General Comments

Tang et al described a new developed geospatial probabilistic estimation package (GPEP) to generate ensemble meteorological datasets based on station observations. They demonstrated good performance of GPEP with some examples. Compared to Gridded Meteorological Ensemble Tool (GMET), GPEP has several improvements, such as multiple selection of spatial interpolation methods, additional meteorological variables, user defined inputs, and etc. However, the computational performance of GPEP can be worse than GMET since GPEP is developed in Python. The computational performance is improved by using multiprocessing package, and they show speed of the simulation is significantly increased as more cores are used. GPEP represents a useful tool for the Earth science community. Specifically, the gridded meteorological variables that generated by GPEP can be used as forcing data for Earth System Models, hydrological models, especially for ungagged regions. The ensemble estimations can be further used for assessing the uncertainty caused by meteorological forcings in the simulation. Although I think this study could be a significant contribution to ensemble geophysical datasets estimation, it can be further improved before publication in Geoscientific Model Development. Please find my major concerns and specific comments in the following.

Major Comments

1. GMEP considers the spatial correlation and correlation between temperature and precipitation. As more variables can be processed and generated in GPEP, are the intercorrelation among variables considered? For example, in the application of Earth System Model, precipitation, humidity, radiation, and temperature are needed. Those variables are correlated to each other in time and space. Is it possible for GPEP to generate those variables together? In addition, how other variables are generated in GPEP is not described. Eq (1) – Eq (4) describes how temperature and precipitation are generated in GMEP. Can the same equations be used for any other variables?

2. What is the roadblock for running GMEP for the High-resolution meteorological forcing ensemble generation? It will be interesting to see if GPEP has better or similar performance than GMEP in such application.

3. There are other 1km reanalysis meteorological datasets that can be used for benchmarking GPEP's simulation for the high-resolution meteorological forcing ensemble generation example (i.e., Sec 4.2). For example, Daily Surface Weather Data on a 1-km Grid for North America (Daymet). Is GPEP better than existing high-resolution reanalysis dataset for reproducing the meteorological variables at the station locations? How is the spatial pattern compared to such high-resolution reanalysis dataset. I think by adding such benchmark and evaluation can give the readers/users more confidence on the application of GPEP.

4. The authors argued GPEP features multiple selection of spatial interpolation method. But still the locally weighted linear/logistic regression were used in all the demonstrations. I think it is necessary for the author to show the application of other spatial interpolation method. Specifically, will using supervised learning method for the spatial interpolation can

improve the performance in Figure 4 (the high-resolution meteorological forcing ensemble generation)?

5. Does GPEP only accept gauge data as input? Can we use gridded reanalysis dataset as input and output at higher spatial resolution? In addition, can GPEP generate the meteorological datasets at sub-daily scale? This can be useful for using GPEP to generate forcings for models, as some models requires sub-daily meteorological forcings.

Specific Comments

Line 32: In my experience, reanalysis dataset commonly has better temporal coverage than the station observation. Please cite relevant reference to support this statement.

Line 46: Full name for HadCRUT4?

Line 88 and Line 89: What are continental EMDNA and global EM-Earth?

Line 98: Is temperature range = maximum temperature – minimum temperature in the day?

Line 235: I wonder the computational performance for large scale application, for example, continental or global simulation.

Section 4.1: I suppose same algorithm was adopted in GPEP. So, is the difference caused by random seed used in GMET? Can the authors further explain the attribution for the differences?

Figure 2: Does subplot (a) represent the mean precipitation of the simulation period, or all the daily precipitations within the simulation period?

Figure 3: What is the source and spatial resolution of the elevation data? Does GPEP estimate the south-north and west -east slopes internally, or the user need to preprocess it? It will be useful for the authors to describe how the slopes were calculated.

Line 317: Could the higher performance in the flat eastern areas due to that there are more stations in this region? Then this raise the question what is the minimum number of stations that needed by GPEP for a good performance? What is the performance of GPEP in data sparse region? I am not asking the authors to run additional simulations for this question, some discussions and perspective from the authors will be very helpful for the readers.

Line 322 – line 325: Can the authors quantify the ensemble spread? For example, the spatial correlation between any two ensemble members.

Line 326: Is there a station fall inside the area selected by Figure 6? If so, how the simulation ensemble compared to observation?