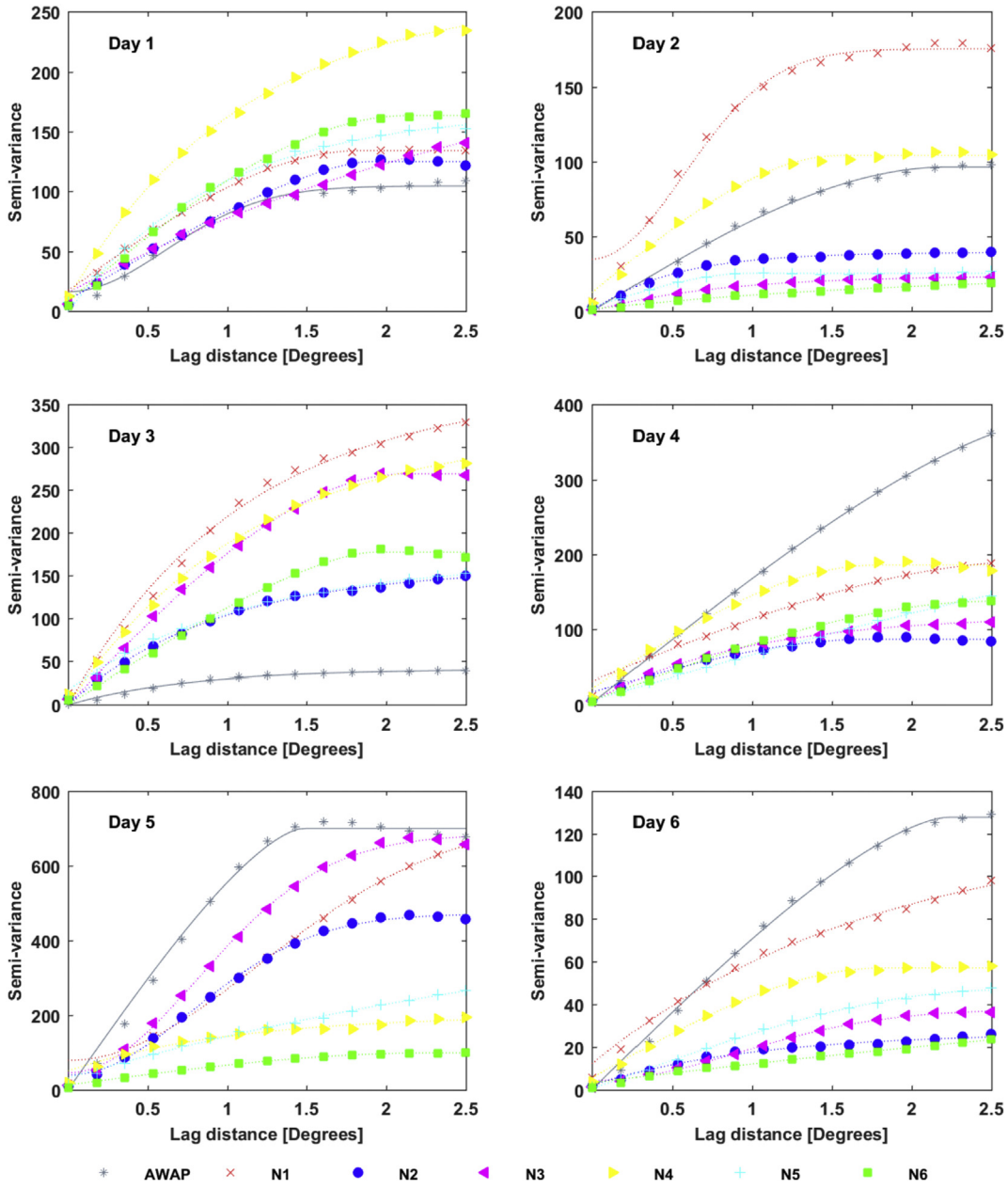

Fig. 7. Empirical variograms (symbols) and fitted variogram models (lines) for AWAP (grey) and model simulations (N1–N6, see legend) for selected days (Day 1–6).

simulations for D03 were regridded to the regular 0.05° latitude and longitude coordinates of AWAP. If aggregating rainfall across the entire domain, rainfall totals and the sequencing of events are overall well captured for the case studies (Fig. 3). For the winter case study (C1), daily rainfall totals are somewhat lower than observed on day 12 for the majority of ensemble members with the exception of best fit ensemble N1; the best fit calculated as the member with the smallest mean absolute error (MAE). In the shoulder season (C2), the overall timing of events are well simulated but magnitudes are somewhat overestimated, the maximum rainfall occurring on day 10 rather than day 11 (the best fit provided by ensemble member N6). The largest error in terms of magnitude is found in the summer case study (C3) on day 6, when rainfall amounts are underestimated by about approximately half of that observed. However, totals on other days are overall well captured, the best fit provided by ensemble member N1. If using the RMSE instead of MAE as a measure of best fit, N3 ranks somewhat higher than N1, reflecting the larger influence of large biases in this metric (Table 2).

The following assessments are based on the 6 selected days (2 non-consecutive days per case study) referred to as Day 1–6 (see section 2.3.1). A map of the daily rainfall total for each selected day is displayed in Fig. 4, where the top panels show observed rainfall totals based on gridded AWAP data and subsequent rows illustrate realisations as simulated by WRF configurations N1 to N6.

For Day 1 the location of rainfall is similar in position, but has a wider footprint compared to observed rainfall (Fig. 4). With the exception of ensemble member N4, simulated rainfall fields appear to be similar in magnitude relative to the observed data. Across ensemble members, those using PBL scheme YSU (N4-6) appear to simulate the higher intensity rainfall further east compared to observed patterns. Day 2 shows rainfall across the northern regions

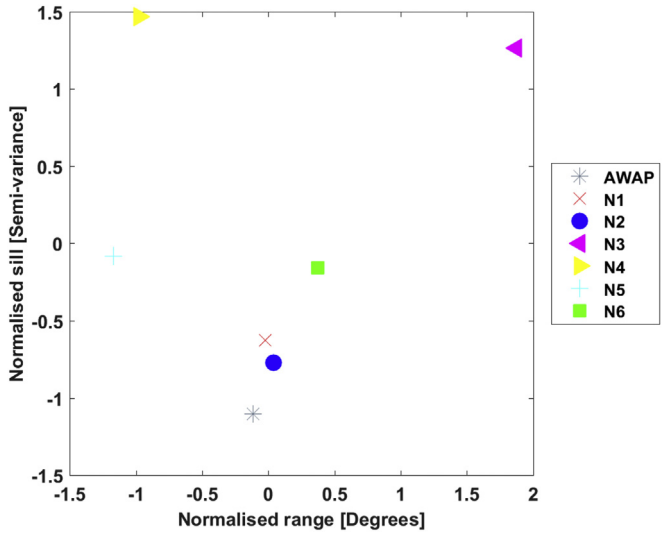

Fig. 8. Scatterplot of the normalised sill (y-axis) and range (x-axis) parameter values for all AWAP and WRF configurations (N1–N6) for selected Day 1. The resulting inverse distance calculated for N1–N6 relative to AWAP are shown in the first row of Table 4.

