Comprehensive Methodology for the Evaluation of High-Resolution WRF Multiphysics Precipitation Simulations for Small, Topographically Complex Domains

IOANNIS SOFOKLEOUS,[a] ADRIANA BRUGGEMAN,[a] SILAS MICHAELIDES,[b] PANOS HADJINICOLAOU,[b] GEORGE ZITTIS,[b] AND CORRADO CAMERA[c]

[a] *Energy, Environment and Water Research Center, The Cyprus Institute, Nicosia, Cyprus*
[b] *Climate and Atmosphere Research Center, The Cyprus Institute, Nicosia, Cyprus*
[c] *Dipartimento di Scienze della Terra "A. Desio," Università degli Studi di Milano, Milan, Italy*

(Manuscript received 28 April 2020, in final form 10 December 2020)

ABSTRACT: A stepwise evaluation method and a comprehensive scoring approach are proposed and implemented to select a model setup and physics parameterizations of the Weather Research and Forecasting (WRF) Model for high-resolution precipitation simulations. The ERA5 reanalysis data were dynamically downscaled to 1-km resolution for the topographically complex domain of the eastern Mediterranean island of Cyprus. The performance of the simulations was examined for three domain configurations, two model initialization approaches and 18 combinations of atmospheric physics parameterizations. Two continuous and two categorical scores were used for the evaluation. A new extreme event score, which combines hits and frequency bias, was introduced as a complementary evaluator of extremes. A composite scaled score was used to identify the overall best performing parameterizations. The least errors in mean daily and monthly precipitation amounts and daily extremes were found for the domain configuration with the largest extent and three nested domains. A 5-day initialization frequency did not improve precipitation, relative to 30-day continuous simulations. The parameterization type with the largest impact on precipitation was microphysics. The cumulus parameterization was also found to have an impact on the 1-km nested domain, despite that it was only activated in the coarser "parent" domains. Comparison of simulations with 12-, 4-, and 1-km resolution revealed the better skill of the model at 1 km. The impact of the various model configurations in the small-sized domain was different from the impact in larger model domains; this could be further explored for other atmospheric variables.

KEYWORDS: Model errors; Model evaluation/performance; Model initialization; Numerical analysis/modeling; Parameterization; Reanalysis data

1. Introduction

Dynamical downscaling of global reanalysis data, with limited area models (LAM) or regional climate models (RCM), is often used for the reconstruction of past weather events and climate conditions at a range of spatial resolutions. At high spatial resolution (≈12 km), simulations driven by reanalysis data reproduce mean precipitation and extremes better than simulations at coarser spatial resolutions (≈50 km; Prein et al. 2016), while convection-permitting resolutions (<4 km) outperform all other resolutions in the simulation of precipitation extremes and the diurnal cycle of convective precipitation (Berthou et al. 2018; Piazza et al. 2019). Hydrological modeling applications for medium-size watersheds, with areas ranging between 10 and 500 km^2, require dynamically downscaled precipitation at resolutions on the order of 1 km with temporal resolutions less than 1 h (e.g., Camera et al. 2020). However, dynamical downscaling is subject to multiple modeling challenges, such as the choice of the size of the gridded model domain, the frequency of updating the model initial conditions and the applicability of the atmospheric physics parameterization schemes (e.g., Giorgi and Gutowski 2015; Kioutsioukis et al. 2016).

The impact of the size of gridded model domains, referred to as domain size hereafter, on precipitation simulations has been extensively studied at coarse spatial resolutions (>40 km) for continental-scale sized study areas. While some studies reported very similar performance with different domain sizes (Colin et al. 2010; Centella-Artola et al. 2015), other studies reported discrepancies in the modeled precipitation, which are caused by the topographical effects induced by the geographical coverage of the various domain sizes (Seth and Giorgi 1998; Leduc and Laprise 2009). The impact of domain size has received less attention for horizontal resolutions below 10 km. Song et al. (2018) found that a horizontally extended domain, which includes steep topography close to the lateral boundaries, may yield adverse effects on regional climate simulations at 6-km resolution. For simulations at 2.5-km resolution, over a region with relatively homogenous flat terrain, Brisson et al. (2016) found that a minimum distance of about 150 km between the evaluation domain and the lateral boundary was required to achieve convergence of modeled precipitation toward observations. According to Rojas and Seth (2003), the choice of the domain size should be associated with the quality of the lateral boundary conditions (LBCs) because RCM simulations in a large domain can partly compensate low-resolution LBCs. Vannitsem and Chomé (2005) suggest,

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-20-0110.s1.

Corresponding author: Ioannis Sofokleous, i.sofokleous@cyi.ac.cy

DOI: 10.1175/JHM-D-20-0110.1

however, that high-resolution LBCs should not always motivate the use of a small domain size. The state-of-the-art ERA5 global dataset, which has higher resolution (≈31 km) than other widely used reanalysis products, such as NCEP-GFS (≈56 km) and ERA-Interim (≈79 km), can serve as a source for such high-resolution LBCs for dynamical downscaling experiments on the impact of domain configuration.

The performance of dynamical downscaling is also related to the frequency of updating the model initial conditions, referred to as model initialization frequency hereafter. Commonly, dynamical downscaling is performed with continuous model simulations, initialized once at the beginning of the simulation period. The deviation of the simulated atmospheric fields from the large-scale driving data is, however, a well-known issue with longer simulation times (Vitart 2004). To overcome this reduction of skill with time, long and continuous model simulations can be divided into segments that are separately initialized. Qian et al. (2003) and Seth et al. (2004) found that the increase in the initialization frequency from 5 months to 30 days and to 10 days led to a reduction in systematic errors of precipitation and an improved spatial distribution of precipitation intensities in simulations over South America. Lo et al. (2008), used the Weather Research and Forecasting (WRF) Model at 36-km horizontal resolution and found that weekly initializations gave a higher skill in simulated precipitation over the United States than monthly initializations. Lucas-Picher et al. (2013) found that dynamical downscaling with the HIRHAM RCM at 12-km resolution over Europe with daily initialization resulted in improved temporal and spatial correlation of precipitation, relative to continuous multiyear simulations, particularly during summer. From the above studies, the positive impact of frequent initializations on precipitation simulations, relative to the long and continuous simulations over continental-scale domains is evident. The same has, however, not been investigated over small domains with sizes on the order of few hundred kilometers.

Additional aspects of the dynamical downscaling process that may lead to errors in the RCM output are the model physics and model horizontal resolution (Buizza et al.1999; Lo et al. 2008). Various studies have evaluated the performance of different physics parameterization schemes for the dynamical downscaling of reanalysis data, in order to reproduce precipitation with the WRF Model over different regions. At horizontal resolutions well above the convection-resolving scales, several studies [Ji et al. (2014), 10 km; Hu et al. (2018), 20 km; Zittis et al. (2014), 50 km; and Katragkou et al. (2015), 50 km] found that planetary boundary layer (PBL) and cumulus schemes have the largest impact on simulated precipitation. Across the gray zone for cumulus schemes, i.e., the spatial resolutions at which the model starts to partially resolve the development of cumulus clouds (between 10 and 4 km), Jeworrek et al. (2019) showed that the effect of the choice of the cumulus parameterization was higher than the effect of the microphysics parameterization choice, for simulations of convective precipitation in the southern Great Plains in the United States. At convection-resolving resolutions, i.e., below 3 km, cumulus schemes are generally not activated, and studies have highlighted the dominant role of different microphysics schemes on the total volume of simulated precipitation (Cassola et al. 2015; Zittis et al. 2017; Mohan et al. 2018). Furthermore, Avolio and Federico (2018) reported the important impact of PBL schemes for simulating heavy rainfall at horizontal resolutions of 1 and 3 km, over coastal areas of southern Italy. All the above studies combined show that there is no single parameterization scheme that exclusively outperforms other schemes in the simulation of precipitation.

A significant portion of research has been devoted to the assessment of the quality of the model simulations, with the use of different model evaluation methods. The most basic form of quantitative verification of continuous gridded variables, such as precipitation for long accumulation times and temperature, is the computation of scalar measures, e.g., mean error (Wilks 2006). More elaborate methods of quantitative evaluation, generally known as spatial verification methods, e.g., "neighborhood" techniques and object-based techniques with focus on precipitation, were summarized by Gilleland et al. (2009). For the evaluation of discrete variables, e.g., occurrence of specific weather events or the distribution of continuous variables over selected classes, various skill scores, originating from contingency tables, have been developed over a period of more than 100 years (Stephenson 2000). Yet, it is not well defined which measures or methods are appropriate for the comparison of modeled data with observations for different applications (Murphy 1991; Wilks 2006; Gilleland et al. 2009). For this reason, adopting the major principle applied in the evaluation of hydrologic models, i.e., the use of multiple and comprehensive evaluation measures, which provide information on different aspects of the quality of the simulations (Gupta et al. 1998), could make the evaluation of simulated precipitation more complete.

Few studies have tested the effect of domain size, initialization frequency and physics parameterizations on the accuracy of precipitation simulations from reanalysis data at 1-km resolution (e.g., Zittis et al. 2017; Avolio and Federico 2018). The present study aims to identify a set of multiphysics configurations of the WRF Model that most closely capture the variable precipitation generating processes for high-resolution downscaling of hindcast precipitation simulations. The individual WRF Model simulations can be used as input for hydrological modeling applications for small and medium-size watersheds. The ERA5 reanalysis is dynamically downscaled with nested domain simulations to 1 km with the WRF Model. The model experiments focus on a small domain (about 230 km × 145 km) with complex topography, encompassing the island of Cyprus, located in the eastern Mediterranean. The specific objectives of this study are (i) to introduce a new evaluation measure for the assessment of model performance for extreme events, (ii) to select a model domain size and number of nested simulation domains from three different domain configurations, (iii) to evaluate the performance of 5- and 30-day model initialization frequencies, (iv) to create a subset of the most skillful physics parameterizations (members) from a composite, multiple-metrics evaluation of an initial 18-member set, comprised of various combinations of physics parameterizations, and (v) to evaluate the performance of multiphysics simulations for different horizontal model

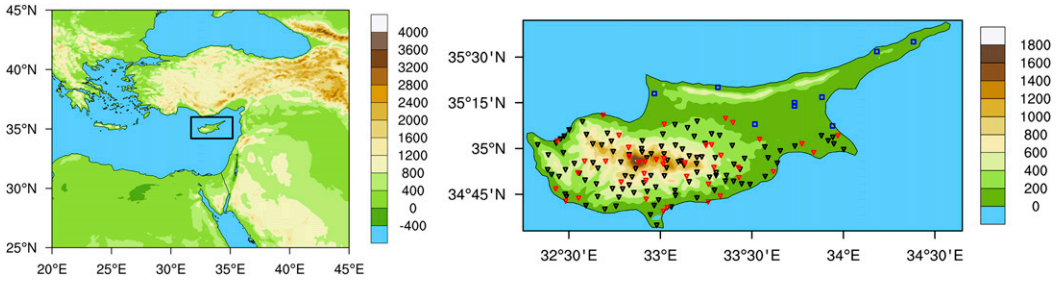

FIG. 1. (left) Elevation (m) and location of the study area in the eastern Mediterranean and (right) the 158 precipitation stations (53 automatic, red triangles; 85 manual gauges, black triangles of the Cyprus DoM; and the 10 SYNOP stations,blue rectangles) over the elevation map (m) of Cyprus.

resolutions. The numerical experiments cover a period of 8 months of the hydrometeorological year 2011–12.

