Supplemental Material

Journal of Hydrometeorology

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

https://doi.org/10.1175/JHM-D-20-0110.1

Supplemental material

1. Contingency table based verification measures/skill scores

The equations of measures and skill scores based on the 2×2 contingency table (Fig. S1) are listed below (Hogan and Mason 2011):

Hit Rate (H): $$H = \frac{a}{a+c} \quad (1)$$

False Alarm Rate (F): $$F = \frac{b}{b+d} \quad (2)$$

Frequency Bias (BIAS): $$BIAS = \frac{a+b}{a+c} \quad (3)$$

Peirce Skill Score (PSS): $$PSS = \frac{a}{a+c} - \frac{b}{b+d} \quad (4)$$

Equitable Threat Score (ETS): $$ETS = \frac{a-a_{ref}}{a+b+c-a_{ref}} \quad (5)$$

where $a_{ref} = (a+b)(a+c)/n$

Extreme Dependency Score (EDS): $$EDS = \frac{2\ln[(a+c)/n]}{\ln(a/n)} - 1 \quad (6)$$

Extreme Event Score (EES):
$$EES = \frac{a}{a+c} \cdot \frac{a+b}{a+c} \qquad for\ b \le c$$
$$EES = \frac{a}{a+c} \cdot \frac{a+c}{a+b} \qquad for\ b > c \quad (7)$$

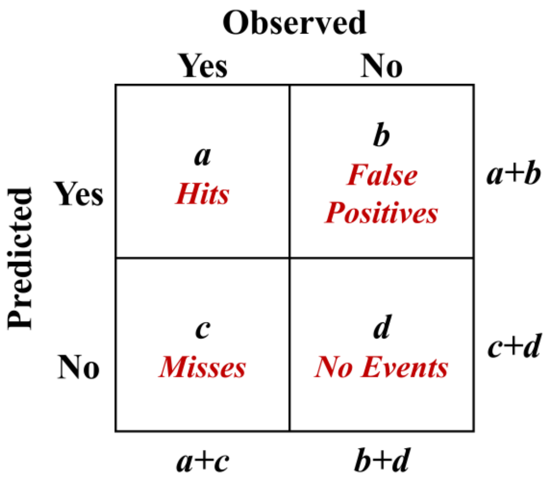

FIG. S1. The 2×2 contingency table with the frequencies of possible outcomes, *a*, *b*, *c* and *d* of a predicted (simulated) variable, in relation to the observations.

2. Properties of the Extreme Event Score (EES)

a. Equitability of Extreme Event Score

A verification measure *S*, defined in terms of the four elements *a*, *b*, *c* and *d* of the contingency table, has an expected value *E(S|p)*, given a particular base rate *p*. The base rate *p* is equal to *(a+c)/n*, where *n* is the sum of the four elements or the sample size. The expected value, generally referred to as expected score, is computed from the summation of all possible contingency tables *m*, each weighted by their probability of occurrence *P(a,b,c,d|p)*. According to Hogan et al. (2010), a verification measure is equitable, when it awards all random forecasts the same expected value S_0:

$$E(S|p) = S_0 \quad \text{for all } p \text{ and } q_s \tag{8}$$

where q_s is the sample forecast rate, which is equal to *(a+b)/n*. These authors calculated the expected score using the equations below:

$$E(S|p) = \sum_{i=1}^{m} S(a_i, b_i, c_i, d_i) P(a_i, b_i, c_i, d_i|p) \tag{9}$$

Equivalently, the expected score can be expressed in terms of the probability of a particular sample forecast rate q_s.

$$E(S|p) = \sum_{q_s} E(S|p, q_s) P(q_s|p) \tag{10}$$

where

$$E(S|p, q_s) = \sum_{i=1}^{m} S(a_i, b_i, c_i, d_i) P(a_i, b_i, c_i, d_i|p, q_s) \tag{11}$$

The probabilities in equations 10 and 11 are calculated as follows.

$$P(q_s|p) = C(n, nq_s) {q_p}^{nq_s} (1 - q_p)^{n(1-q_s)} \tag{12}$$

and

$$P(a_i, b_i, c_i, d_i|p, q_s) = \frac{C(n, nq_s) C(n - np, nq_s - a)}{C(n, nq_s)} \tag{13}$$

where $C(x,y)=x!/[y!(x-y)!]$ and q_p is the population forecast rate of occurrence (different from the sample forecast rate of occurrence q_s), ranging anywhere between 0 and 1, inclusive.