of D03 with larger totals over the Great Dividing Range. Simulated rainfall fields generally have good position of rainfall, but the extent is typically smaller and for many ensemble members the high rainfall amounts are greater than the observed amounts. For Day 2, rainfall patterns of pairs with the same MP scheme appear similar, i.e. larger magnitudes and extent when using WDM6 (N1 and N4), compared to Thompson (N2 and N5), and even more so in comparison to Milbrandt (N3 and N6). Simulated rainfall fields for Day 3 typically have larger magnitudes compared to observed rainfall (particular for N1, N3, and N4) with a larger extent, more so for ensemble members using PBL scheme YSU compared to MYNN. Day 4 and 5 show heavy rainfall events that are underestimated by all ensemble members in spatial extent and magnitudes. In terms of location, simulations do a reasonable job for Day 4, but for Day 5 heavy rainfall to the east of Melbourne is not captured by any model. For Day 5 there is a clear difference between simulations using PBL scheme MYNN and YSU, where the latter appears to generate a more wide spread and less intense rainfall field. Simulations for Day 6 do not capture the location of the observed rainfall, but rainfall magnitudes are overall reasonable.

The quantile–quantile plots illustrate the agreement in terms of distributional characteristics between observed and simulated rainfall fields for each of the selected days (Fig. 5). Though some plots show large similarities between observed and simulated rainfall, a 2 sample Kolmogorov Smirnov test ($\alpha = 0.05$) rejected the null hypotheses that samples were drawn from the same population for all days. Many of the features noted by a visual inspection of the maps are evident in these plots. Day 1 shows large similarity in observed and simulated magnitudes, with the exception of some overestimation of rainfall above 50 mm in N4. Day 2 shows some overestimation of higher magnitudes for ensemble members using MP scheme WDM6 (N1 and N4), and to a lesser extent when using Thompson in combination with MYNN (N2). Other ensemble members show a slight underestimation of magnitudes. For Day 3, all simulated fields overestimated observed magnitudes, more so by N1, N3 and N4. For Day 4–6, all ensemble members typically underestimated observed rainfall magnitudes, less so for N1. For Day 5, simulated and observed rainfall distributions appear to be more similar for simulations using PBL scheme MYNN (N1–3).

Using the FSS metric and variogram model parameters, we attempt to qualify and quantify characteristics of the simulated patterns relative to the observed patterns. The FSS metric was calculated using rainfall totals exceeding the 90th percentile for each simulation and a range of neighbourhood sizes (15–195 km) (Fig. 6); thus focussing on areas with heavier rainfall. Following Mittermaier and Roberts (2010), if the observed fractional rainfall across the region is small then acceptable scores are considered FSS>0.5. From this perspective, acceptable skill according to this measure is only achieved for Day 2 and 4 at all scales and for scales above 100 km for Day 3 and 5. No acceptable skill was achieved for any ensemble member at any scale for Day 1 and 6. Though there is spread amongst ensemble members, the variation amongst ensemble members is much less compared to skill given by any particular ensemble member across Days 1–6. Amongst the ensemble members, FSS scores for N4 tend towards the higher end, and those associated with N6 (Day 5 being a clear exception) towards the lower end of the range. If considering the best scoring members at 100 km (typical scale for mesoscale convective systems), the WDM6 MP scheme appears to perform well; the highest score given in combination with PBL scheme MYNN (N1) on Day 1 and 2, and with PBL scheme YSU (N4) on Day 5 and 6 (this configuration also gave the second best score for Day 4) (Table 3).

To assess the spatial characteristics of the daily rainfall omnidirectional (considering dependence in all directions) empirical semi-variograms was calculated for each simulated rainfall field (Day 1–6) and theoretical semi-variogram models fitted (Fig. 7). For Day 1 and 2, the simulated fields show more variability compared to observed fields (as shown by a larger sill for simulations), whilst on Day 4, 5 and 6 the observed variability is typically larger, particularly for Day 4 and 6. Using an inverse distance metric, a weight representing the proximity of the sill and range of simulated fields to those of the observed rainfall in a normalised Euclidean parameter space was calculated for all selected days. A graphical demonstration of the metric is shown in Fig. 8 for Day 1. Markers

Table 4
Weights derived from variography analysis, calculated as the inverse distance between AWAP and model output in a coordinate space defined by the variogram parameters (sill + nugget and range). Yellow marking note the highest scoring model simulation, whilst red marks the second highest scoring simulation.

Case	Day	N1	N2	N3	N4	N5	N6
1	1	2.05	2.72	0.32	0.37	0.68	0.94
1	2	0.43	0.39	0.44	1.42	0.50	0.82
2	3	0.33	0.90	0.30	0.38	0.75	0.39
2	4	0.50	0.30	0.29	0.43	0.44	0.39
3	5	4.07	1.01	3.07	0.49	0.40	0.40
3	6	3.51	0.40	0.44	0.55	0.47	0.33