2. Study area and observational datasets

a. Study area and climate

The simulations geographically focus on Cyprus, an island located in the eastern part of the Mediterranean Sea, approximately at latitude 35°N and longitude 33°E (Fig. 1). The climate of the island is characterized as Mediterranean. Precipitation over the island exhibits a substantial annual and interannual variability, which is typical for the Mediterranean region (Hoerling et al. 2012). The mean annual precipitation (1980–2010) ranges from between 240 and 400 mm in the lowlands to nearly 1100 mm in the Troodos Mountains, which is the main mountain range of the island (Camera et al. 2014). The steeply sloping Troodos mountain range covers the central part of the island and has its highest point at Mount Olympus at 1952 m above mean sea level. The complex terrain is responsible for the orographic enhancement of precipitation over the high-elevation regions.

b. Daily precipitation dataset CY-OBS

Precipitation observations at 148 manual and automatic stations from the Department of Meteorology of Cyprus (DoM), for the area under the control of the Republic of Cyprus, were used to create a gridded daily dataset of observed precipitation of Cyprus (CY-OBS). The dataset has spatial resolution of 1 km and covers the period 2011–12. Precipitation observations from 10 additional locations in the northern part of the island from 6-hourly synoptic observation (SYNOP) reports were also included. The locations of the precipitation measurements are shown in Fig. 1. The total precipitation over Cyprus for the hydrologic year 2011–12 from the CY-OBS dataset is shown in Fig. 2.

For the interpolation of the observed data, inverse distance weighting (IDW) was used for small-scale events and geographically weighted regression (GWR) for large-scale events, similar to Camera et al. (2014). GWR is a more complex formulation than the least squares regression, with the developed regression model varying in space, i.e., the weight of independent geographic variables of the regression can change from location to location. For the development of CY-OBS, the geographical variables used in the regression are the elevation, the distance to the coast, the east coordinate, the north coordinate, the distance from the main mountain ridge to the east, and the distance from the main mountain ridge to the west (Camera et al. 2014).

3. Study period, model setup, and evaluation approach

a. Stepwise evaluation approach

The dynamical downscaling was conducted with the WRF-ARW model (Skamarock et al. 2008) version 4.0. Initial and boundary conditions were provided by the ERA5 reanalysis dataset at 31-km resolution. Boundary conditions were updated every 6 h. The number of model vertical levels was set to 40. The length of the spinup period was set to 6 h. Comparison of simulated precipitation obtained with different spinup times that led to the choice of 6 h is provided in Fig. S5 in the online supplemental material.

A comprehensive, stepwise evaluation approach, illustrated in Fig. 3, was developed to evaluate the performance of the model to simulate precipitation in the study period from October 2011 to May 2012 (section 3c) for different model setup options and 18 combinations of model physics parameterizations (members) of the WRF Model (WRF18). For the stepwise methodology, the total simulation period of eight months was divided into two periods. Different model setups and configurations were tested in period 1 and the best-performing setup and five model parameterizations were

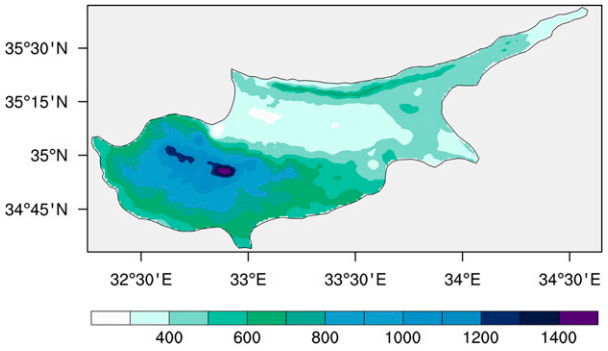

FIG. 2. Total precipitation (mm) over Cyprus for the hydrologic year October 2011–September 2012 from the CY-OBS dataset.

Simulation experiments	# experiments	Calibration period
Step 1 **3 domain setups** × 1 initialization × 18 members × 1 month	54	Jan 2012
Step 2 1 domain setup × **2 initializations** × 18 members × 2 months	72	Jan 2012, May 2012
Step 3 1 domain setup × 1 initialization × **18 members** × 3 months	54	Oct 2011, Jan 2012, May 2012
		Validation period
Step 4 1 domain setup × 1 initialization × 5 members × 5 months	25	Nov2011, Dec2011, Feb2012, Mar 2012, Apr 2012

Model configurations tested

Domain setup	Initialization frequency	Physics parameterisations
12-4-1 6-1a 6-1b	5-days 30-days	18 members

FIG. 3. The stepwise approach followed for the selection of WRF setup options and physics parameterizations for precipitation simulations.

subsequently evaluated in period 2. These two periods can be considered equivalent to calibration and validation periods.

In step 1 of the stepwise approach, three domain configurations were tested for 31-day continuous simulations of January 2012 for all 18 members. The boundaries of the three domain configurations are shown in Fig. 4, while the domain dimensions and time step used in each domain configuration are shown in Table 1. The three domain configurations are a 12–4–1 configuration with three downscaling steps, at 12-, 4-, and 1-km resolution and 6–1a and 6–1b configurations with two downscaling steps, both at 6- and 1-km resolution. All configurations used one-way nesting. The domain configuration that achieved the better skill was selected for use in the next step of the stepwise approach. The details of the evaluation method are described in section 4c.

In the second step of the evaluation approach, the effect of 5- and 30-day model initialization frequencies was investigated for the 18 members, for January and May 2012. The differences in the performance of each member between the two initialization approaches were evaluated likewise to the three domains setups in step 1. The initialization frequency with the best performance, based on the same evaluation methods used for the domain configurations, was selected for step 3.

In the third step of the stepwise approach, monthly simulations with the selected model domain configuration and initialization frequency were conducted with the 18 members for October 2011, thereby completing the simulations of the calibration period (October 2011, January 2012, May 2012) (Fig. 3). The 18 members were ranked for four model evaluation measures and for a composite scaled score (see section 4b) for the entire calibration period. Five members were selected,

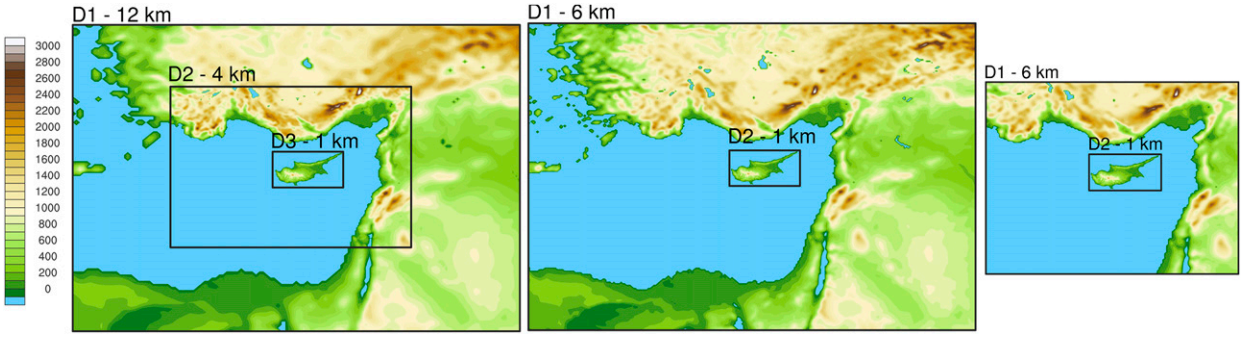

FIG. 4. (left) Elevation maps (m) and boundaries of the three-nested WRF domains with horizontal resolutions of 12 km (domain 1), 4 km (domain 2), and 1 km (domain 3) and (center),(right) of the two two-nested WRF domains with horizontal resolutions of 6 km (domain 1) and 1 km (domain 2). The domain sizes are presented in Table 1.

TABLE 1. Model domain dimensions, resolution, and time step for the three domain configurations.

Domain configuration	Domain 1	Domain 2	Domain 3
12–4–1			
Spatial resolution (km)	12	4	1
Domain size (km^2)	1488 × 1248	796 × 652	225 × 145
Time step (s)	72	24	6
6–1a			
Spatial resolution (km)	6	1	
Domain size (km^2)	1488 × 1248	235 × 145	
Time step (s)	36	6	
6–1b			
Spatial resolution (km)	6	1	
Domain size (km^2)	826 × 768	235 × 145	
Time step (s)	36	6	

the first four based on best performance for each of the four evaluation measures and the fifth member based on the composite scaled score of the four measures. The performance of the selected members was then evaluated with simulations of the five months of the validation period (November 2011, December 2011; February 2012, March 2012, April 2012) in the fourth step of the stepwise approach (Fig. 3).

b. *Physics parameterizations*

The WRF physics parameterizations were selected based on their performance in previous investigations over the study area (Zittis et al. 2017; Tymvios et al. 2018). The different members are all based on the same model configuration, except of the parameterization schemes. Table 2 lists the used parameterization schemes and Table 3 presents the combinations of different parameterization schemes, i.e., surface layer, PBL, cumulus physics (convection) and microphysics, resulting in 18 members. Microphysics and cumulus physics are parameterized with three options each. The parameterizations of shortwave and longwave radiation and the land surface model are the same for all 18 members (Table 2).

c. *Selection of the study period*

The study period consists of the eight wettest months of the hydrologic year 2011–12, from October 2011 to May 2012. This year was selected because the total precipitation was 31% higher than the long-term average precipitation over Cyprus with a larger number of precipitation events to simulate.

The length of the model calibration period was limited to three months to make efficient use of computational resources for the testing a comprehensive set of domain configurations, initialization frequencies and physics parameterizations. The type and number of precipitation events that occurred in these months of the selected hydrologic year are representative of the events that could occur throughout any hydrologic year. For the evaluation of domain configurations (step 1) only simulations in January were performed, because this month contained a substantial number of rainfall events that are affected by large-scale phenomena. Simulations in two months, January and May, were performed for the evaluation of initializations (step 2) because the impact of initialization frequency may be different for convective precipitation (May) than for large-scale winter precipitation (January) (Lucas-Picher et al. 2013). The evaluation of the 18 members (step 3) was done for three months, which include large-scale precipitation events (January) and events related to strong convective conditions (May) and weak convective conditions (October). A similar range of atmospheric conditions and rainfall events occurred in the five months of the validation period, with almost equal total precipitation over Cyprus, i.e., 250 mm in the calibration period and 283 mm in the validation period.

