In the abstract, the use of the word “forecast” is troublesome. Climate models do not produce forecasts, at least not in any statistical sense. Climate models produce projections based on assumptions about the greenhouse gas emissions associated with a given scenario, along with other forcings. These scenarios may or may not be connected to any reality. As such it is crucial to be careful with language, especially when trying to lay a foundation for a probabilistic interpretation.

The data set does not appear to include multiple ensemble members for each model. How would this approach change with the inclusion of ensemble members for each model? Similarly, is it required that the models be bias corrected, which is a part of the downscaling procedure used in this data set? What is the impact of the biases in the target dataset used in the downscaling?

There is some inconsistency in the choice of model for Tx90p. In line 95, the authors note that Tx90p is constrained to be between [0,100]. (It is not mentioned, but it is also a discrete variable with a finite set of possible values as it is defined on line 55.) The authors then choose to use a Gaussian distribution, rather than the truncated distribution as suggested in the manuscript near line 100 or some other distribution more in line with a discrete variable. The Gaussian might offer a reasonable approximation, but this should be explored further and justified accordingly.

Multi-model ensembles pose significant challenges, and the authors have chosen a fairly simplistic approach of a mixture model with equal model weights based on Weigel, et al., 2010. While the Weigel paper is an important contribution to the discussion on model weighting, much work has been done in the interim, including by a co-author on the Weigel paper. While a technical note is likely not the right place to delve into the many issues, it is remiss to downplay them particularly when the paper is trying to establish a
probabilistic framework. There are many approaches to model weighting, including Bayesian model averaging and the more recent climate model weighting by independence and performance (ClimWIP), among many others, that address skill and model dependence. However, the issue with model weighting is not only about adjusting weights to account for skill or dependence between models due, for example, to shared components, but to whether the multi-model ensemble actually represents the range of uncertainty needed to justify a probabilistic statement (e.g., Chandler, 2013). At the very least, the authors should acknowledge the complexity of the issue and more recent work in the area, as well as the challenges that go beyond equal vs unequal weights and impact the very premise of their work.

The OEP formula near line 145 requires the assumption of independent events. There are many physical processes that exist on time scales longer than a year, and the authors have worked to include them through modeling of the mean and filtering. Can you assume years are independent? (This is somewhat different than using the OEP from catastrophe modeling point of view where different disasters are more naturally modeled as independent of each other.)

Near line 165, the authors note that they remove inter-annual variability using a moving window filter. However, they already have the quadratic best-fit - is this not enough? Or is there other temporal variability they are trying to keep (e.g., decadal)? A bit more detail is necessary here, including the type of filter.

The authors use a “best-fit” (least-squares?) quadratic curve to remove the trend. For certain climate variables, the GEV is then fit via maximum likelihood to the residuals and then the location parameter is shifted by the value of the best-fit curve. The best-fit curve should represent the mean of the distribution. Given that the location parameter of the GEV is not the mean (the mean of the GEV is a function of all three parameters), this could have some unintended consequences, including distributions at points in time that are not valid due to the finite lower/upper bound on GEV depending on the value of the shape parameter. If it is appropriate to place a mean on the location parameter, this can easily be done directly through maximum likelihood (or in a Bayesian framework) and such an approach is easily accomplished through, for example, the extRemes package in R.