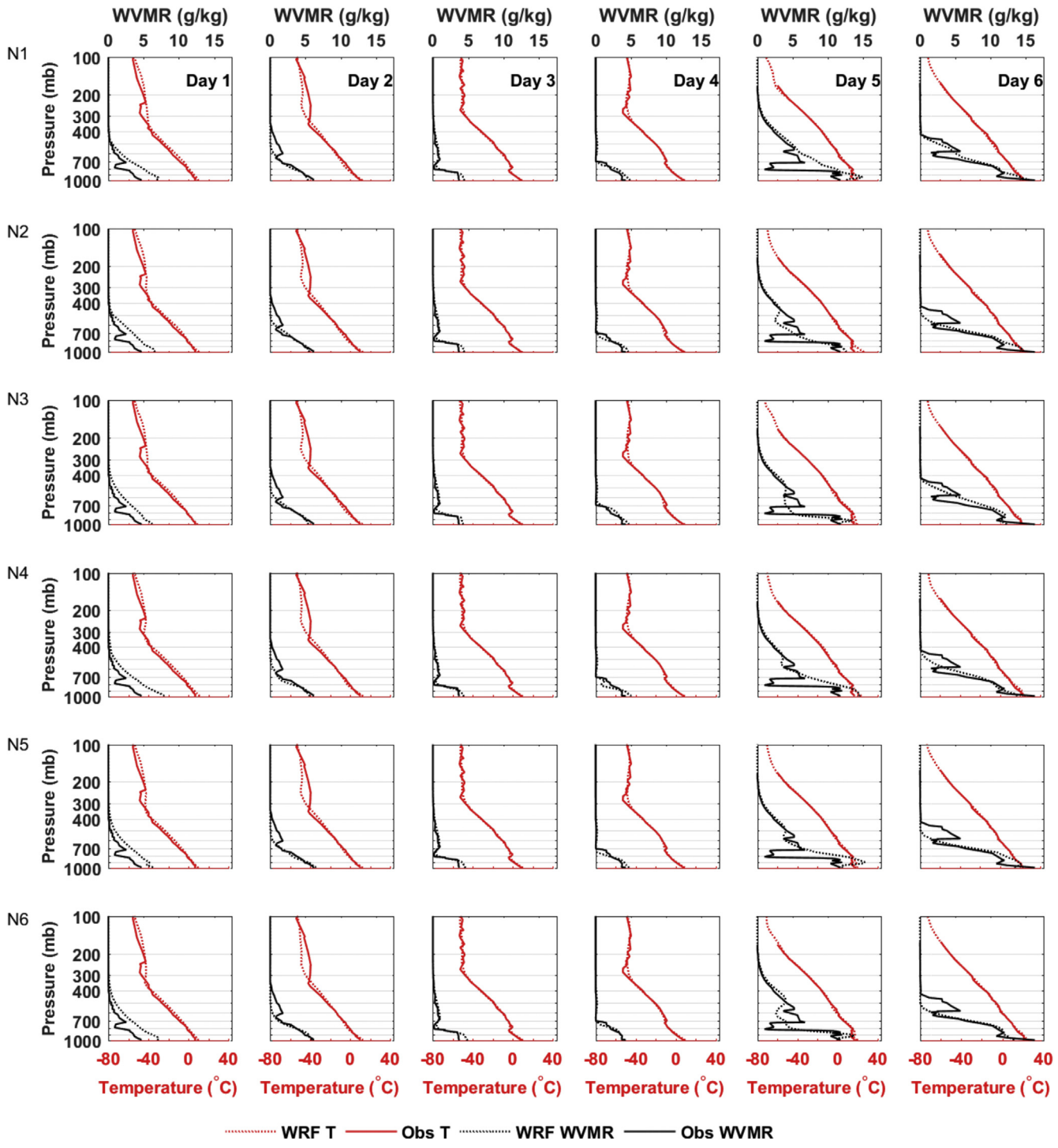

Fig. 9. Sounding data from Melbourne airport (full line) and corresponding values for closest matching grid cell (dotted line) of mixing ratio for water vapour (WVMR, g/kg) in black and temperature (°C) in red for each selected rainfall day (Day 1–6) and ensemble member (N1–6). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

show the similarity of the normalised sill and range values of WRF configurations (N1–N6) relative to those of the observed rainfall. The graph shows that markers for N1 and N2 are closest in 'parameter space' to the observed marker, resulting in larger metric values (inverse distance) relative to other WRF configurations (row 1 of Table 4). The variography metrics for Day 1–6 are summarised in Table 4 and shows that simulations of N1 are frequently most similar, or second most similar to observed rainfall.

3.2. Pressure level analysis

To gain a richer understanding of how the simulations differ in a physical sense, moisture and temperature characteristics of the WRF configurations are assessed relative to each other. Grounding the comparison to 'reality' is made by comparing output on pressure levels with observed sounding data; noting that whilst the sounding does represent observations of the atmosphere, the data

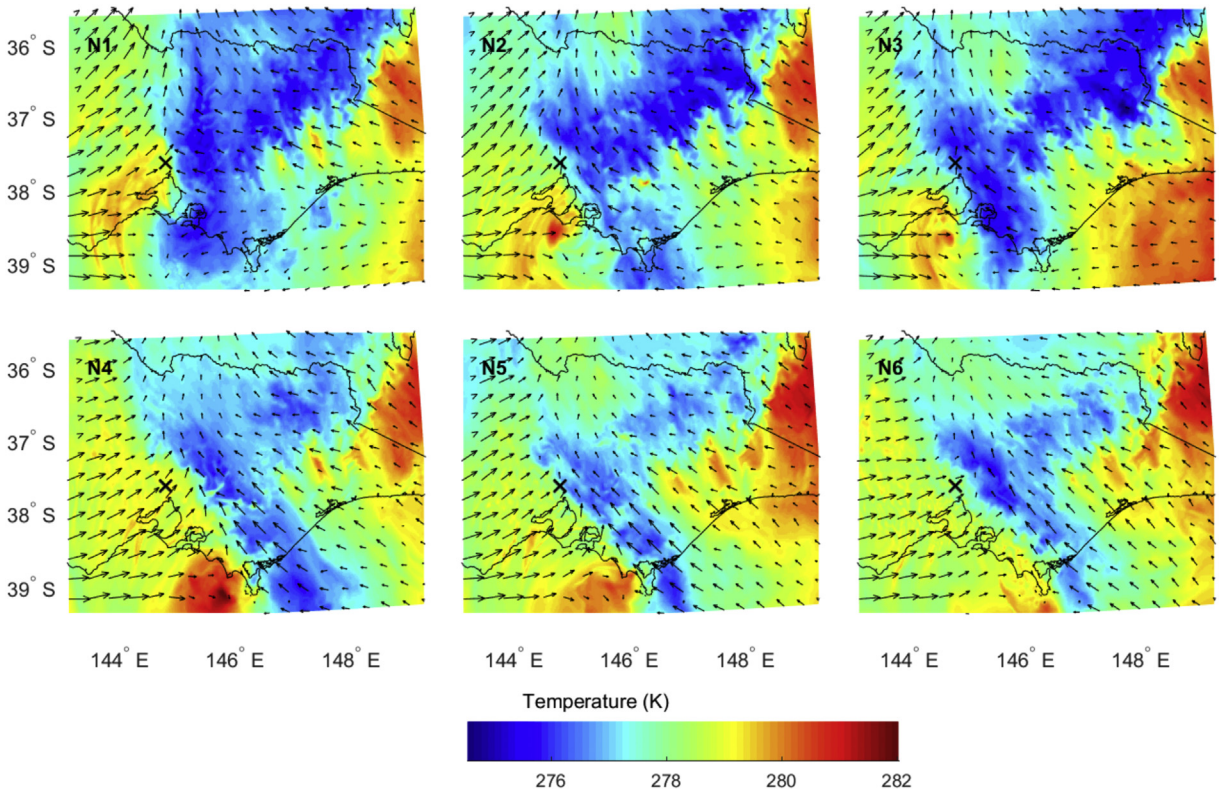

Fig. 10. Temperature (K) at 900 mb pressure level for selected Day 1 (10 am EST) and all ensemble members N1–6. Wind direction is overlayed to indicate air-mass movement. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