4. Evaluation measures

a. *New deterministic verification score for categorical variables*

Various simple and complex skill scores, derived from the counts of simulated and observed event pairs, are used as accuracy measures for discrete variables, e.g., number of events that fall within a specified class (Stephenson 2000). The 2 × 2

TABLE 2. The physics parameterization schemes selected for the WRF precipitation simulations over Cyprus.

Parameterization type (No. of schemes)	WRF parameterization name
Longwave radiation (1)	RRTM longwave scheme (Mlawer et al. 1997)
Shortwave radiation (1)	Dudhia shortwave scheme (Dudhia 1989)
Surface layer (2)	MM5 similarity scheme (MM5; Zhang and Anthes 1982)
	Eta similarity (Eta; Janjić 1994)
Land surface model (1)	Noah land surface model (Noah LSM; Tewari et al. 2004)
Planetary boundary layer (2)	Yonsei University scheme (YU; Hong et al. 2006)[a]
	Mellor–Yamada–Janjić (MYJ; Janjić 1994)[b]
Microphysics (3)	WRF single-moment 6th class (WSM6; Hong and Lim 2006)
	WRF double-moment 6th class (WDM6; Lim and Hong 2010)
	Ferrier (Rogers et al. 2001)
Cumulus (3)	Kain–Fritch (KF; Kain 2004)
	Betts–Miller–Janjić (BMJ; Janjić 1994)
	Grell–Freitas (GF; Grell and Freitas 2014)

[a] The Yonsei University PBL scheme can only be used in combination with the MM5 surface layer scheme.
[b] The Mellor–Yamada–Janjić PBL scheme can only be used in combination with the Eta similarity surface layer scheme.

TABLE 3. The combinations of different schemes of four types of physics parameterizations used for the 18 WRF members.

Member	Microphysics[a]	Cumulus[b]	PBL[c]	Surface layer[d]
T1	6	2	2	2
T2	6	2	1	91
T3	6	1	2	2
T4	6	1	1	91
T5	6	3	2	2
T6	6	3	1	91
T7	5	1	1	91
T8	5	1	2	2
T9	5	3	2	2
T10	5	3	1	91
T11	5	2	2	2
T12	5	2	1	91
T13	16	2	1	91
T14	16	2	2	2
T15	16	1	1	91
T16	16	1	2	2
T17	16	3	1	91
T18	16	3	2	2

[a] Microphysics schemes: Ferrier (5), WSM6 (6), WDM6 (16).
[b] Cumulus physics schemes: KF (1), BMJ (2), GF (3).
[c] Planetary boundary layer physics schemes: YU (1), MYJ (2).
[d] Surface layer physics schemes: Eta similarity (2), MM5 similarity scheme (91).

contingency table (Fig. 5) contains the four components, i.e., hits (a), false positives (b), misses (c), and correct nonevents (d), which these skill scores are generally based on. A new extreme event score (EES), based on three of these four components, is proposed here for the evaluation of extreme events in deterministic model simulations. The new measure combines into a single value two simple scalar attributes of the 2×2 contingency table, namely, the hit rate (H) and the frequency bias (BIAS), which are frequently used for the evaluation of categorical variables (Wilks 2006).

The equations of H and BIAS are

$$H = \frac{a}{a+c}, \tag{1}$$

$$\text{BIAS} = \frac{a+b}{a+c}. \tag{2}$$

The proposed EES is given by

$$\begin{aligned} \text{EES} &= H \times \text{BIAS}, \quad \text{for} \quad \text{BIAS} \le 1, \\ \text{EES} &= H \times \text{BIAS}^{-1}, \quad \text{for} \quad \text{BIAS} > 1. \end{aligned} \tag{3}$$

The H and the BIAS are incorporated in the definition of EES in Eq. (3), as the first and the second multiplicative term, respectively. Thus, EES, through H, determines the ability of the model to yield a number of correctly simulated extremes, relative to the total number of extremes; through the second term, which corresponds to BIAS or BIAS^{-1}, EES penalizes the model proportionally with the degree of underestimation of the number of extremes (BIAS < 1) or overestimation of the number of extremes (BIAS > 1). The value of EES ranges from 0 (no skill), when no hits are achieved by the model, to 1 (perfect skill), when the model is unbiased and correctly predicts all events.

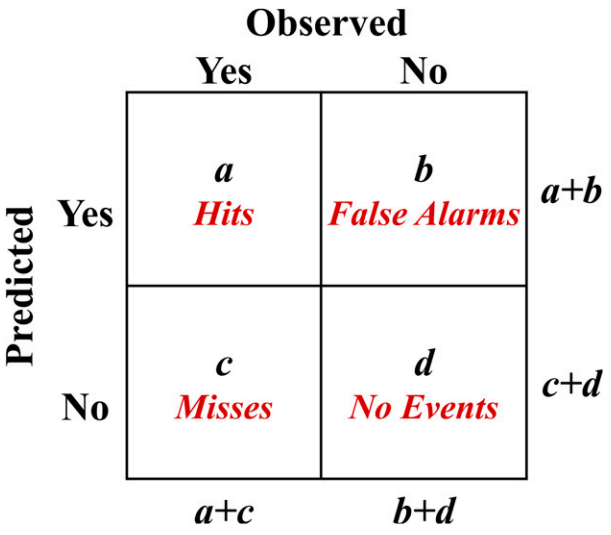

FIG. 5. The 2×2 contingency table with the frequencies of possible outcomes, a, b, c, and d of a predicted (simulated) variable, in relation to the observations.

The proposed EES was compared against three other skill scores and three scalar attributes of the contingency table that can together describe the full joint distribution of any 2×2 contingency table (Wilks 2006). The skill scores are the Peirce skill score (PSS; Peirce 1884) also known as Hanssen–Kuiper's discriminant, the equitable threat score (ETS), also known as Gilbert skill score (Gilbert 1884) and the recently proposed extreme dependency score (EDS; Stephenson et al. 2008). The three scalar attributes of the contingency table are H, false alarm rate (F), and BIAS. The equations for these scores are available in supplemental material section 1. More details can be found in Hogan and Mason (2012). All scores were computed for the precipitation simulations from the 18 multiphysics runs for January and May 2012 with hits counted when daily precipitation exceeds 30 mm at any grid cell of the WRF domain. The properties of the EES were further evaluated against the same three skill scores for a large number of random contingency tables (see supplemental material section 2). The equitability of the EES was also tested, following the method described by Hogan et al. (2010). Two contrasting scores were selected for inclusion into the set of the evaluation measures (section 4b) that were thereafter used to evaluate the performance of the different WRF setup options and parameterizations.

b. *Evaluation measures for continuous variables and composite score*

In addition to the categorical scores used to evaluate the model's performance for extreme precipitation events, two measures were selected to evaluate errors in the volume of precipitation. Considering the simulated precipitation as a continuous variable on the gridded WRF domain, the two measures are the total period Bias (mm) and the mean absolute

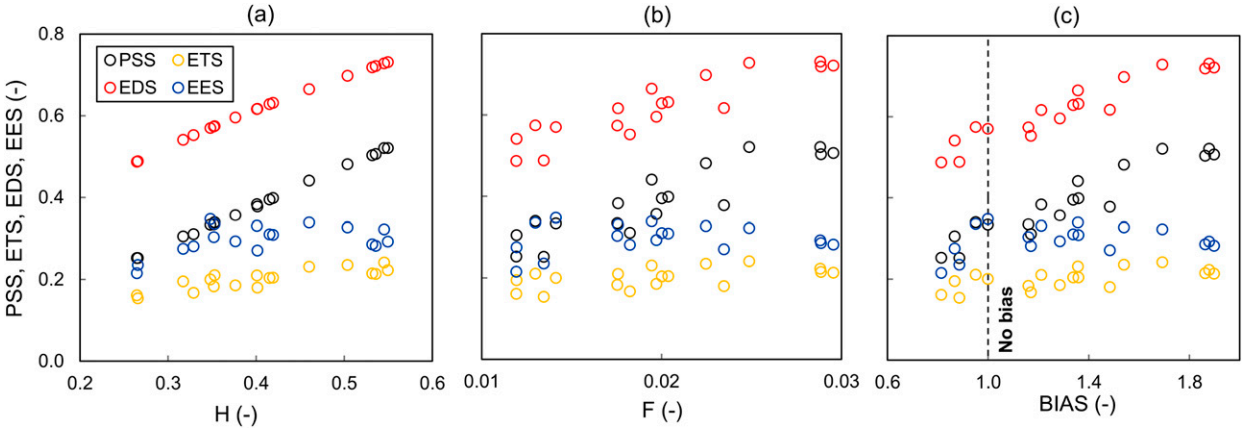

FIG. 6. Scatterplots of Peirce skill score (PSS), equitable threat score (ETS), extreme dependency score (EDS), and extreme event score (EES) against (a) hit rate (H), (b) false alarm rate (F), and (c) frequency bias (BIAS). The scores are computed for daily precipitation events above 30 mm, for the 18-member simulations of January and May 2012.

error (MAE) (mm day^{-1}) of daily precipitation. Total Bias, which is different from the frequency "BIAS" derived from the contingency tables, and MAE are first computed over all grid cells and all days. Total Bias is subsequently summed for all days and averaged over all grid cells whereas MAE is averaged over both the number of days and the number of grid cells. The former measure gives the magnitude and sign of error in total precipitation in a month or within the simulation period and the latter measure evaluates the mean model performance on a daily basis across all precipitation regimes. The different properties of the simulations evaluated by the selected measures are all important aspects, relevant to the water balance, in hydrological modeling.

Adapted from Young (2006), a composite scaled score (CSS_i), which combines the values of Bias, MAE, and the two categorical scores for extreme events into a single score, was applied in the current study to rank the 18 members:

$$\mathrm{CSS}_i = \frac{1}{N_s}\sum_{s=1}^{N_s}\left(\frac{x_{s,i} - x_{s,\mathrm{worst}}}{x_{s,\mathrm{best}} - x_{s,\mathrm{worst}}}\right), \tag{4}$$

where i is the index identifying the member, s is the index of the statistical measure out of a number of N_s (four) measures, $x_{s,i}$ is the value of measure s obtained by member i, and $x_{s,\mathrm{worst}}$ and $x_{s,\mathrm{best}}$ are the worst and the best values for measure s among all 18 members. The proposed score ranges between 0 and 1. If a member achieves the best values for all evaluation measures among all members, then its CSS is equal to 1. The CSS can effectively rank different configurations because it combines in a single value multiple measures that evaluate various desired properties of the precipitation simulations.

c. Statistical analysis

To assess the uncertainty in the obtained values of total Bias and MAE and the two categorical measures, the bootstrapping method was applied (Efron and Tibshirani 1993). The results of the model evaluation measures described above are given by the median value of the sampling distribution of each measure, obtained with bootstrapping. The bootstrap was applied 1000 times on each sample, consisting of daily gridded precipitation maps (WRF output and CY-OBS). The sample had a size equal to the number of days in the simulation period. For instance, to obtain the distribution of one evaluation measure for one member in January, the 31 pairs of daily observed and simulated precipitation maps were randomly drawn with replacement 1000 times and the value of a measure was computed 1000 times. The 90% confidence interval (CI), i.e., the interval between the 5% lower and 95% upper level of the sampling distribution of each evaluation measure, was used as an estimate of the uncertainty in the values of the measures.

