Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-506-RC2, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Comparison of statistical downscaling methods for climate change impact analysis on drought" *by* Hossein Tabari et al.

Anonymous Referee #2

Received and published: 11 December 2020

In general, I found this manuscript difficult to follow. If I weren't supposed to review it, I would have stopped reading it fairly quickly. This is not so much a question about the English but the way the material is structured and that it presumes that the reader already is familiar with the work. I also think I have some misgivings about the results and have a few issues with the nomenclature. There is no proper evaluation of the methods, and I think the authors could give more background and discuss their efforts in the context of other relevant papers.

L.40-48: It may be of interest to note that the occurrence of droughts also may be a consequence of a reduction in the global area with 24-hr rainfall (DOI: 10.1088/1748-9326/aab375). L.66: The statement 'As there is no single best downscaling method' may be correct generally for a range of situations, but for specific and limited problems,

there may be some method that is superior and provides the most reliable answer. L.73/79: ensemble sizes of 14 or 25 members are on the low end for giving robust information, due to pronounced and stochastic/chaotic decadal variations (Deser et al., 2012) and 'the law of small numbers'. If the ensemble is to be regarded as a statistical sample, then it typically needs more than 30 data points to get results that are not heavily affected by random sampling fluctuation, and preferably more than 100 members to get more robust statistics. L83: If 'precipitation time series produced by GCMs as sole predictor' also is used for model calibration, then there seems to be a problem because the GCMs are not synchronised with the observations in the real world. L92: Both GCMs and RCMs have a minimum skillful scale (DOI: 10.2151/jmsj.2015-042) which implies that single grid box values should not be used as a predictor. The difference between downscaling and bias adjustment is that the former utilises the scales that the models are able to skillfully reproduce and known dependencies between large and small scales to infer changes in the local (small-scale) climate whereas the latter is about modifying individual grid-box values (below the minimum skillful scales) to match the statistical properties of observations from similar locations. L100: I found the section difficult to follow and at times there seemed to be too little information. A rule of thumb is it is annoying to read a paper where you need to read another paper before you understand what it does, and the reader should be able to reproduce the results based on a detailed 'recipe' explaining the method. Reproducibility is an important issue. L101: State that the four methods are explained in more detail below. Also, I'm not sure why there are four methods - what is the merit of each? Sure, I get it that they may give different results, but so what? L107: Insufficient information about the BC methods? L115: A change factor of mean cannot really be classified as a downscaling method as it does not involve different spatial scales. L180: Perhaps an illustration will make it easier to follow. L199: Why not downscale the parameters of the pdf for the dry-spell duration? If there is a probability p that it rains on a specific day, then the pdf for the duration of dry spells ought to follow a binomial distribution with one parameter p = 1/mean(spell lengths). This approach has been tried for heatwaves in India (DOI:

10.5194/ascmo-4-37-2018). Because the statistics of spells (duration of events) tend to follow the binomial distribution, it is advisable to include the mean duration - not just its median (which is usually not a quantity that corresponds to the parameters of a pdf). L200-206: There is a subtle but important difference between a rain gauge measurement (a point measurement) and gridbox values from GCMs (area mean estimates), which also has implications for threshold values. This is also expected to be affected by different spatial resolutions (Table 1). For a coarser model, local 'convective' would be smudged out over a larger area (I guess it is parameterised) and expected to have a different statistical characteristic to a very tiny spatial sample (rain gauge measurements with diameters of the order of centimetres). This in itself is a justification for bias adjustment, but nevertheless makes it difficult to compare dry and wet days in models with different spatial resolution. L222: Pmax is the max monthly maximum precipitation or the mean monthly maximum? L235: Why is a two-tailed test used here? L237: It is important with a proper evaluation of the methods involving e.g. cross-validation and historical data, which seems to be missing. E.g., can the methods reproduce historical changes/variations in dry-spell lengths? L253: 'medium' should be 'median'? It's better to use the mean because the statistics of dry-spell length is expected to follow a distribution that is not too different to the binomial distribution for which the mean is connected to its only parameter. L258: Define 'DSL'. L269: It's annoying to stumble across acronyms like 'RI' which I then need to scroll up to remind me of what it stands for. The same goes with the other acronyms. I think it's a bad habit (and sloppy) to use many acronyms and abbreviations and if there is no need to do so, it's generally better to spell out the words to make it easier for the reader and ensure that (s)he focuses on the message rather than trying to decipher the text. The punishment is of course revisions... I think that the ensemble used here is too small to make any judgement about future climate due to pronounced stochastic local variability on decadal scales (Deser et al., 2012) and the law of small numbers. This stochastic nature should call for the use of probabilistic projections. Furthermore, I would not call the efforts discussed here 'downscaling' but they are more along the line of 'prediction' or 'bias correction' in

my opinion (there is no scaling dependencies being utilised between large and small). It can be shown that the large-scale circulation (SLP anomalies) tend to determine whether there is rain or no rain for a location, and downscaling would involve using the large-scale circulation as a means to infer the wet-day frequency and maybe also dry-spell lengths. L380: Need to define 'accurately' and provide evidence for the statement.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-506, 2020.