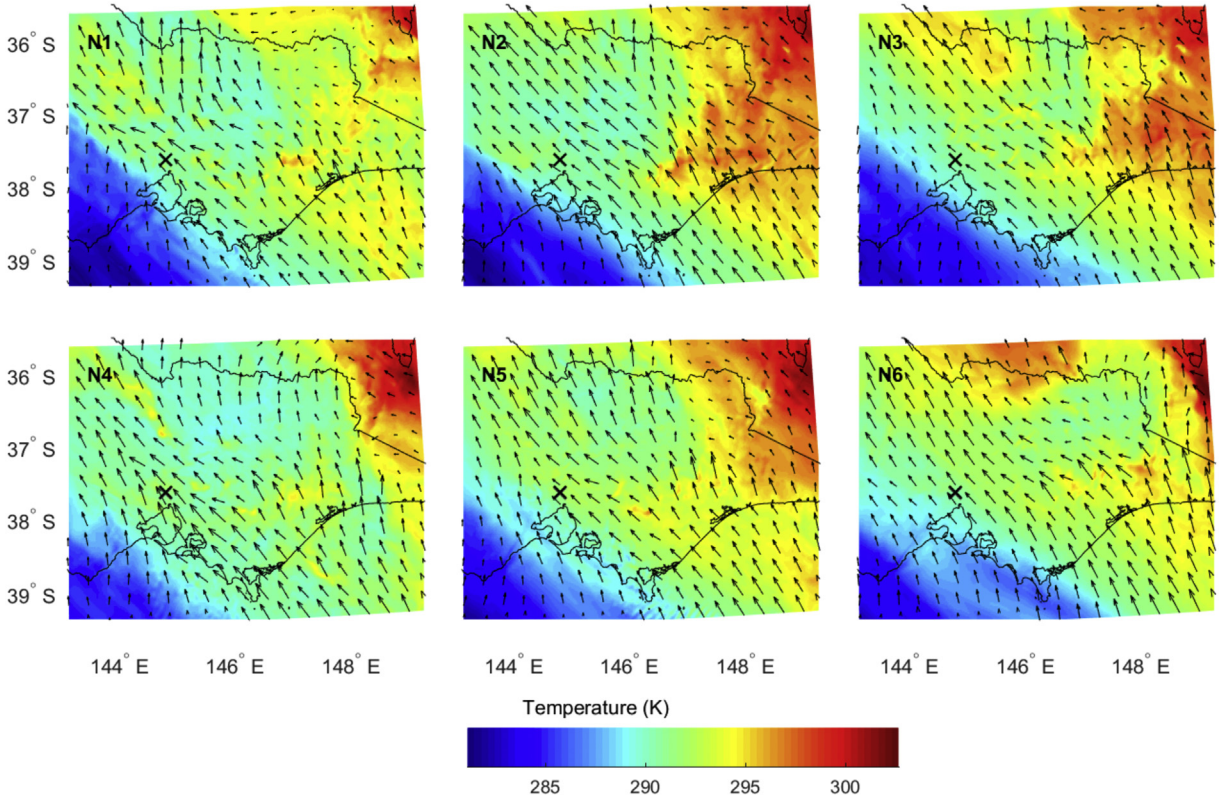

Fig. 11. Temperature (K) at 900 mb pressure level for selected Day 5 (11 am EDT) and all ensemble members N1–6. Wind direction is overlayed to indicate air-mass movement. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

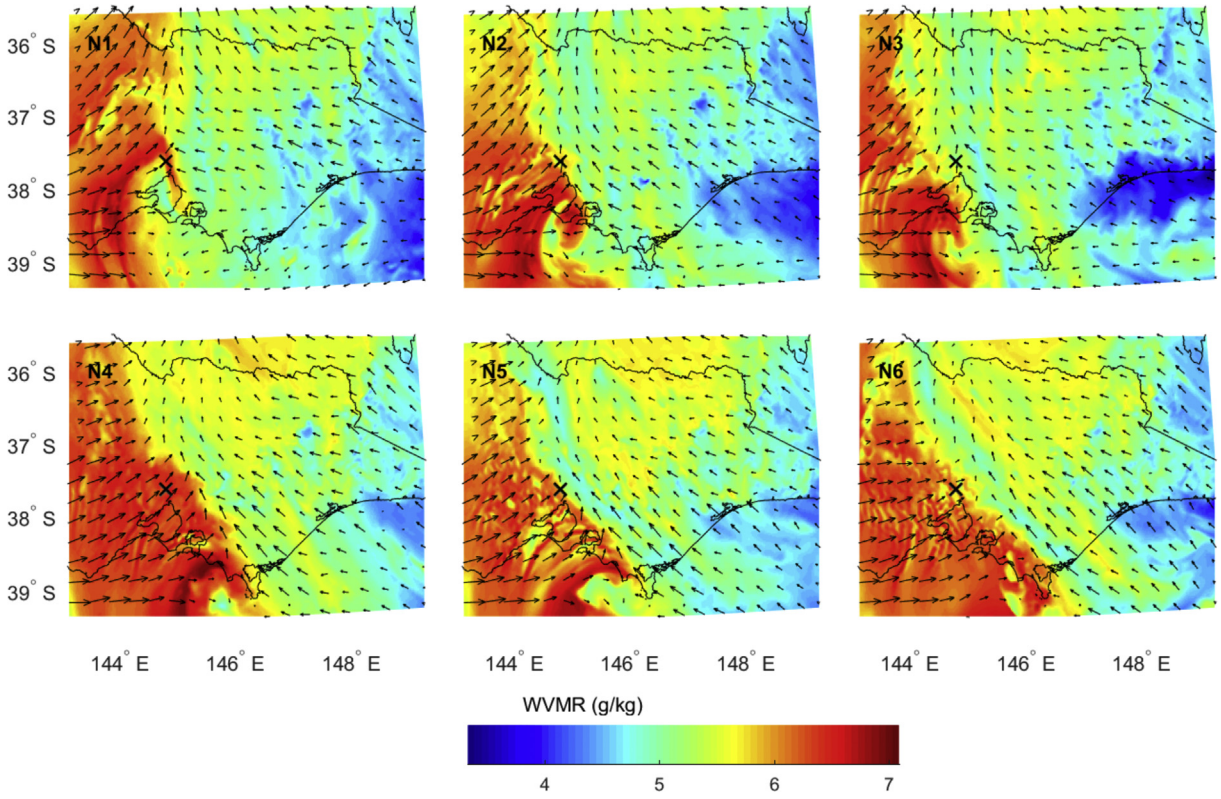

Fig. 12. Water Vapour Mixing Ratio (WVMR, g/kg) at 900 mb pressure level for selected Day 1 (10 am EST) and all ensemble members N1–6. Wind direction is overlayed to indicate air-mass movement. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

includes a location bias due to spatial drift as the balloon rises (McGrath et al., 2006). By viewing the sounding data in combination with maps of temperature and moisture content (water vapour and rain water mixing ratio) at relevant pressure levels, it is possible to give the vertical assessment a spatial context.