For the investigation of the improvement in the model performance of each member, with the 12–4–1 domain configuration relative to the 6–1a or 6–1b configuration, or with the 5-day initialization relative to the 30 day, the method described in Hogan and Mason (2012) was used to evaluate the median difference in the values of each evaluation measure between the compared configurations or initializations. The difference

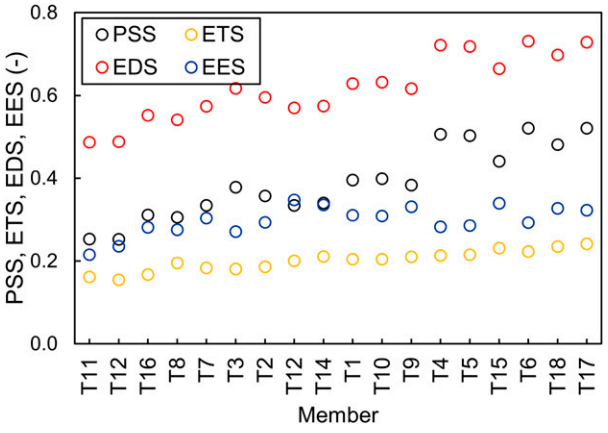

FIG. 7. Values of the extreme event scores (see Fig. 6), sorted in increasing order of the composite scaled score [Eq. (4)] of the 18 WRF members.

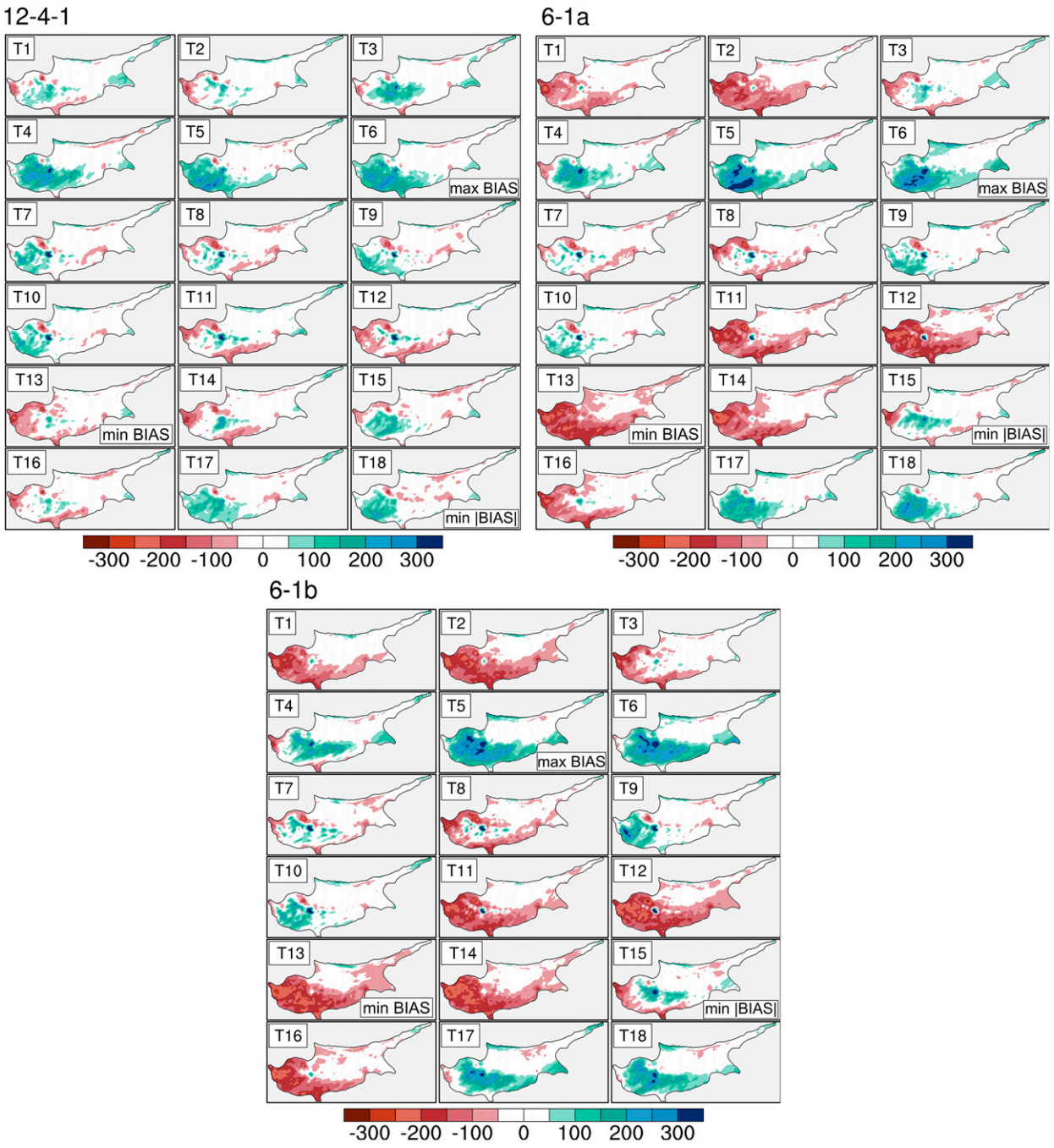

FIG. 8. Bias (mm) of WRF simulated total precipitation for January 2012 for the 18 WRF members and 30-day initialization, for the three different domain configurations: (a) 12–4–1, (b) 6–1a, and (c) 6–1b. Three labels indicate the members with the spatially average maximum positive (max BIAS), the maximum negative (min BIAS), and the least bias (min |BIAS|).

in the value between two domain configurations or the two initializations for each member was bootstrapped as described above. The improvement or degradation of the performance of a member is considered statistically significant with the one type of domain configuration or initialization relative to the other, when the bootstrapped difference of a particular evaluation measure is different from zero at the 90% CI.

5. Results and discussion

a. Evaluation of the new EES and selection of categorical scores

The EES was found equitable for base rates (observational rate of occurrence) less than or equal to 0.05 with any sample size (sum of the four elements of the contingency table).

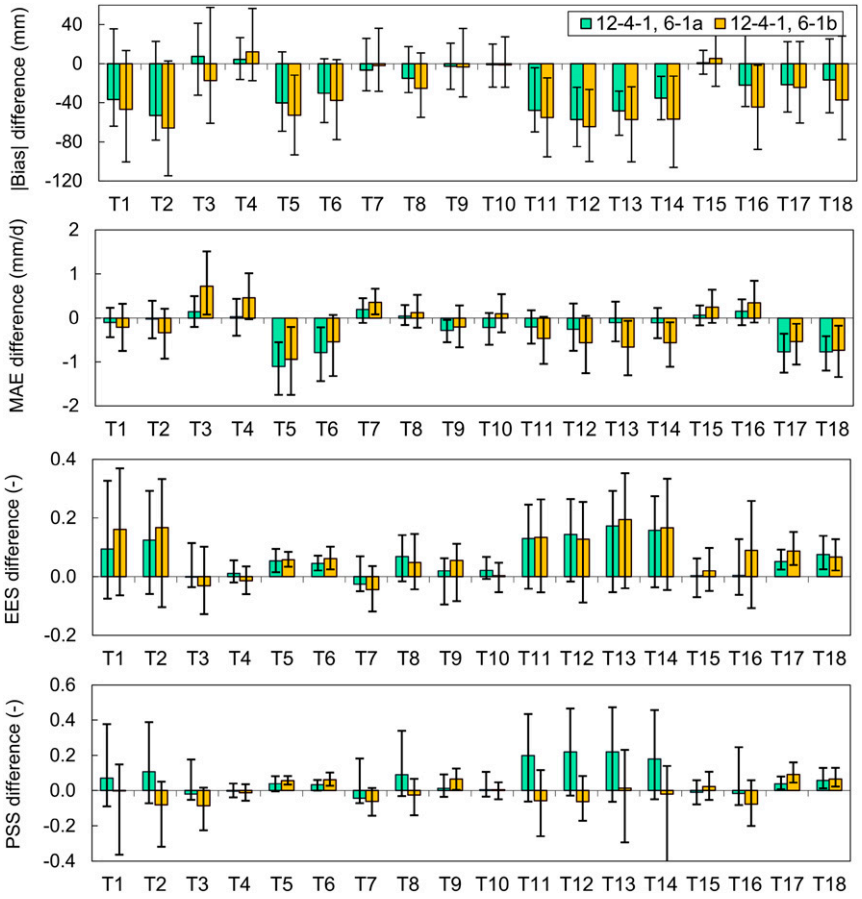

FIG. 9. Differences between the values of four evaluation measures (|Bias|, MAE, EES, and PSS) of the simulations performed with the 12–4–1 and 6–1a (green) and 12–4–1 and 6–1b (yellow) domain setups and the 90% confidence intervals for January 2012 for all 18 WRF members.

Thus, EES can be considered a suitable evaluator for extremes. The equitability of ETS depends on the sample size, which must exceed 30. The EDS is equitable for base rates equal or larger than 0.1 with sample size larger than 1000. The PSS is the only truly equitable score, among the four tested scores, i.e., equitable for all base rates and sample sizes. The property of equitability as well as the comparison of the EES against other skill scores for a large number of random contingency tables are visualized in sections 2a and 2b of the supplemental material.

The functioning of the proposed EES in the WRF precipitation simulations of this study is illustrated in Fig. 6 with three scatterplots of values of the EES and the three other skill scores against the *H*, *F*, and the BIAS. The scatterplot of PSS, ETS, EDS and EES against *H* in Fig. 6a reveals a linearity in the relationship of *H* with each of the PSS and EDS. The former relationship is rather obvious when looking at the formulation of PSS, which is the *H* minus *F*. The latter relationship can be deduced from the formulation of EDS, which depends on the logarithms of the hits, while *H* linearly depends on the number of hits [Eq. (1)]. The ETS and the EES vary less than 0.1 for all 18 members and their relative change with *H* is smaller than the relative change of PSS and EDS with *H*. The relationship of PSS and EDS against *F*, in Fig. 6b, is not linear, yet the two scores exhibit a general increase with increasing *F*, while the values of ETS and EES again exhibit a very small variation. Similar findings for the relationship of the four scores with *H* and *F*, can also be seen in Fig. 6c against BIAS.

