Point-by-Point response to Reviewer Comments
Biometrics MS #150926M
May 26, 2016

We wish to thank the associate editor and two reviewers for their thorough reviews and constructive suggestions. To address the concerns raised in this review we have carefully edited the manuscript to clarify issues regarding the skew-t parameterization and other notational issues. We have also added some new simulation results to explore more complicated max-stable processes, and a discussion of the properties of the spatiotemporal extension of our model. We hope you agree that these changes have improved the manuscript.

Response to the comments of the Associate Editor

Major comments

1. The model in (1) has a non-standard skew distribution notation. As far as I could check, the parameter σ^2 should have an inverse gamma with parameters $a = b$, and not arbitrary and different a and b. The equality is important in the derivations and it is one of the reasons we get $2a$ degrees of freedom. Indeed, some reserachers adopt $a/2$ and $a/2$ (rather than a and b) such that they end up with a degrees of freedom. Can you provide a web-supplement proof that your more general model with $a \neq b$ is a skew-t distribution with $2a$ degrees of freedom, or provide a reference for this result?

2. In the simulations, you adopted a prior distribution for the a and b parameters in $\sigma_t^2(\mathbf{s}) \sim IG(a, b)$ as $a \sim$Gamma$(0.1, 0.1)$ and b with a discrete distribution over a mesh from 0.1 to 10 with spacing 0.1. The parameter $\sigma_t^2(\mathbf{s})$ is clearly a crucial parameter as it controls the scale. Why this choice? Trial and error? What have you used in the real dataset? Is this supposed to be a default choice?

3. A very confusing issue for me was the adoption of the new parametrization $\lambda = \lambda_1\lambda_2$. Typically, users adopt a prior such as $\lambda \sim N(0, V)$ with a large and known value for V. One consequence is that the posterior distribution of λ has non negligible probability mass around 0, making the skew model irrelevant. When you adopted the parametrization $\lambda = \lambda_1\lambda_2$, you "solved" this issue by making zero an impossible value under the prior. The parameter λ_1 is equal to 1 or $+1$ and it basically selects the semi-axis where the slant parameter will live. The parameter $\lambda_2 \sim IG(a, b)$ and hence $\lambda_2 > 0$ and bounded away from zero, in practice. The value $\lambda = 0$ is not supported by the prior. Under your parametrization, you are artificially avoiding your model to become a simple gaussian process.

   What, exactly, is the convergence problem you had with the parameter before adopting this reparametrization? It affected only ?

   Incidentally, you used the same notation $IG(a, b)$ for two different purposes: for λ_2 in page 10, and for σ^2 in page 4. Please, keep them different as there are a lot of things going on in your model.

4. In page 6, section 3.1, you say that you "impute the censored values as a step in the MCMC algorithm". Why do you need these pseudo-values, specially considering that you do have the real below threshold observations? Is it because you need them follwoing to your tail model (the skew-t model)? But why do you need all the uncensored pseudo data?

Anyway, you should make clear that trading a focus only on the extreme tail data you loose efficiency by decreasing the number of (real) data available. Censoring on purpose seems a rather drastic technique.

5. The temporal dependence structure is induced by putting AR(1) models on some of the skew-t parameters. It is not clear what sort of temporal dependence that induces on the data.

Minor comments

1. Page 3, line 11: "... with computing on the order of Gaussian models "
2. Page 3, line 13: ... as well as TO make predictions ...
3. Page 3, line -2: ... compared TO Gaussian AND max-stable ...
4. Page 8, line -1: a partition in the $\mathbf{w}$ space induces a partition in the $\mathbf{w}$. The property (7) is used in the $\mathbf{w}$ space. Is this property still respected in the $\mathbf{w}$ space since the transformation is non-linear and acts independently in each coordinate?
5. Page 9, line -5: "the marginal distributions are Gaussian": which distributions, those of $Y(\mathbf{s})$?
6. The hierarchical model in page 10 has $\sigma_t(\mathbf{s})$ multiplying only $v_t(\mathbf{s})$ and leaving out the term $|z_t(\mathbf{s})|$. I think this is a typo.
7. In page 12, line -1: you mention the partition label switching problem appearing during the MCMC. This was not consequential for the inference? Can you expand on this?
8. There are several papers using skew distributions in the spatial context. You should refer to the most relevant ones in your related work section.