Fig. 9 shows the sounding data for selected rainfall days (Day1–6) and ensemble members (N1–6). In all simulations, observed and simulated temperatures show similar vertical profiles, with the exception of simulated temperatures being somewhat higher or lower than observed at pressure levels above 400 mb on Day 1 and 2 and somewhat higher surface temperatures on Day 5 for ensemble member N2. Less agreement is shown in moisture profile. For Day 1 all simulations tend to overestimate the amount of water vapour in the lower levels of the atmosphere and for Day 2, 3 and 4 simulations capture the overall characteristics of the observed profile with some common discrepancies. These include underestimation at elevations above 600 mb for Day 2 and overestimation at near surface levels (1000 mb) in Day 3. The warm summer-atmosphere of Day 5 and 6 show a greater moisture content and larger vertical variation compared to other selected days. The graphs show larger variation amongst the ensemble members than for previous days and no ensemble member shows a close resemblance to observed, though some agree more than others at different levels in the atmosphere. For Day 5, MP scheme WDM6 gives the smoothest vertical profile of all MP schemes (more so in combination with PBL scheme MYNN; N1) but does not capture the lower moisture levels around 800 mb. The Thompson MP scheme gives a similar profile to that of WDM6 for Day 5, but with underestimation at upper troposphere levels (more so when in combination with PBL scheme MYNN; N2) and overestimation of moisture content at lower levels when in combination with PBL scheme YSU (N5). The MP scheme Milbrandt appear to show the largest similarity to observed, particularly at lower levels in the atmosphere (>800 mb). For Day 6, the moisture profile of ensemble members using MP scheme WDM6 again show a 'smooth' characteristic (as for Day 5, more so when in combination with PBL scheme MYNN; N1). Furthermore, ensemble members using MP scheme Thompson and Milbrandt (in combination with YSU) all underestimate moisture content above 600 mb (N2, N5 and N6). In summary, ensemble members show similar biases towards observed profiles, though there are exceptions. Such as the mixed simulated representation of the zone of lower moisture content at 800 mb level on Day 5 and at 550 mb on Day 6.

To assess physical differences amongst the simulations maps of temperature and moisture content at a near surface level (900 mb) were mapped for all selected days (at 10 am EST for Day 1 and 2, and 11 am EDT for Day 3–6) for the D03 region (a marker indicating the location of the flux tower measurements is plotted in Fig. 9). Here, results are only displayed for Day 1 and 5, but maps for other days are available as supplementary material (including positive vertical winds).

The focus on lower levels is motivated by an interest to assess the influence of the physics schemes on the heat and moisture fluxes from the surface into the boundary layer of the atmosphere; reflecting skill in processes relevant to the models ability to simulate convective movement. However, note that whilst these maps elucidate differences amongst simulations, skill is difficult to attribute to individual configurations in absence of observational data. Some indication of skill may be gleaned from comparing the agreement in physical properties at the lower levels with the simulated rainfall response and the corresponding observed pattern (e.g. Fig. 4).

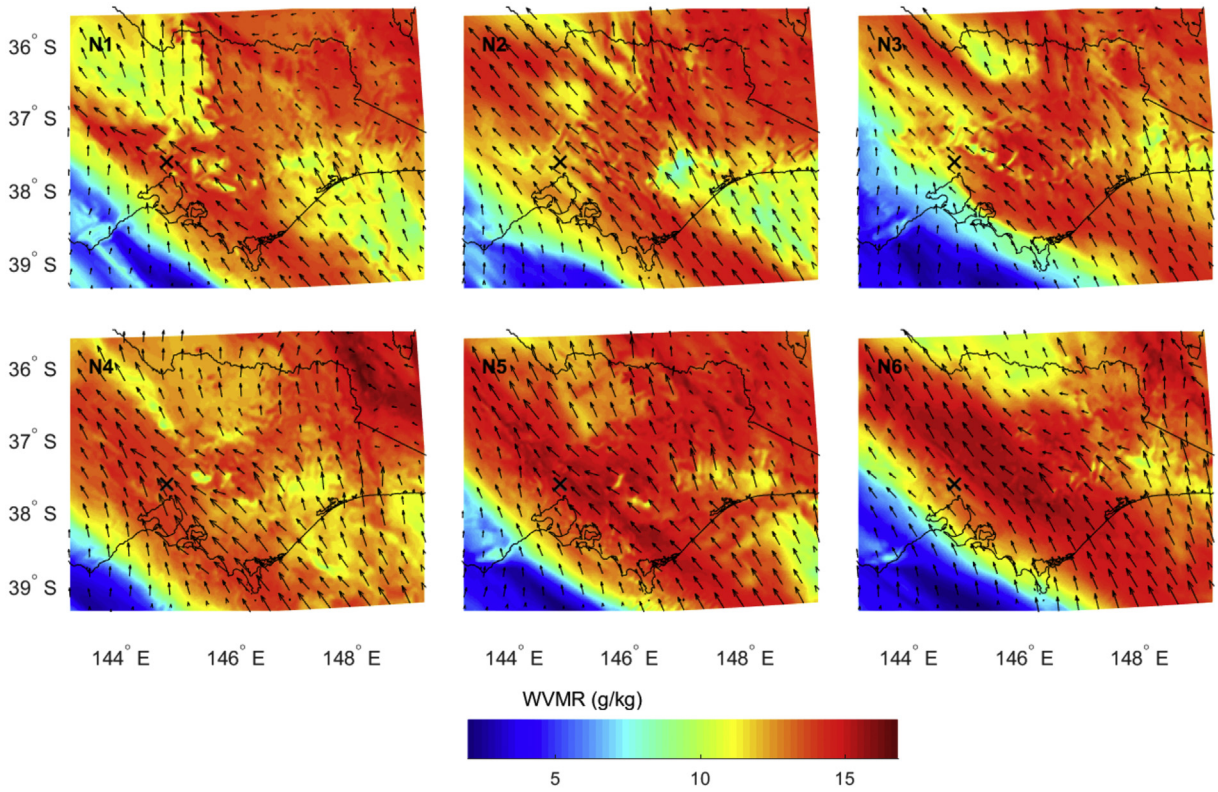

Fig. 13. Water Vapour Mixing Ratio (WVMR, g/kg) at 900 mb pressure level for selected Day 5 (11 am EDT) and all ensemble members N1–6. Wind direction is overlayed to indicate air-mass movement. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

A comparison of near surface temperature fields for the selected days show that simulations using the MYNN PBL scheme typically have a larger range in temperatures with greater spatial variability compared to those using YSU (Figs. 10 and 11). Temperatures in MYNN are typically a bit cooler (as shown for Day 1, but also visible in Day 5 particularly in the cooler air mass). This characteristics of greater spatial variability and variable range in MYNN simulations is also seen in the moisture fields, albeit not as clearly as for temperature (Figs. 12 and 13). For example, note the higher water vapour content of YSU compared to MYNN (particularly in Day 5 relative to Day 1). The warmer and more moist simulations of N4 to N6 may be related to the stronger mixing by the non-local YSU scheme relative to the local MYNN scheme; a local response in accordance with findings by Coniglio et al. (2013).