A further comparison of the values of the four scores, plotted in increasing order of the composite scaled score for the 18 WRF members, is presented in Fig. 7. The figure shows that the choice of either PSS or EDS, both accounting for the number of hits, would result in the same ranking of different simulations. We selected the truly equitable PSS instead of the less common and not truly equitable EDS. The comparison of EES with ETS, both accounting for hits and BIAS, reveals that ETS has lower variability than EES, indicating that EES provides more information on the relative performance of different simulations than ETS. Thus, the proposed EES was selected as the second score for the evaluation of extremes.

b. WRF ensemble selection

1) EFFECT OF DOMAIN CONFIGURATION

The gridded monthly Bias for each of the 18 multiphysics WRF members (Fig. 8) with the three domain configurations

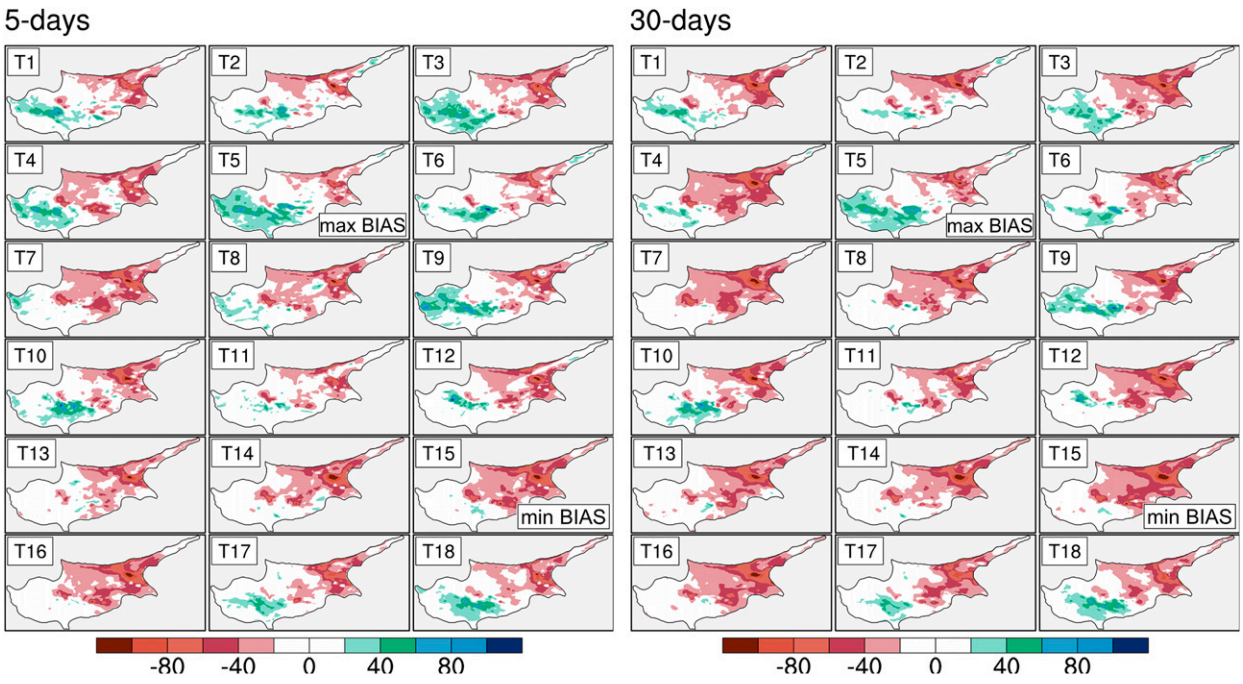

FIG. 10. Bias (mm) of WRF simulated total precipitation for May 2012 for the 18 WRF members for (left) the 5-day initialization and (right) the 30-day continuous run. Similar to Fig. 8, two labels indicate the members with the spatially average minimum (min BIAS) and maximum bias (max BIAS). The member with the least bias coincides with the member with the maximum bias.

shows spatial patterns of underestimation or overestimation over the island. The largest Bias occurs in the wettest part of the island, i.e., the southwest, which is the windward side of Troodos Mountains. On average, the 12–4–1 configuration stands out with consistently lower Bias for all 18 members compared to the Bias of the members with the 6–1a and 6–1b configurations. Underestimations, e.g., for members T11, T12, T13, and T14, and the overestimations, e.g., for members T5, T6, and T17, range from −18 to −45 mm and from 17 to 45 mm, respectively, in the 12–4–1 configuration. The monthly Bias of the corresponding members in 6–1a ranges from −69 to −93 mm and from 41 to 76 mm, and in 6–1b from −77 to −102 mm and from 48 to 84 mm, respectively.

The performance of each member of the domain configuration with the largest extent and three nested domains (12–4–1) versus that of the two nested-domain configurations (6–1a and 6–1b) is indicated by the difference in the values of four evaluation measures in Fig. 9. Comparison of the differences in the value of total Bias in absolute values (|Bias|) confirms the relative higher skill of the 12–4–1 configuration. The values of total |Bias| obtained with the 12–4–1 domain minus the value obtained with the 6–1a (6–1b) domain is negative, almost up to −80 mm, in 15 (16) out of 18 members, indicating the lower total |Bias| achieved with the 12–4–1 setup. A similar comparison of the median differences of MAE, EES, and PSS between domains, member by member, is also shown in Fig. 9. The comparison reveals that MAE is on average lower, up to 1 mm, with 12–4–1 and that EES and PSS scores, computed for the threshold of 30 mm, are on average higher, up to 0.2, with 12–4–1 domain.

Considering uncertainty, by the 90% CI, it becomes less evident whether or not there is statistical significance in the difference among the different domains, i.e., whether the 90% CI of the median difference is above or below zero. Yet, for the total |Bias| there are four (six) members, i.e., T11, T12, T13, and T14 (T5, T11, T12, T13, T14, and T16) with statistically significant differences favoring the 12–4–1 against 6–1a (6–1b), relative to zero (zero) statistically significant cases favoring the 6–1a (6–1b) against 12–4–1. Likewise, for MAE there are five (six) members favoring the 12–4–1 against 6–1a (6–1b), relative to zero (two) members favoring the 6–1a (6–1b) against 12–4–1. For EES and PSS, the 12–4–1 achieves significantly better results than 6–1a (6–1b) for four (five) members.

The 12–4–1 configuration achieves the overall higher number of significantly improved measures of total |Bias|, MAE, EES, and ETS than any of the other two domain configurations and it is selected for the following simulations of the stepwise approach. The selected domain configuration has a similar extent and number of nested domains as the domain configurations used in other studies over Cyprus (Zittis et al. 2017; Tymvios et al. 2018). These studies used WRF to downscale the ERA-Interim reanalysis data (≈79 km) and the NCEP-GFS analysis (≈56 km), both at a lower resolution than ERA5 reanalysis used in this study (≈31 km). Still, with the high-resolution LBCs of ERA5, the two less computationally demanding domain configurations, 6–1a and 6–1b, could not reproduce precipitation over the island with the same skill as the 12–4–1 domain configuration. These results support the recommendation of Vannitsem and Chomé (2005), at least for the small geographic area of this study. These authors suggested that computational savings and improved simulation

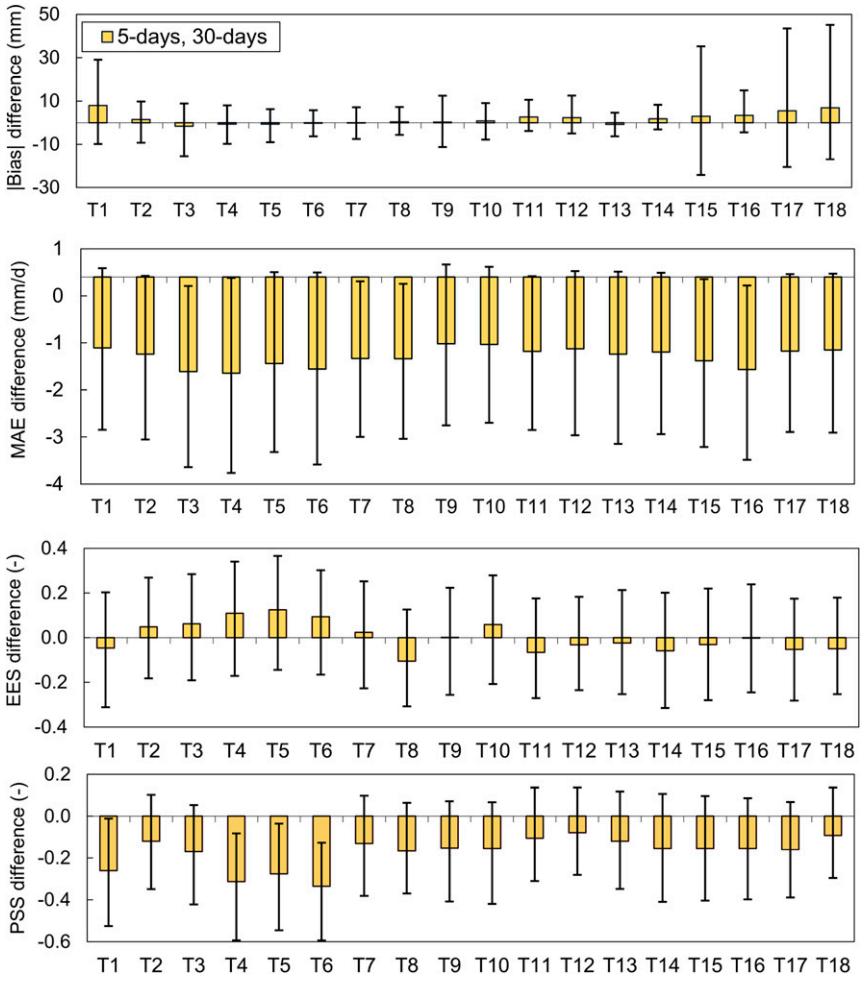

FIG. 11. Differences between the values of four evaluation measures (|Bias|, MAE, EES, and PSS) of the simulations performed with the 5-day initialization and 30-day initialization and the 90% confidence intervals for January and May 2012 for all 18 WRF members.

results can be achieved by an increase in the resolution of LBCs combined with a reduction in the size of the domain. This improvement is, however, seen for RCM applications with typical domains sizes on the order of few thousand kilometers or larger and not with smaller domains size, according to the authors.

2) EFFECT OF INITIALIZATION FREQUENCY

The spatial distribution of the Bias of the precipitation for May 2012 of the 18-member simulations initialized with two different frequencies are shown in Fig. 10. Similar to January (maps not shown), large difference in the spatial Bias between the 5-day and 30-day initializations is not evident. The simulations for May exhibit, on average, smaller absolute but larger relative Bias than January. The Bias in May is mainly negative, compared to both positive and negative Bias for different members in January, and is found over the northern and eastern part of the island. Nonetheless, there are particular members, e.g., T3 and T5, which exhibit additionally large positive Bias on the southwestern part of the island for both initializations. The average Bias over the domain ranges from −0.4 mm (T5) to −29 mm (T15) for the 30-day continuous simulation, and from −2 mm (T9) to −24 mm (T15) for the 5-day initialized simulations.