Response to the comments of the Reviewer 1

Major comments

1. The overall modeling goal is never fully clear. In the abstract you write "We present a new method based on the spatial skew-t process." But a new method for what? Reading the paper there are hints that you the aim of the modeling procedure might be to predict the probability of exceeding 75ppb (e.g. first line of the abstract), but this is then taken as one of your modeling thresholds (p.3 l.5–6; p. 15 l.5). In a new version of the paper, the modeling aims should be stated clearly upfront so that the results can be assessed agains these aims.

2. There is a lot of confusion about the "threshold", and what this means in various contexts. Sometimes you discuss this as fixed and specified by regulation (e.g. p.3 l.5) but other times you refer to the role of censoring as not allowing low-moderate values influence the fit (e.g. Section 3.1 or the stated goal of modeling "spatial extremes" in the first line of Section 3) even though you state that the constant threshold is not extreme everywhere (e.g. in the Abstract, and p.15 l.3 where in fact the highest threshold you use in the application is the 0.06 quantile at some sites, and so hardly the tail!). The role of the threshold needs to be much more clearly explained in the context of the application (fixed thresholds of interest), and modeling the extremes.

3. Section 3.2: you introduce a Poisson process as a device for allowing long range asymptotic independence. In fact, after its first mention on p.7 l.1 you never refer to a Poisson process again, but instead use a fixed number of knots, i.e. a uniform distribution. Amazingly this is not even commented on! The purpose of the Poisson process seems to be for the proof in Web Appendix C, so if spatial long-range independence relies on this feature (which, assuming a constant rate would eventually guarantee an incease in the number of points, as opposed to taking a fixed number of draws from a uniform distribution) then the discrepancies between these two assumptions ought to be thoroughly discussed.

   The proof in Web Appendix C also needs some greater attention to detail. It looks to me as if using defintion (7) and for a fixed radius h then $\mathbf{s}_1$ and $\mathbf{s}_2$ need not be in different partitions almost surely, although this may be the case as $h \to \infty$. Perhaps I am wrong, but either way it would be helpful to have the full proof spelt out to avoid any confusion.

4. Section 3.3: As you note, there are several places where you could introduce temporal dependence. Are you able to analyze any theoretical properties of the suggested form, or even how this matches the data?

5. Section 5: The simulation study needs more details. For example, when fitting datasets (1)–(4) to model (6) (max-stable), were the margins of this estimated or assumed fixed? Also you fit trend terms in some models when the mean is constant, and do not comment on the rationale for this or what happens to the estimates of these parameters (are β_{lat} and β_{long} close to 0 and β_{int} close to 10?)

   Furthermore the range of models simulated from is rather limited; for (4) you could simulate from a range of different max-stable processes (even if you dont fit these, it might be interesting to see how the skew-t model handles them).

   It would be nice to see alternative model evaluations to the Brier score, or at least perhaps gain some sense of whether there is strong spatial variation in the score. If the model is much better at predicting in some locations than others, then one may want to be wary of using it in some areas...

6. Section 6: why do you only use July data? Is this a computational restriction, due to availability of data, or due to particular interest in that month? As you point out, if you used more data then you would be able to say more about the extremal dependence, so this point is rather pertinent. Indeed, although Figure 6 shows that two particular sites close together are relatively strongly dependent, we see nothing about the general picture of $\chi(h)$. I agree this would be difficult to assess with so few data, but its not clear whether you have more data available to help assess this (even if you didnt fit the model to all of them).