Following the same principle of investigating the equitability of a verification measure, as in Hogan et al. (2010), the expected score of random forecasts of *EES*, as well as of *PSS*, *ETS* and *EDS*, is plotted in Fig. S2. The expected score is plotted against the sample size *n* and four different base rates *p* and population forecast rates of occurrence q_p. The selected base rates of 0.1, 0.05 and 0.02 correspond to the frequency of observed events $(a + c)/n$. This particular selection of base rates (less or equal to 0.1) was made in order to investigate the equitability of *EES* for less frequent and extreme events.

It can be seen in Figure 2 that *PSS* and *ETS* are equitable for all sample sizes, with an expected score equal to zero, except for sample sizes less than 30 for *p*=0.1 for *ETS*. The expected score of *EES* as well as of *EDS* are not constant for different base rates. The expected value of the *EES* is constant around 0.04 for *p*=0.05. For the case of rarer events, i.e., *p*=0.02, the expected score of *EES* is constant and equal to 0.02 for all *n*. This contrasts with the *EDS*, which does not stabilize to a constant value, neither for *p*=0.05, nor for *p*=0.02, for the shown range of sample sizes. From the behaviour of the expected value in Fig. S2, the *EES* cannot be seen as truly equitable. Yet for the case of small base rates, corresponding to rare events, which the proposed score is intended to evaluate, the *EES* is equitable.