Viewing the patterns of temperature and moisture fields with maps of daily rainfall totals (Fig. 4) suggests an explanation for differences in simulated rainfall responses for Day 1, less so for Day 5. For Day 1, rainfall of MYNN PBL simulations (N1–3) give highest rainfall rates along the southern coastline around 144° E, whilst the YSU PBL simulations indicate the highest rainfall rates around 146° E (Figure 4, 1st column). Both temperature and moisture fields (and overlayed wind fields) for Day 1 show that the rainfall response follows an air mass boundary that is differently simulated by the two schemes (Figs. 10 and 12). In this instance, the location of the event in observed rainfall is more similar to that of the MYNN realisations (N1–N3). Day 5 offers a more complex synoptic situation for the simulated region compared to Day 1, as described in section 2.3.1. The temperature and moisture maps clearly show the demarcation between the moister air mass to the north and a cooler and drier air mass to the south (Figs. 11 and 13). For Day 5, WRF simulations show a simulated rainfall response that appear strongly influenced by the PBL scheme (Fig. 4). The simulations using the MYNN scheme (N1–N3) show heavy local rainfall compared to the widespread, but not as heavy, rainfall totals of the YSU simulations. Observed rainfall for Day 5 (Fig. 4) appear to agree more with simulations of MYNN, though the snapshots of temperature and moisture content from near midday does not offer a clear indication as to why.

For both days, maps of rain water mixing ration (RWMR, kg/kg) give an insight to characteristics of the MP schemes (Figs. 14 and 15). Higher rain water concentrations (and somewhat lesser spatial extent of the rainfall event) are found for the WDM6 schemes (N1 and N4), less so for Thompson (N2 and N5) and even lesser so for Milbrandt (N3 and N6). Without radar informed rainfall depths, it is difficult to assess which configuration is more similar to observed rainfall intensities, noting that N1 and to a lesser degree N4 scored well on the variography metric that looks as the spatial variance and dependence structure (Table 4).

4. Discussion and conclusions

Ensemble members assessed here differ only in two aspects: having a different combination of PBL and MP physics schemes. All tested schemes are standard options of WRF and as such are all expected to do well. The purpose of this assessment is to identify a configuration with characteristics that are desirable for water resource impact assessments using easy to implement metrics and analyses. Qualities that are deemed relevant are: timing of events (an indication of how the model simulates the movement of mass in the model domain); distributional qualities (ensuring that the full range of observed magnitudes are simulated); spatial extent and spatial characteristics (simulating rainfall processes that

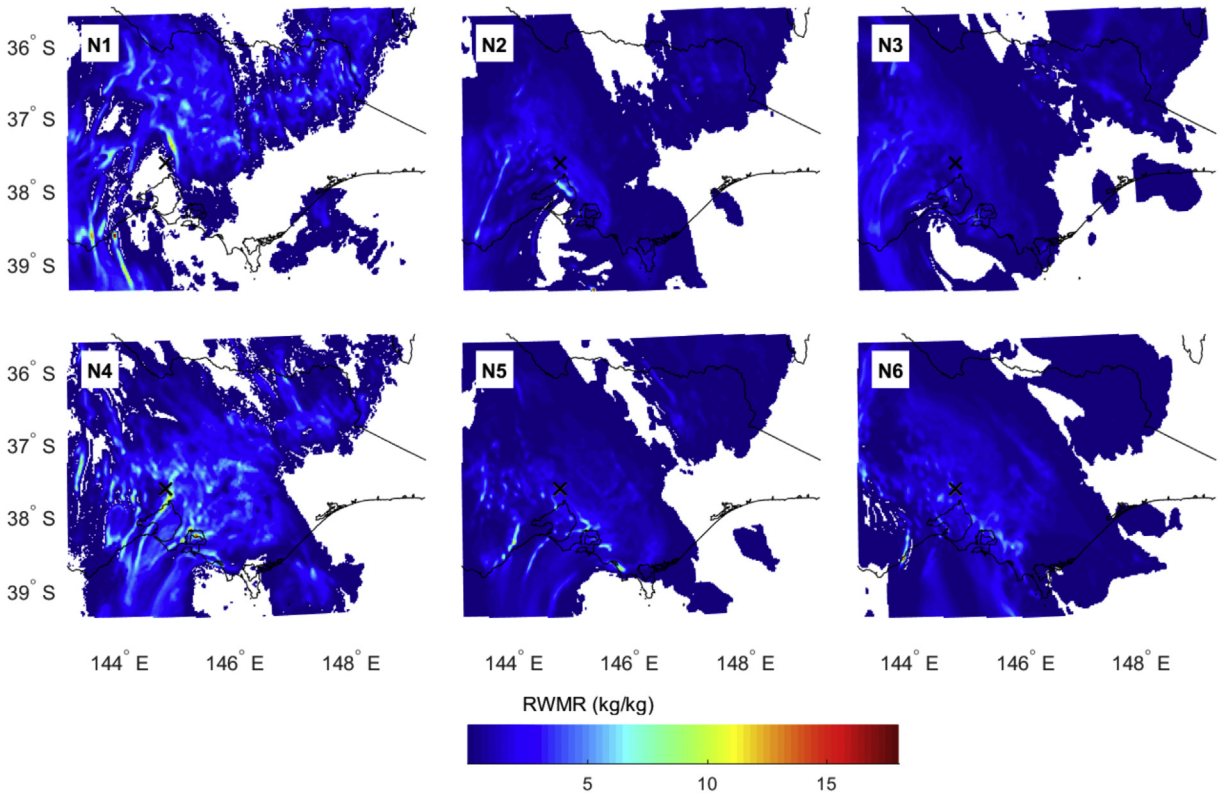

Fig. 14. Rain Water Mixing Ratio (RWMR, kg/kg) at 900 mb pressure level for selected Day 1 (10 am EST) and all ensemble members N1–6. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

adhere to observed patterns of variability and extent). Further analysis looked at differences in physical properties of the simulations, using atmospheric sounding data and horizontal temperature and moisture fields for the low level atmosphere.

Three two week case study simulations were completed for each of the considered WRF configurations, each representing a different season and synoptic conditions for the geographical area of interest. Subsequently two non-consecutive days of the heavier rainfall days within the simulation period were selected for a more detailed spatial analysis (discrepancies due to different use of physics schemes being easier to identify for the heavier events). All evaluations used only the innermost simulation domain (D03), to maximise the influence of the MP scheme on the rainfall estimates.

Different metrics and analyses were tested to see how well they captured desired characteristics in the simulated rainfall fields. A metric is considered meaningful if it captures aspects that are deemed relevant to the intended application (as mentioned above), easily understood and directly relates to conclusions drawn when visually comparing simulated and observed rainfall fields.