A similar member by member and evaluation measure by evaluation measure comparison of the model performance with the two initializations, as with the three different domains, for the results of January and May combined is shown in Fig. 11. In terms of the difference in total |Bias| between 5-day and 30-day, on average, the |Bias| is larger with the 5-day initialization, with the difference of the 30-day |Bias| from the 5-day |Bias| ranging from −1 to 8 mm. Unlike total |Bias|, the MAE is higher with the 30-day initialization for all 18 members, in the entire 90% CI. Differences of EES between 5-day and 30-day do not favor any of the two initializations. The PSS has negative differences for all members, which implies that PSS obtained with the 30-day initialization is higher (better model skill) than that obtained with the 5-day initialization. This difference is significant for four members (T1, T4, T5, T6). Based on these results, the 5-day initialization, which requires about 4% more computational resources than the 30-day continuous runs due to the additional time for spinup with

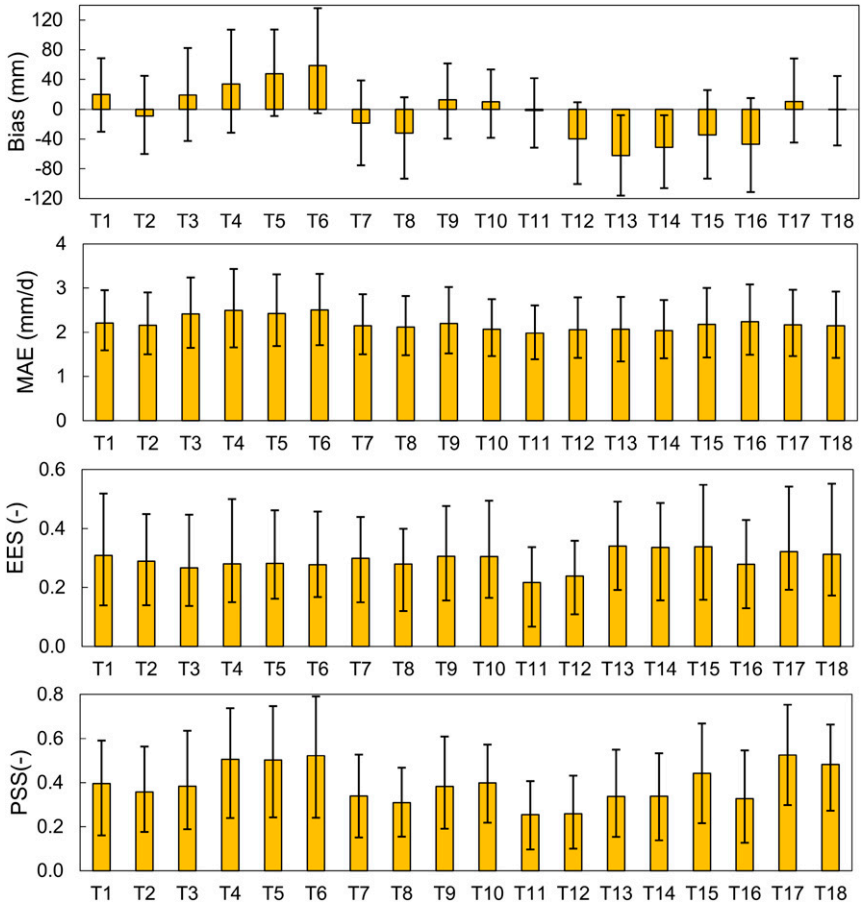

FIG. 12. Median values and 90% confidence intervals of four evaluation measures (Bias, MAE, EES, and PSS) of the simulations performed with 30-day initialization and the 12–4–1 domain configuration for October 2011 and January and May 2012 for all 18 WRF members.

each initialization, is found not to add value to the simulations, relative to the 30-day continuous run.

These results do not fully agree with the conclusions of other studies, which were, however, obtained for much larger domain sizes (Seth et al. 2004; Berckmans et al. 2017). The improved spatial and temporal correlations between simulated and observed 6-hourly precipitation obtained with 7-day versus a 30-day initialization by Lo et al. (2008) and for daily precipitation with 1-day initialization versus multiyear continuous simulations by Lucas-Picher et al. (2013) do not compare with the results for 5-day initialization in the small domain of this study. These findings suggest that the small size of a domain, such as the size of the outer domain in this study (1500 km × 1200 km), compared to the larger domain sizes of typical RCM applications, allows the control of the LBCs to be strong enough on the simulated atmospheric features, so that the positive impact of more frequent initialization does not exceed the impact of LBCs.

3) EVALUATION AND SELECTION OF WRF MEMBERS

The four evaluation measures of the simulations with the 18 members with the 12–4–1 domain configuration and the 30-day initialization, for the months of October, January and May combined, are shown in Fig. 12. The average precipitation of the three months over Cyprus is both overestimated and underestimated. The highest total positive Bias of the three months was obtained with T6 (65 mm) and the highest negative Bias was obtained with T13 (−48 mm). The lowest Bias was obtained with T11 and T18, which both achieve low Bias in each of the three months (not shown here). Further comparison of the members with the median values of MAE, EES, and PSS reveals little information on the best performing members. Overall, there are small differences in the median values of the four evaluation measures and the 90% CIs indicate the large uncertainty in these measures for all members.

The CSS and the scaled scores (SS) of the four performance measures, which are the four components of CSS, are presented in Fig. 13. Members T18, T11, T13, and T17 have the highest scores for total Bias, MAE, EES, and PSS, respectively. The highest CSS (0.8) is achieved by members T15, T17, and T18. Thus, T15 is selected as the fifth member. The use of the different criteria ensures that the selected members capture together all properties of the

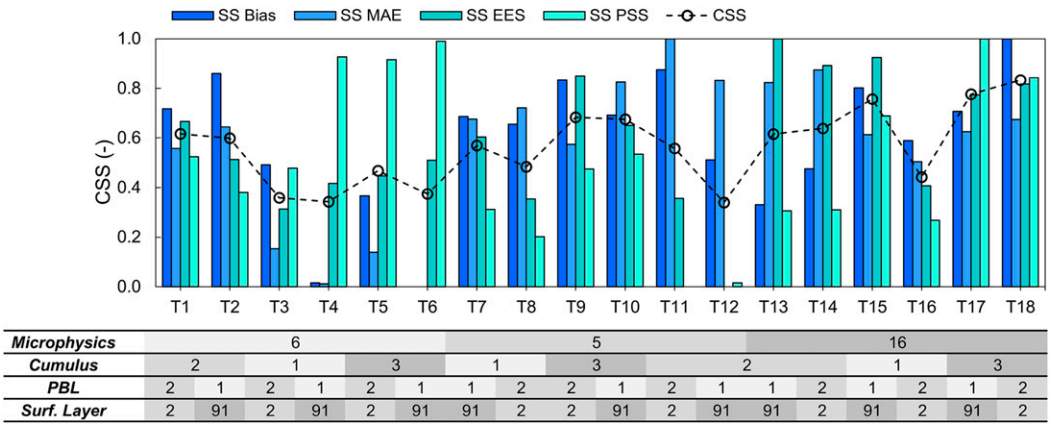

FIG. 13. Scaled scores (SS) of the four evaluation measures (Bias, MAE, EES, and PSS) and the composite scaled score (CSS), for the 3-month calibration period (October 2011, January 2012, and May 2012), for the 18 WRF members and the codes of the parameterizations of each member displayed below.

simulations, i.e., the volume of daily and monthly precipitation as well as the extreme amounts, better than any other combination of members. The most evident impact of a physics parameterization type on the obtained CSS of the 18 members is that of the microphysics parameterizations (Fig. 13). Members with WDM6 (16) microphysics scheme achieve a CSS of 0.68, with Ferrier (5) a CSS of 0.55, while four out of the six members that use the WSM6 scheme (6) are all ranked last, with an average CSS of 0.46 for the six members. The performance of the three different cumulus schemes, seen from the CSS of the members based on these schemes also exhibits a trend. Members with GF (3) obtain an average CSS of 0.64, with BMJ (2) a CSS of 0.56 and with KF (1) a CSS of 0.49. These results indicate that the different cumulus schemes affect the precipitation in the 1-km resolution domain on which the evaluation is made, even though this type of scheme is turned off in this domain. The average CSS of the members using the two pairs of the PBL/surface layer parameterization schemes YU (1)/MM5 (91) and MYJ (2)/Eta (2) schemes have CSS of 0.56.

The above findings on the cumulus and microphysics parameterizations are generally in line with the findings of Zittis et al. (2017) for extreme events over the same area. These authors showed that the Ferrier and WDM6 schemes performed best out of five microphysics schemes and the BMJ scheme performed best out of four cumulus schemes for simulations on a 4-km grid. Hong et al. (2010) showed that the WDM6 scheme outperforms WSM6 for convective precipitation over different regions, which agrees with the comparison of the two microphysical schemes adopted here. Katragkou et al. (2015) and Avolio and Federico (2018) found that the bias in the long-term precipitation over Europe and the bias of extreme precipitation in short-duration events in southern Italy were less with the BMJ than with the KF scheme, in line with the results of the current study. For a one-year simulation period and a 5-km grid over Greece, Politi et al. (2018) found that WSM6 performed better than Ferrier, contrary to the findings herein. Jeworrek et al. (2019) identified the GF cumulus scheme as the best performing among other cumulus schemes for convective precipitation simulations over Southern Great Plains, similar to the finding here.

c. *Validation of the selected five members (WRF5) performance*

The values of the four evaluation measures for the selected five members (T11, T13, T15, T17, T18) in the 3-month calibration period are compared against the values of the measures for the same members in the 5-month validation period in Fig. 14. The sign of percent Bias, but not the magnitude, in the two periods, is interestingly the same between the same members in the two periods. In absolute values, the members have together a larger Bias in the validation period. This is most likely due to the higher total precipitation in the validation period (283 mm against 250 mm). The median value of MAE is lower for all members in the validation period, with the 90% CIs overlapping for all members in the two periods. The comparison of the obtained median EES and PSS shows that nearly all members perform better in the validation period, with the 90% CIs also, on average, larger. The higher values of EES and PSS in the validation period can be related to the greater number of extreme event occurrences, above the 30 mm day^{-1} threshold, in the validation period (≈14 000) relative to the calibration period (≈13 000). The validation results indicate that the five selected members have retained their overall performance from the calibration period.

d. *Effect of horizontal resolution*

The impact of the horizontal resolution of the three domains of the 12–4–1 domain configuration on the average precipitation simulated by the selected five members (T11, T13, T15, T17, T18) over the area of Cyprus is seen in Table 4. The simulations at 1-km resolution are the wettest compared to the coarser resolutions, but are generally closer to the