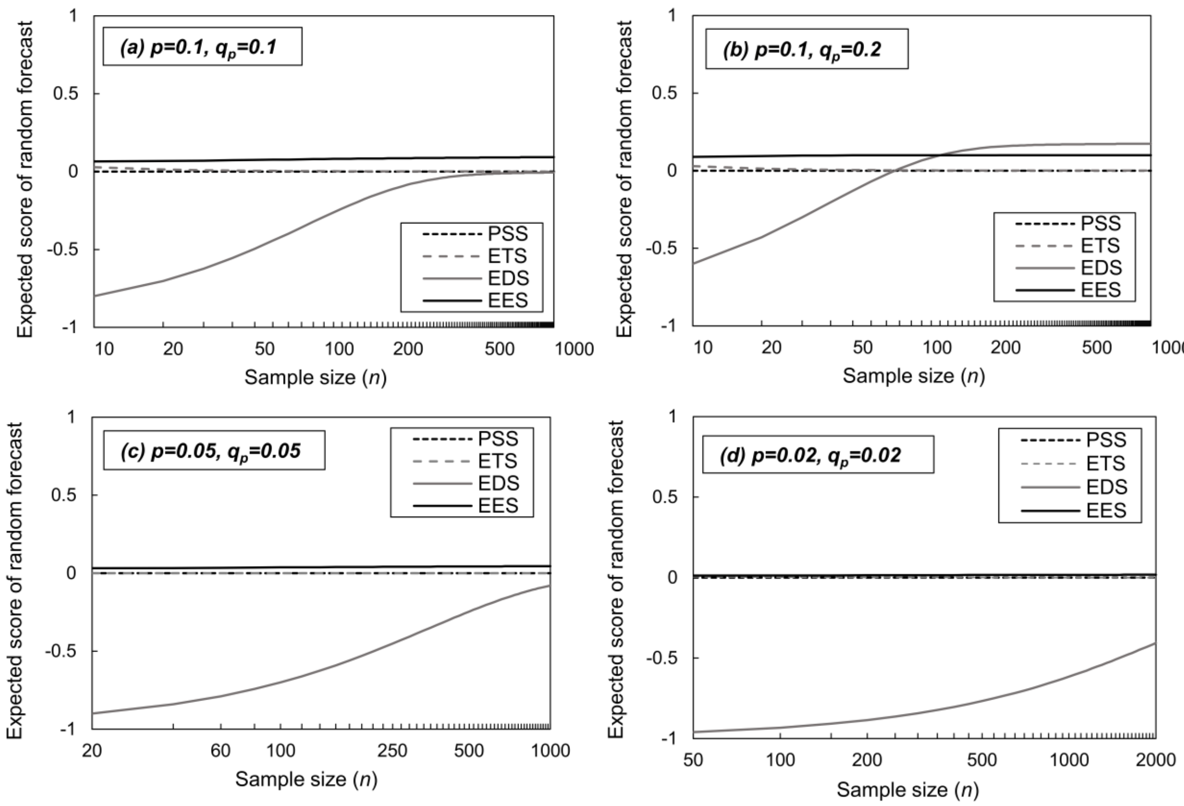

FIG. S2. Expected score of random forecasts for the Pierce Skill Score (*PSS*), the Equitable Threat Score (*ETS*), the Extreme Dependency Score (*EDS*) and the Extreme Event Score (*EES*) for different sample sizes and for four different pairs of base rate p and population forecast rate q_p.

b. Evaluation of Extreme Event Score against other verification measures

The proposed *EES* and the three other skill scores of section 2a are plotted against the three scalar attributes of the contingency table, i.e., Hit Rate *H*, *BIAS* and False Alarm Rate *F* in Fig. S3, for all 91 000 possible contingency tables of size 1000 (n) and base rate p=0.1. The *EES* is equal to *H* (1:1 line) when the random forecasts are unbiased. For all other cases of *BIAS* ≠ 1, the *EES* is less than *H* for *BIAS*>1 and higher than *H* for *BIAS*<1. The plot of *EES* versus *BIAS* has a maximum (*EES*=1) at *BIAS*=1 (unbiased forecast), while *EES* decreases with increasing underestimation or overestimation of extreme events by the model. In relation to *F*, the *EES* maximizes at *F* equal to 0 and minimizes at *F* equals 1. The changes in the values of *ETS* for different *H*, *BIAS* and *F* follow similar patterns as the changes in the values of *EES*. The direction of change of the values of *PSS* is the same as for *EES*. The *PSS* exhibits, however, a linear relation with *H*, *BIAS* and *F*, which is not seen with the *EES* or *ETS* and this dependence is explained by the formulation of *PSS* (equation 4). The *EDS* has a clear pattern for its change with *H*, with a unique value of *EDS* corresponding to a unique value of *H* and this is neither seen with *BIAS* nor with *F*.

The relation of *EES* against the other three skill scores *PSS*, *ETS* and *EDS* is linear or nearly linear in certain cases as seen in Fig. S4. There is however a large variation in the relative change of the *EES* with these three scores, which is explained by the formulations of the four scores (equations 4-7).