The timing of rainfall events was assessed comparing totals of simulated and observed rainfall for the innermost domain D03. Generally the MAE and RMSE gave similar ranking amongst the ensemble members; though for case study 3 one day with very large observed rainfall that was not well captured by either ensemble member caused a difference in ranking when using the different metrics. Perhaps for such small samples, the MAE is the more robust measure as single events can otherwise have very large impact on the metric. The quantile–quantile plots provided a good visual assessment of distributional similarities. A two sample Kolmogorov Smirnov was applied as a goodness-of-fit test, but as all tests gave very low p-values, the test did not provide useful information to differentiate amongst simulations. Other metrics were more informative. The FSS scores provided skill information on positioning of heavy rainfall centres (90th percentile exceedances), and the resolution for which skilful simulations are achieved. Whilst the FSS scores reduce the double penalty influence in comparison to grid based evaluations, the metric is still a function of position. As there is generally large similarity amongst ensemble members in terms of positioning, the FSS scores typically indicated similar skill across the ensemble. Greater separation in skill was provided by the variography metric, combining the sill and range parameters of the semi-variogram. Based on information contained within the sample assessed (no grid cell-to-grid cell comparison) the metric captured generic features of the spatial rainfall pattern. Overall, the variography analysis proved useful in terms of summarising spatial characteristics as seen in the mapped rainfall patterns, always giving highest rank to the pattern visually identified to be most similar to that of observed. Whilst providing an opportunity to look beyond the rainfall fields, the sounding profile did not easily provide information that could qualify or quantify skill around the desired criteria. The vertical profiles provided verification of the simulated atmospheric structure at a single location, but typically required further analysis of 2 dimensional horizontal or vertical fields to better understand the context. The main benefit from viewing the pressure level fields was a better understanding of different characteristics of the WRF configurations. The MYNN PBL simulations appeared somewhat cooler and drier compared to the YSU simulations and simulations using the WDM6 scheme stood out as having somewhat greater localised rainfall intensities compared to those using the Thompson and Milbrandt MP schemes. Without a formal assessment, however, this paper is not in a position to attribute skill to these

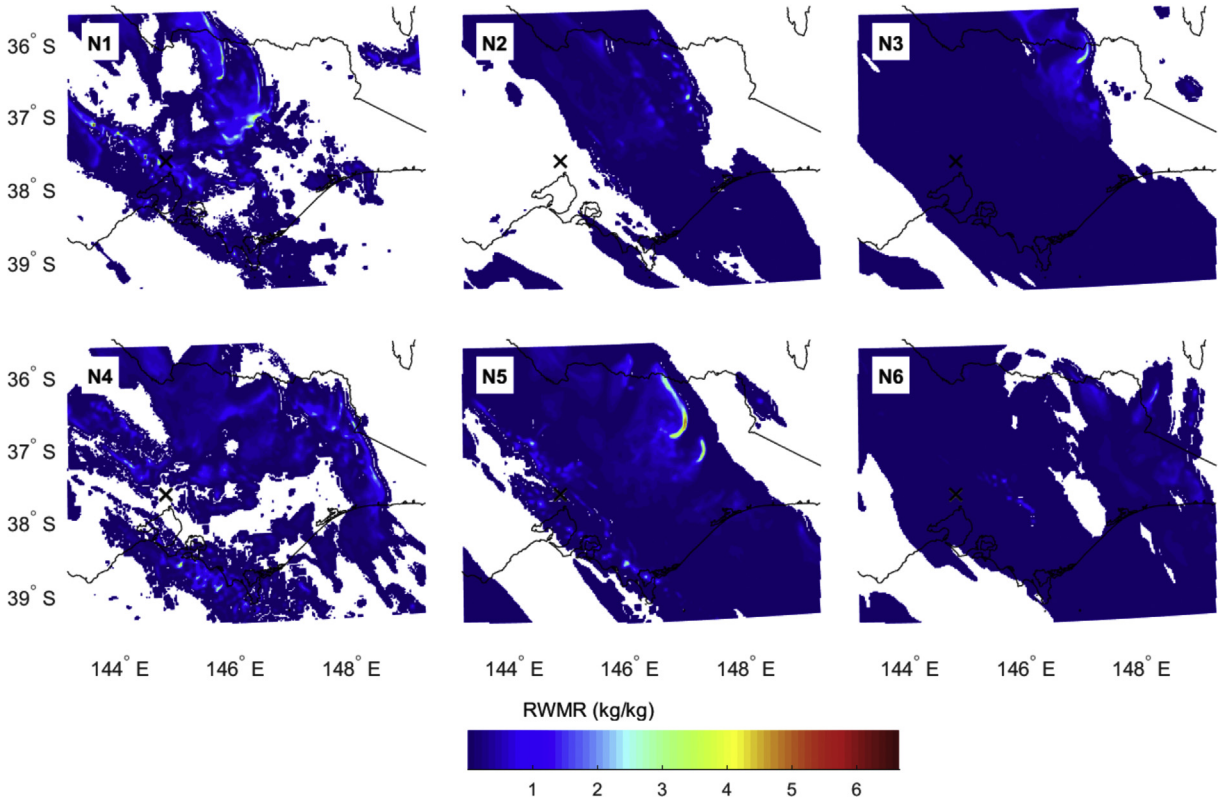

Fig. 15. Rain Water Mixing Ratio (RWMR, kg/kg) at 900 mb pressure level for selected Day 5 (11 am EDT) and all ensemble members N1–6. The markers show the location of the sounding station (black) and the nearest grid cell coordinate (grey); note that because locations are very close the markers overlap.

characteristics; noting that the stronger mixing by YSU in comparison to MYNN was also noted by Coniglio et al. (2013) and in general by Evans et al. (2012).

Drawing on information based on these metrics, ensemble members performing somewhat better than others could be identified for each of the studied rainfall days. More often than not, the N1 ensemble member was identified as performing well, particularly when using the variography metric. Hence, when considering magnitude, timing, distribution and spatial skill, this experiment points toward N1 having the type of qualities desired for impact studies in the water resource sector.

Acknowledgements

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. The work is conducted as part of the Victorian Climate Initiative (VicCI) funded by the Victorian Government's Department of Environment, Land, Water and Planning (DELWP), the Australian Bureau of Meteorology and the Australian Commonwealth Science, Industry and Research Organisation (CSIRO).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2016.01.012.

References

Burton, A., Glenis, V., Jones, M.R., Kilsby, C.G., 2013. Models of daily rainfall cross-correlation for the United Kingdom. Environ. Model. Softw. 49, 22–33.
Caldwell, P., Chin, H.-N.S., Bader, D.C., Bala, G., 2009. Evaluation of a WRF dynamical downscaling simulation over California. Clim. Change 95.
Casati, B., 2010. New developments of the intensity-scale technique within the spatial verification methods intercomparison project. Weather Forecast. 25, 113–143.
Chiles, J.-P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, Inc, New York, USA.
Chotamonsak, C., Salathe Jr., E.P., Kreasuwan, J., Chantara, S., Siriwitayakorn, K., 2011. Projected climate change over Southeast Asia simulated using a WRF regional climate model. Atmos. Sci. Lett. 12.
Coniglio, M.C., Correia Jr., J., Marsh, P.T., Kong, F., 2013. Verification of convection-allowing WRF model forecasts of the planetary boundary layer using sounding observations. Weather Forecast. 28, 842–862.
Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Holm, E.V., Isaksen, L., Kallberg, P., Koehler, M., Matricardi, M., Mcnally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J.N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. Royal Meteorol. Soc. 137, 553–597.
Del Genio, A.D., Wu, J., Chen, Y., 2012. Characteristics of mesoscale organization in WRF simulations of convection during TWP-ICE. J. Clim. 25.
Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York.
Diggle, P.J., Ribeiro Jr., P.J., 2007. Model-based Geostatistics. Springer-Verlag New York, New York, USA.
Done, J., Davis, C.A., Weisman, M., 2004. The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. Atmos. Sci. Lett. 5.
Ebert, E.E., 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol. Appl. 15, 51–64.
Ebert, E.E., Gallus Jr., W.A., 2009. Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. Weather Forecast. 24, 1401–1415.
Ekstrom, M., 2014. Test of WRF Physics Schemes for Project 6 of the Victorian Climate Initiative: An Outline of Selected Physical Parameterisation Schemes and Other Runtime Options. CSIRO Water for a Healthy Country Flagship, Australia.
Ekström, M., Grose, M.R., Whetton, P.H., 2015. An appraisal of downscaling methods used in climate change research. WIREs Clim Change 6, 301–319. http://

