

The manuscript presents a comprehensive 30-year climatological evaluation (1991–2020) of precipitation over the Sierra de Guadarrama (Spain) using the WRF model at resolutions of 9, 3, and 1 km. The authors employ a high-density observational network, including the GuMNet high-altitude stations, to validate the model's performance in a region of complex orography.

The study provides valuable insights into the *added value* of kilometer-scale modeling. Key findings include the model's ability to better represent altitudinal gradients at 1 km, despite a systematic overestimation of wet-day frequency and accumulations. Notably, the authors observe that the 3 km resolution (WRF2) often yields better statistical agreement than the 1 km domain (WRF3).

This work represents an important computational effort and provides a valuable dataset for mountain hydrology and regional climate studies in the Iberian Peninsula. However, from the perspective of model development, the physical configuration chosen for the primary 30-year climatology appears sub-optimal based on the authors' own sensitivity tests. The manuscript requires a major revision to address physical inconsistencies in the *convective gray zone* and to strengthen the benchmarking of its added value.

## Major Comments

**I. Physics-Dynamics Coupling and Cumulus Parameterization.** The authors use the New Tiedtke cumulus scheme across all domains, including the 1 km resolution (WRF3). As established in recent literature (e.g., Tomassini et al., 2023), the range 1-5 km is the *convective gray zone* where convection is partially resolved.

Unless I am misinterpreting the results presented in Figure 11, Simulation 3 (New Tiedtke-CPS, where parameterization is off at 1 km) appears to reduce the precipitation overestimation reported in the Reference run. In Panel (b), the WRF3 results (blue dots) for high-accumulation sites seem better aligned with the 1:1 observational diagonal than in Panel (a).

The authors admit that "other parameterizations or CPS approaches may be more adequate" at 1 km. Could the authors clarify if I have overlooked a specific reason why the Reference configuration was preferred for the 30-year study despite their own sensitivity tests showing that scale-aware schemes or a CPS-off approach would have yielded superior results?

**II. Benchmarking against High-Resolution Reanalysis (CERRA).** The study evaluates *Added Value* solely against ERA5 (~31 km).

Given that ERA5 fundamentally misses the sharp orographic peaks of the Sierra de Guadarrama, proving that WRF is "better" than a 31 km global model is a "low bar" that provides limited insight into the actual merit of the 1 km simulation. I would suggest expanding the benchmarking beyond ERA5. It would be very helpful for the reader to see how both the 3 km (WRF2) and 1 km (WRF3) climatologies compare to an existing regional reanalysis such as CERRA (5.5 km). Since the results show that WRF2 often matches or exceeds WRF3 in statistical skill, it would be valuable to demonstrate if these kilometer-scale simulations provide any physical value that a regional reanalysis does not already offer.

**III. Microphysics Sensitivity.** The manuscript identifies a persistent overestimation of wet-day frequency, reaching a factor of 2 in summer. While the study tests various cumulus schemes, the Thompson microphysics remains constant throughout all experiments.

The overestimation of low-intensity events is often a symptom of grid-scale condensation and autoconversion processes, which are inherently microphysical. Considering that this overestimation is the primary driver for the overstated precipitation totals, would the authors expect substantial differences in these results if a different microphysics scheme were used? It would be scientifically valuable to discuss whether the current results—specifically the peaks of 1800 mm/year at the summits—are physically robust or if they might be sensitive to the choice of microphysics (e.g., using a different two-moment scheme or an alternative single-moment scheme like WSM6).

### **Minor Issues**

In Figures 5 and 11, the scatter plots have inconsistent ranges (e.g., in Fig. 5 (Top), Y-axis goes to 2000 mm while X-axis ends at 1450 mm) and non-square aspect ratios. For accurate visual assessment of the 1:1 line, the authors should use identical ranges for both axes and a perfectly square frame. This ensures that the reader can correctly perceive the magnitude of model biases.

### **References**

Tomassini, L., *et al.* (2023). *Confronting the Convective Gray Zone in the Global Configuration of the Met Office Unified Model*. *Journal of Advances in Modeling Earth Systems*, **15**, e2022MS003418.