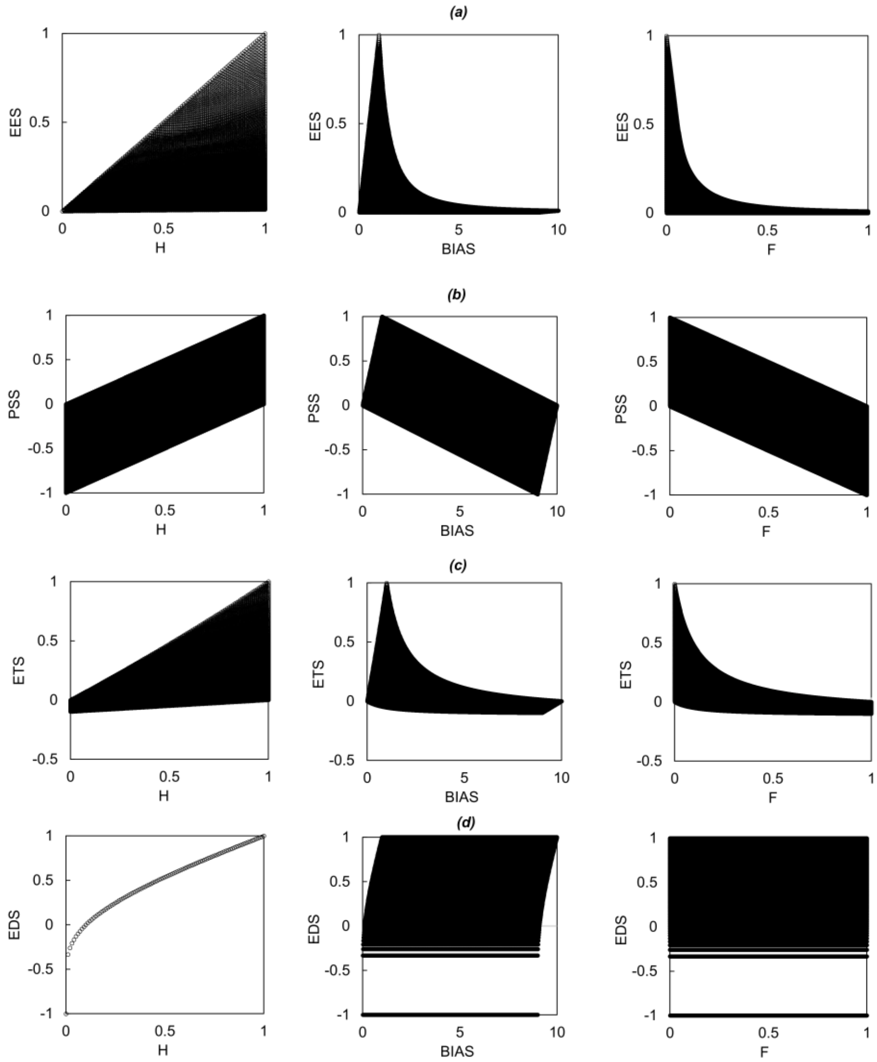

FIG. S3. Values of the *EES*, *PSS*, *ETS* and *EDS* against Hit Rate (*H*), *BIAS*, and False Alarm rate *F* for all 91 000 possible contingency tables of size 1000 and base rate p=0.1.

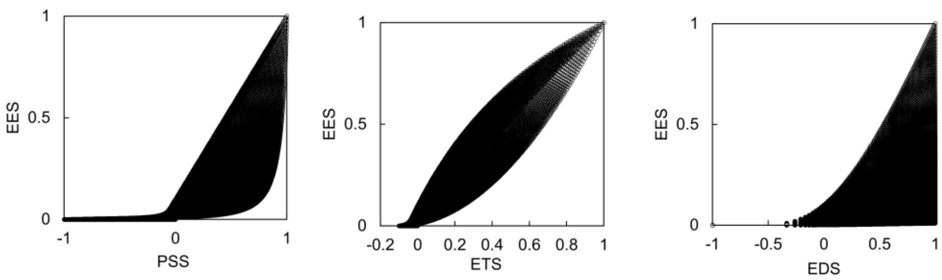

FIG. S4. Values of the *EES* against *PSS, ETS* and *EDS* for all 91 000 possible contingency tables of size 1000 and base rate p=0.1.

3. Spinup time in WRF simulations

Some studies recommend a spinup time of 12 hours in order to prevent instabilities in WRF (Kleszek et al. 2014; Bonekamp et al. 2018). The length of spinup is not however a fixed value because it is associated with the quality of the model input data (Warner and Peterson 1997).

To quantify any instabilities occurring in the simulated precipitation in the initial time steps due to the choice of spinup time, the hourly domain-average accumulated precipitation is computed for various lengths of spinup in the first 24 hours for two cases. First, the simulated precipitation with a spinup time of 6 and 12 hours (6h, 12h) is evaluated against observations for the 1st of January 2012 (Fig S5a). Secondly, simulated precipitation obtained with spinup times of 6 and 12 hours as well as 14 days (14d) is evaluated for the 15th of January 2012 (Fig. S5b). The simulation results during the spinup time are omitted from the accumulated precipitation shown in Fig. S5.