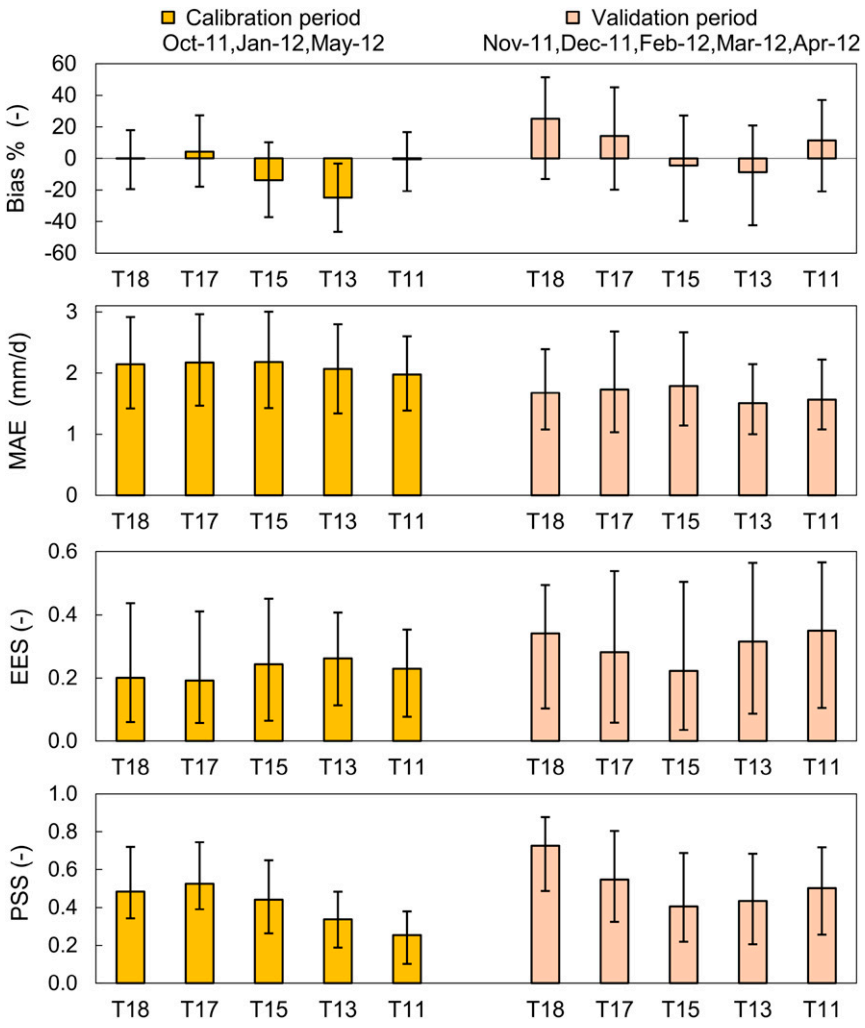

FIG. 14. Median values of four evaluation measures (percent Bias, MAE, EES, and PSS) and 90% confidence intervals for the selected five WRF members for the 3-month calibration period and the 5-month validation period.

observations, regarding monthly amounts. Percent Bias below ±10% is observed with the highest-resolution grid in four out of eight months, as well as in the total of the eight months. The percent Bias is less than ±20% with the 4-km and the 12-km grid only in January. For five out of eight months, the least Bias is achieved with the 1-km grid, for one month with the 4-km grid (May) and for three months with the 12-km grid (November, April, and May). In the months October, April, and May with low total precipitation, i.e., less than 40 mm, the total precipitation values can be more indicative of the model performance than the percent Bias. The absolute Bias ranges from 7 to 17 mm in these months. For November, neither the absolute Bias (40 mm) nor the percent Bias (48%) favors the 1-km grid and best results are obtained with the 12-km grid. From the above, it is evident that the 1-km grid outperforms the coarser resolutions. These results on the model horizontal resolution agree with findings of other studies, which highlighted the added value of the high-resolution convection-permitting simulation results, particularly for small spatial and temporal scales, for extreme events and for areas with steep orography (Chan et al. 2014; Cassola et al. 2015; Prein et al. 2015; Zittis et al. 2017).

6. Summary and conclusions

A comprehensive stepwise evaluation approach has been implemented to select the model domain configuration, initialization frequency and best performing physical parameterization schemes for precipitation simulations with the WRF Model over a small domain, centered on Cyprus. A new model evaluation score, the EES, was proposed for evaluating categorical variables, and particularly extreme events. The EES was applied for daily amounts that exceed 30 mm. The comparison of EES with other commonly used verification scores showed that the EES is a complete evaluator of the simulation of extremes, accounting for hits and bias in a single value. The

TABLE 4. Mean values of monthly and total period precipitation over Cyprus and percent bias from the selected five members for 8 months of simulations for domain resolutions of 12, 4, and 1 km.

		WRF5		
	CY-OBS (mm)	12 km	4 km	1 km
October 2011				
Monthly mean (mm)	13	22	24	20
Percent bias (%)		68	86	54
November 2011				
Monthly mean (mm)	81	105	107	120
Percent bias (%)		29	32	48
December 2011				
Monthly mean (mm)	66	47	42	69
Percent bias (%)		−28	−37	5
January 2012				
Monthly mean (mm)	200	166	172	198
Percent bias (%)		−17	−14	−1
February 2012				
Monthly mean (mm)	85	61	68	81
Percent bias (%)		−28	−20	−5
March 2012				
Monthly mean (mm)	33	22	26	31
Percent bias (%)		−32	−22	−7
April 2012				
Monthly mean (mm)	17	8	7	7
Percent bias (%)		−55	−59	−56
May 2012				
Monthly mean (mm)	37	23	23	20
Percent bias (%)		−37	−37	−45
8 months total				
Mean (mm)	532	455	469	546
Percent bias (%)		−15	−12	3

selection of the five most skillful members from a set of WRF multiphysics configurations was based on the highest rank from a CSS and from two categorical scores as well as two continuous scores. The two categorical scores, i.e., PSS and EES, evaluated extreme daily precipitation and the two continuous scores, i.e., total Bias and MAE evaluated total precipitation amounts in the simulation period and mean daily amounts, respectively.

A three-nested domain configuration with a 1488 km × 1248 km outer domain and 12-, 4-, and 1-km grids was found to outperform a similarly sized two-nested domain configuration with 6- and 1-km grids, and an 826 km × 768 km domain with 6- and 1-km grids, for the downscaling of 31-km ERA5 reanalysis data. The average 30-day bias of the three-nested domain was 15 mm lower than the bias of 6–1a and 19 mm lower than the bias of 6–1b. The three nested domains also achieved lower MAE and both higher EES and PSS than the two nested domains. The results showed that two two-nested domain configurations cannot downscale the high-resolution ERA5 reanalysis as accurately as the three-nested domain configuration, at least for the month of January 2012, when a series of multiple large-scale precipitation events was observed. A better understanding of the effect of the size and the number of nested domains on simulated precipitation, with additional sensitivity tests in different study periods, could reduce the uncertainty of precipitation simulations, resulting from the domain configuration in future studies.

The 5-day initialization frequency, compared to 30-day continuous simulations without reinitializations, did not yield significant improvements in the precipitation simulations in the small-sized domain of this study, as seen from the total Bias, EES and PSS, despite the on average lower MAE by 1.5 mm day^{-1} of the 5-day initialization. These results are different from the findings of continental-scale studies, which found improved precipitation simulations with 1-day and 7-day initializations relative to 30 days. Thus, the small impact of more frequent initialization, as seen in this study, suggests that the control of the LBCs on the simulated atmospheric features is stronger for small domains than for larger domains. In this respect, future studies on similar small inner domains (less than 500 km × 500 km) could investigate the impact of LBCs with different model initializations and outer domain sizes on other simulated atmospheric features that are relevant for precipitation.

The evaluation of the WRF Model based on multiple physics parameterizations showed that the CSS of the different members was mostly related to the microphysics schemes, underscoring the important impact of this type of parameterization on high-resolution simulated precipitation. The Ferrier and WDM6 microphysics schemes exhibited similar performance and outperformed WSM6. The cumulus parameterization scheme was also found to have an effect on the simulated precipitation on the 1-km grid, even if it was turned off on the 1-km domain and turned on only on the coarser-resolution domains, i.e., 4 and 12 km. The GF cumulus scheme outperformed KF and BMJ schemes.

The use of four evaluation measures and the CSS, which integrates the four evaluation measures, facilitated the objective selection of a skillful 5-member set of WRF multiphysics configurations. The selected members preserved their performance in a 5-month validation period, compared to a 3-month calibration period. The evaluation of precipitation simulations for different model horizontal resolutions showed that the average of the five members had the least Bias in the 1 km, relative to 4 and 12 km. This reflects the ability of high-resolution precipitation simulations to yield the least water volume errors, which is especially important for hydrologic modeling studies and water balance investigations. Overall, the use of multiple and comprehensive evaluation measures for the assessment of WRF performance allowed a more complete evaluation of the different properties of simulated precipitation, such as daily and monthly volumes and daily extremes, for different dynamical downscaling options and model configurations. The stepwise approach proposed in the paper can be applied to select an efficient set of WRF multiphysics configurations that accounts for these properties and that can be used as input for hydrologic applications.

Acknowledgments. The authors thank Dr. Filippos Tymvios and his colleagues from the Department of Meteorology of Cyprus for sharing the precipitation data. Ioannis Sofokleous received computational resources by the Cy-Tera Project, which is co-funded by the European Regional Development Fund

and the Republic of Cyprus through the Research and Innovation Foundation (*Project Cy-Tera* NEA ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31).

Data availability statement. Data that were dynamically downscaled in this study were ERA5 reanalysis data, which are openly available at Climate Data Store at https://cds.climate.copernicus.eu. Observed datasets developed in this study are subject to restrictions due to confidentiality agreements with the Cyprus Department of Meteorology. Request access is available at the Department of Meteorology website at www.moa.gov.cy/moa/ms/ms.nsf.

REFERENCES

Avolio, E., and S. Federico, 2018: WRF simulations for a heavy rainfall event in southern Italy: Verification and sensitivity tests. *Atmos. Res.*, **209**, 14–35, https://doi.org/10.1016/j.atmosres.2018.03.009.

Berckmans, J., O. Giot, R. De Troch, R. Hamdi, R. Ceulemans, and P. Termonia, 2017: Reinitialized versus continuous regional climate simulations using ALARO-0 coupled to the land surface model SURFEXv5. *Geosci. Model Dev.*, **10**, 223–238, https://doi.org/10.5194/gmd-10-223-2017.

Berthou, S., E. J. Kendon, S. C. Chan, N. Ban, D. Leutwyler, C. Schär, and G. Fosser, 2018: Pan-European climate at convection-permitting scale: A model intercomparison study. *Climate Dyn.*, **55**, 35–59, https://doi.org/10.1007/S00382-018-4114-6.