dx.doi.org/10.1002/wcc.339.
Emmanuel, I., Andrieu, H., Leblois, E., Flahaut, B., 2012. Temporal and spatial variability of rainfall at the urban hydrological scale. J. Hydrol. 430, 162–172.
Evans, J., Ekstrom, M., Ji, F., 2012. Evaluating the performance of a WRF physics ensemble over South-East Australia. Clim. Dyn. 39, 1241–1258.
Fowler, H., Ekstrom, M., Kilsby, C., Jones, P., 2005. New estimates of future changes in extreme rainfall across the UK using regional climate model integrations. 1. Assessment of control climate. J. Hydrol.Hydrol. 300, 212–233.
Fowler, H.J., Ekström, M., 2009. Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. Int. J. Climatol. 29, 385–416.
Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., Ebert, E.E., 2009. Intercomparison of spatial forecast verification methods. Weather Forecast. 24, 1416–1430.
Gilleland, E., Ahijevych, D.A., Brown, B.G., Ebert, E.E., 2010. Verifying forecasts spatially. Bull. Am. Meteorol. Soc. 91, 1365–1373.
Gilmore, J., Evans, J., Sherwood, S., Ekström, M., Ji, F., 2015. Extreme precipitation in WRF during the newcastle east coast low of 2007. Theor. Appl. Climatol. 1–19.
Heikkila, U., Sandvik, A., Sorteberg, A., 2011. Dynamical downscaling of ERA-40 in complex terrain using the WRF regional climate model. Clim. Dynam. 37.
Hong, S.-Y., Noh, Y., Dudhia, J., 2006. A new vertical diffusion package with an explicit treatment of entrainment processes. Mon. Weather Rev. 134, 2318–2341.
Iacono, M.J., Delamere, J.S., Mlawer, E.J., Shephard, M.W., Clough, S.A., Collins, W.D., 2008. Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models. J. Geophys. Res. Atmos. 113.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, USA.
Janjic, Z.I., 1994. The step-mountain ETA coordinate model – further developments of the convection, viscous sublayer, and turbulence closure schemes. Mon. Weather Rev. 122, 927–945.
Janjic, Z.I., 2000. Comments on “Development and evaluation of a convection scheme for use in climate models”. J. Atmos. Sci. 57, 3686.
Jankov, I., Gallus, W.A., Segal, M., Shaw, B., Koch, S.E., 2005. The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. Weather Forecast. 20.
Ji, F., Ekström, M., Evans, J., Teng, J., 2014. Evaluating rainfall patterns using physics scheme ensembles from a regional atmospheric model. Theor. Appl. Climatol. 115, 297–304.
Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Aust. Meteorol. Oceanogr. J. 58, 233–248.
Kain, J.S., Weiss, S.J., Levit, J.J., Baldwin, M.E., Bright, D.R., 2006. Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: the SPC/NSSL spring program 2004. Weather Forecast. 21.
Kendon, E.J., Roberts, N.M., Fowler, H.J., Roberts, M.J., Chan, S.C., Senior, C.A., 2014. Heavier summer downpours with climate change revealed by weather forecast resolution model. Nat. Clim. Change 4, 570–576.
Kendon, E.J., Roberts, N.M., Senior, C.A., Roberts, M.J., 2012. Realism of rainfall in a very high-resolution regional climate model. J. Clim. 25, 5791–5806.
Lepioufle, J.M., Leblois, E., Creutin, J.D., 2012. Variography of rainfall accumulation in presence of advection. J. Hydrol. 464, 494–504.
Leung, L.R., Kuo, Y.-H., Tribbia, J., 2006. Research needs and directions of regional climate modeling using WRF and CCSM. Bull. Am. Meteorol. Soc. 87, 1747–1751.
Li, J., Hsu, K., Aghakouchak, A., Sorooshian, S., 2015. An object-based approach for verification of precipitation estimation. Int. J. Remote Sens. 36, 513–529.
Lim, K.-S.S., Hong, S.-Y., 2010. Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models. Mon. Weather Rev. 138, 1587–1612.
Liu, P., Tsimpidi, A.P., Hu, Y., Stone, B., Russell, A.G., Nenes, A., 2012. Differences between downscaling with spectral and grid nudging using WRF. Atmos. Chem. Phys. 12.
Ma, Z., Fei, J., Huang, X., Cheng, X., 2012. Sensitivity of tropical cyclone intensity and structure to vertical resolution in WRF. Asia Pac. J. Atmos. Sci. 48.
Mcgrath, R., Semmler, T., Sweeney, C., Wang, S., 2006. Impact of balloon drift errors in radiosonde data on climate statistics. J. Clim. 19, 3430–3442.
Milbrandt, J.A., Yau, M.K., 2005. A multimoment bulk microphysics parameterization. Part I: analysis of the role of the spectral shape parameter. J. Atmos. Sci. 62, 3051–3064.
Mittermaier, M., Roberts, N., 2010. Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. Weather Forecast. 25, 343–354.
Mittermaier, M., Roberts, N., Thompson, S.A., 2013. A long-term assessment of precipitation forecast skill using the fractions skill score. Meteorol. Appl. 20, 176–186.
Nakanishi, M., Niino, H., 2006. An improved mellor-yamada level-3 model: its numerical stability and application to a regional prediction of advection fog. Bound. Lay. Meteorol. 119, 397–407.
Ncar, 2013. ARW Version 3 Modelling System User's Guide. Mesoscale & Microscale Meteorology Division, National Centre for Atmospheric Research.
Roberts, N.M., Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Weather Rev. 136, 78–97.
Skamarock, W.C., Klemp, J.B., 2008. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. J. Comput. Phys. 227, 3465–3485.
Thompson, G., Field, P.R., Rasmussen, R.M., Hall, W.D., 2008. Explicit forecasts of Winter precipitation using an improved bulk microphysics scheme. Part II: implementation of a new snow parameterization. Mon. Weather Rev. 136, 5095–5115.
Westra, S., Fowler, H.J., Evans, J.P., Alexander, L.V., Berg, P., Johnson, F., Kendon, E.J., Lenderink, G., Roberts, N.M., 2014. Future changes to the intensity and frequency of short-duration extreme rainfall. Rev. Geophys. 52, 522–555.
Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences. Elsevier, US.