According to Fig. S5, the 6h spinup does not result in any instabilities in the modeled precipitation in the first simulation hours after this initial time. The accumulated precipitation is rather similar with that from the longer spinup times in two different dates. Based on these results the 6h spinup is found adequate. The use of the high-resolution (31 km) reanalysis dataset as initial and boundary conditions for the WRF simulations may be the reason why a 6h spinup produces similar results to the suggested 12h spinup (Kleszek et al. 2014; Bonekamp et al. 2018). Givati et al. (2012), Kotlarski et al. (2012) and Tymvios et al. (2018) used also a 6h spinup for their WRF simulations with similar or larger domain sizes than the domain size of this study.

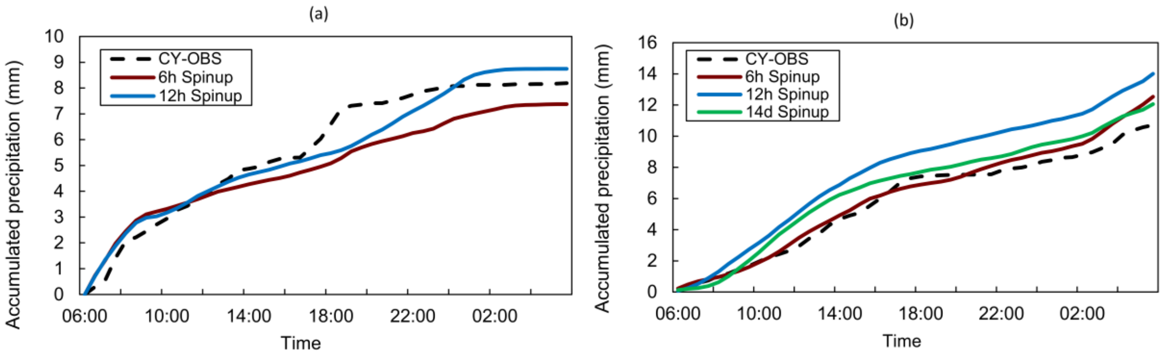

FIG. S5. Hourly observations from the CY-OBS dataset and domain-average accumulated precipitation from the WRF simulations (T11 parameterization) (a) with 6-hour (6h) and 12-hour (12h) spinup times on the 1st of January 2012 and (b) with 6-hour, 12-hour and 14-days (14d) spinup times on the 15th of January.

References

Bonekamp, P. N. J., E. Collier, and W. W. Immerzeel, 2018: The Impact of Spatial Resolution, Land Use, and Spinup Time on Resolving Spatial Precipitation Patterns in the Himalayas. *J. Hydrometeor.*, **19**, 1565–1581, https://doi.org/10.1175/JHM-D-17-0212.1.

Givati, A., B. Lynn, Y. Liu, and A. Rimmer, 2011: Using the WRF model in an operational streamflow forecast system for the Jordan River. *J. Appl. Meteor. Climatol*, **51**, 285, http://dx.doi.org/10.1175/JAMC-D-11-082.1

Hogan, R.J. and I.B. Mason, 2012: Deterministic forecasts of binary events. *Forecast verification: a practitioner's guide in atmospheric science*, Jollife, I.T., and D.B. Stephenson, Ed., John Wiley & Sons, 31–59.

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

Kleczek, M. A., G.-J. Steeneveld, and A. A. Holtslag, 2014: Evaluation of the weather research and forecasting mesoscale model for GABLS3: Impact of boundary-layer schemes, boundary conditions and spin-up. *Bound.-Layer Meteor.*, **152**, 213– 243, https://doi.org/10.1007/s10546-014-9925-3.

Kotlarski, S., S. Hagemann, P. Krahe, R. Podzun, and D. Jacob, 2012: The Elbe river flooding 2002 as seen by an extended regional climate model. *J. Hydrol.*, **472-473**, 169-183, http://dx.doi.org/10.1016/j.jhydrol.2012.09.020

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, http://dx.doi.org/10.1016/j.atmosres.2017.09.011.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A Tutorial on Lateral Boundary Conditions as a Basic and Potentially Serious Limitation to Regional Numerical Weather Prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2618, https://doi.org/10.1175/1520-0477(1997)078<2599:ATOLBC>2.0.CO;2.