Brisson, E., M. Demuzere, and N. Van Lipzig, 2016: Modelling strategies for performing convection-permitting climate simulations. *Meteor. Z.*, **25**, 149–163, https://doi.org/10.1127/metz/2015/0598.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Camera, C., A. Bruggeman, P. Hadjinicolaou, S. Pashiardis, and M. A. Lange, 2014: Evaluation of interpolation techniques for the creation of gridded daily precipitation ($1 \times 1\,km^2$); Cyprus, 1980–2010. *J. Geophys. Res. Atmos.*, **119**, 693–712, https://doi.org/10.1002/2013JD020611.

——, ——, G. Zittis, I. Sofokleous, and J. Arnault, 2020: Simulation of extreme rainfall and streamflow events in small Mediterranean watersheds with a one-way coupled atmospheric-hydrologic modelling system. *Nat. Hazards Earth Syst. Sci.*, **20**, 2791–2810, https://doi.org/10.5194/nhess-2020-43.

Cassola, F., F. Ferrari, and A. Mazzino, 2015: Numerical simulations of Mediterranean heavy precipitation events with the WRF model: A verification exercise using different approaches. *Atmos. Res.*, **164–165**, 210–225, https://doi.org/10.1016/j.atmosres.2015.05.010.

Centella-Artola, A., M. A. Taylor, A. Bezanilla-Morlot, D. Martinez-Castro, J. D. Campbell, T. S. Stephenson, and A. Vichot, 2015: Assessing the effect of domain size over the Caribbean region using the PRECIS regional climate model. *Climate Dyn.*, **44**, 1901–1918, https://doi.org/10.1007/s00382-014-2272-8.

Chan, S. C., E. J. Kendon, H. J. Fowler, S. Blenkinsop, N. M. Roberts, and C. A. Ferro, 2014: The value of high-resolution Met Office regional climate models in the simulation of multihourly precipitation extremes. *J. Climate*, **27**, 6155–6174, https://doi.org/10.1175/JCLI-D-13-00723.1.

Colin, J., M. Déqué, R. Radu, and S. Somot, 2010: Sensitivity study of heavy precipitation in limited area model climate simulations: Influence of the size of the domain and the use of the spectral nudging technique. *Tellus*, **62A**, 591–604, https://doi.org/10.1111/j.1600-0870.2010.00467.x.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman & Hall, 456 pp.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Giorgi, F., and W. J. Gutowski Jr., 2015: Regional dynamical downscaling and the CORDEX initiative. *Annu. Rev. Environ. Resour.*, **40**, 467–490, https://doi.org/10.1146/annurev-environ-102014-021217.

Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, https://doi.org/10.5194/acp-14-5233-2014.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, 1998: Towards improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.*, **34**, 751–763, https://doi.org/10.1029/97WR03495.

Hoerling, M., J. Eischeid, J. Perlwitz, X. Quan, T. Zhang, and P. Pegion, 2012: On the increased frequency of Mediterranean drought. *J. Climate*, **25**, 2146–2161, https://doi.org/10.1175/JCLI-D-11-00296.1.

Hogan, R. J., and I. B. Mason, 2012: Deterministic forecasts of binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jollife and D. B. Stephenson, Eds., John Wiley & Sons, 31–59.

——, C. Ferro, I. Jolliffe, and D. Stephenson, 2010: Equitability revisited: Why the ''equitable threat score'' is not equitable. *Wea. Forecasting*, **25**, 710–726, https://doi.org/10.1175/2009WAF2222350.1.

Hong, S. Y., and J. O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *Asia-Pac. J. Atmos. Sci.*, **42**, 129–151.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **34**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

——, K.-S. S. Lim, Y.-H. Lee, J.-C. Ha, H.-W. Kim, S.-J. Ham, and J. Dudhia, 2010: Evaluation of the WRF double-moment 6-class microphysics scheme for precipitating convection. *Adv. Meteor.*, **2010**, 707253, https://doi.org/10.1155/2010/707253.

Hu, X. M., M. Xue, R. A. McPherson, E. Martin, D. H. Rosendahl, and L. Qiao, 2018: Precipitation dynamical downscaling over the Great Plains. *J. Adv. Model. Earth Syst.*, **10**, 421–447, https://doi.org/10.1002/2017MS001154.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Jeworrek, J., G. West, and R. Stull, 2019: Evaluation of cumulus and microphysics parameterizations in WRF across the convective gray zone. *Wea. Forecasting*, **34**, 1097–1115, https://doi.org/10.1175/WAF-D-18-0178.1.

Ji, F., M. Ekström, J. P. Evans, and J. Teng, 2014: Evaluating rainfall patterns using physics scheme ensembles from a

regional atmospheric model. *Theor. Appl. Climatol.*, **115**, 297–304, https://doi.org/10.1007/s00704-013-0904-2.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

Katragkou, E., and Coauthors, 2015: Regional climate hindcast simulations within EURO-CORDEX: Evaluation of a WRF multi-physics ensemble. *Geosci. Model Dev.*, **8**, 603–618, https://doi.org/10.5194/gmd-8-603-2015.

Kioutsioukis, I., A. de Meij, H. Jakobs, E. Katragkou, J. F. Vinuesa, and A. Kazantzidis, 2016: High resolution WRF ensemble forecasting for irrigation: Multi-variable evaluation. *Atmos. Res.*, **167**, 156–174, https://doi.org/10.1016/j.atmosres.2015.07.015.

Leduc, M., and R. Laprise, 2009: Regional climate model sensitivity to domain size. *Climate Dyn.*, **32**, 833–854, https://doi.org/10.1007/s00382-008-0400-z.

Lim, K. S., and S. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models. *Mon. Wea. Rev.*, **138**, 1587–1612, https://doi.org/10.1175/2009MWR2968.1.

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

Lucas-Picher, P., F. Boberg, J. H. Christensen, and P. Berg, 2013: Dynamical downscaling with reinitializations: A method to generate finescale climate datasets suitable for impact studies. *J. Hydrometeor.*, **14**, 1159–1174, https://doi.org/10.1175/JHM-D-12-063.1.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Cloug, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Mohan, P. R., C. V. Srinivas, V. Yesubabu, R. Baskaran, and B. Venkatraman, 2018: Simulation of a heavy rainfall event over Chennai in Southeast India using WRF: Sensitivity to microphysics parameterization. *Atmos. Res.*, **210**, 83–99, https://doi.org/10.1016/j.atmosres.2018.04.005.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, https://doi.org/10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454, https://doi.org/10.1126/science.ns-4.93.453-a.

Piazza, M., A. E. Prein, H. Truhetz, and A. Csaki, 2019: On the sensitivity of precipitation in convection-permitting climate simulations in the Eastern Alpine region. *Meteor. Z.*, **28**, 323–346, https://doi.org/10.1127/metz/2019/0941.

Politi, N., P. T. Nastos, A. Sfetsos, D. Vlachogiannis, and N. R. Dalezios, 2018: Evaluation of the AWR-WRF model configuration at high resolution over the domain of Greece. *Atmos. Res.*, **208**, 229–245, https://doi.org/10.1016/j.atmosres.2017.10.019.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

——, and Coauthors, 2016: Precipitation in the EURO-CORDEX 0.11° and 0.44° simulations: High resolution, high benefits? *Climate Dyn.*, **46**, 383–412, https://doi.org/10.1007/s00382-015-2589-y.

Qian, J., A. Seth, and S. Zebiak, 2003: Reinitialized versus continuous simulations for regional climate downscaling. *Mon. Wea. Rev.*, **131**, 2857–2874, https://doi.org/10.1175/1520-0493(2003)131<2857:RVCSFR>2.0.CO;2.

Rogers, E., T. Black, B. Ferrier, Y. Lin, D. Parrish, and G. DiMego, 2001: Changes to the NCEP Meso Eta Analysis and Forecast System: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3DVAR analysis. Accessed 09 March 2020, https://www.emc.ncep.noaa.gov/mmb/mmbpll/mesoimpl/eta12tpb/.

Rojas, M., and A. Seth, 2003: Simulation and sensitivity in a nested modeling system for South America. Part II: GCM boundary forcing. *J. Climate*, **16**, 2454–2471, https://doi.org/10.1175/1520-0442(2003)016<2454:SASIAN>2.0.CO;2.

Seth, A., and F. Giorgi, 1998: The effects of domain choice on summer precipitation simulation and sensitivity in a regional climate model. *J. Climate*, **11**, 2698–2712, https://doi.org/10.1175/1520-0442(1998)011<2698:TEODCO>2.0.CO;2.

——, M. Rojas, B. Liebmann, and J. H. Qian, 2004: Daily rainfall analysis for South America from a regional climate model and station observations. *Geophys. Res. Lett.*, **31**, L07213, https://doi.org/10.1029/2003GL019220.

Skamarock, W., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Song, I. S., U. Y. Byun, J. Hong, and S. H. Park, 2018: Domain-size and top-height dependence in regional predictions for the Northeast Asia in spring. *Atmos. Sci. Lett.*, **19**, e799, https://doi.org/10.1002/asl.799.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221–232, https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2.

——, B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**, 41–50, https://doi.org/10.1002/met.53.

Tewari, M., and Coauthors, 2004: Implementation and verification of the unified NOAH land surface model in the WRF model. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 14A.2, https://ams.confex.com/ams/pdfpapers/69061.pdf.

Tymvios, F., D. Charalambous, S. Michaelides, and J. Lelieveld, 2018: Intercomparison of boundary layer parameterizations for summer conditions in the eastern Mediterranean island of Cyprus using the WRF-ARW model. *Atmos. Res.*, **208**, 45–59, https://doi.org/10.1016/j.atmosres.2017.09.011.

Vannitsem, S., and F. Chomé, 2005: One-way nested regional climate simulations and domain size. *J. Climate*, **18**, 229–233, https://doi.org/10.1175/JCLI3252.1.

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779, https://doi.org/10.1175/MWR2826.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Young, A. R., 2006: Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model. *J. Hydrol.*, **320**, 155–172, https://doi.org/10.1016/j.jhydrol.2005.07.017.

Zhang, D., and R. A. Anthes, 1982: A high-resolution model of the planetary boundary layer—sensitivity tests and comparisons with SESAME-79 data. *J. Appl. Meteor.*, **21**, 1594–1609, https://doi.org/10.1175/1520-0450(1982)021<1594:AHRMOT>2.0.CO;2.

Zittis, G., P. Hadjinicolaou, and J. Lelieveld, 2014: Comparison of WRF model physics parameterizations over the MENA-CORDEX domain. *Amer. J. Climate Change*, **3**, 490–511, https://doi.org/10.4236/ajcc.2014.35042.

——, A. Bruggeman, C. Camera, P. Hadjinicolaou, and J. Lelieveld, 2017: The added value of convection permitting simulations of extreme precipitation events over the eastern Mediterranean. *Atmos. Res.*, **191**, 20–33, https://doi.org/10.1016/j.atmosres.2017.03.002.