Minor comments

1. p.5l.6 $\gamma \to 1 - \gamma$.

2. p.8 l.8 should $w_{t,1}$ etc. be bold?

3. p. 8 final line: the phrase "use a copula" could be omitted as it seems superfluous and may introduce unnecessary confusion. Giving adequate details on the transformation and dependence structure is sufficient.

4. p. 12 final line: you fit the correct model in some of these cases, so is it still difficult to tell whether the parameters converge even in this situation?

5. p.13 l.13 and l.15 please give references for these tests

6. p.14 eq.(20): Do you specify the mean of Y, or use the covariate specification given in (1)? These dont look to be the same if $\lambda \neq 0$.

Response to the comments of the Reviewer 2

Major comments

1. Throughout the paper I see an inappropriate use of terminologies which gives an idea that the paper is not carefully written. Examples are:

   i. Page 4 at the beginning of Section 2.1 you write that $Y(s)$ is an observation at the spatial location s. Then you use the same symbol to define a random field (spatial random process) in equation (1). I think you mean that $y(si), i = 1, 2, \ldots$ is an observation while $Y(s;) \equiv Y(s)$ is a random function for s varying on a certain set. A distinction between observations and random objects which represent the phenomena that you are interested on, must be done.

   ii. Page 5, equation (3). You used the $h = ||\mathbf{s}_1\mathbf{s}_2||$ (above equation (4)) to denote the distance between two specific fixed points and then $h = ||\mathbf{st}||$ to denote the distance between two generic locations. The latter (more general) should also be used in equation (3) to avoid confusion;

   iii. Page 7, below equation (7). The way you defined the partition is not clear to me. Suppose that we consider two spatial knots. Then, according to your definitions we have the two partitions:

   $$P_k = \{s \in \mathcal{D} : k = \arg\min_i ||sw_i||\}, \quad \text{with} \quad k = 1, 2.$$

   The index of the partition is also used to denote the value of the radius. Does this mean that we also consider all the values $s\mathcal{D}$ such that for a given w_k we have that $\min ||sw_k|| \leq k$? In this simple example, is $\mathcal{D}$ formed by two disjoint sets?

   iv. Second, you said: "all $z(s)$ and $\sigma(s)$ for sites in the sub-region k are assigned common values z(s) = zk and (s) = k". I guess you mean that for all $s \in P_k$, with $k = 1, 2, \ldots, K$, the functions $z(s)$ and (s) are equal to the constants z_k and σ_k, respectively.

   v. Finally, from the process definition in (6) $Y(s)$, the partitions in (7) and the elicitation of the random function in (8), why should we come to the conclusion that for locations that belong to different partitions, e.g. $s_i \in P_k$ and $s_j \in P_l$ with $P_k \neq P_l$, their distribution is not skew-t anymore? You have only specified that within the same partition the distribution is skew-t, but what happens in the contrary case? In addition, from (9) it seems that for locations that belong to different partitions the distribution of the process is Gaussian. Is this correct? A clear explanation of this part is needed, adding the required steps.

2. Noting differences in notation between Beranger et. al (2015), Azzalini (2013), and Azzalini and Capitanio (2003), the reviewer asks

   i. Regardless of the different parametrization, i.e. you used $\sigma^2 \sim IG(a, b)$ instead of $VG(\nu/2, \nu/2)$, the formula should be equivalent. It seems that there is a missing term in your formula, if not you should explain why there is such a difference and refer to where you got that expression. My suggestion is that you should explain well, with the two steps (1) and (2) above, the process construction;

   ii. Second, in your equation (1), $Y(s)$ is a stochastic process, a random function defined on $\mathcal{D}$. Mathematically speaking, what is $\mathbf{X}(s)^\top$? You only say that it is "a set of spatial covariates at site s", but this is not enough. I guess that $\mathbf{X}(s)^\top$ should be a vector-valued or a matrix-valued function (a vector or matrix of functions). But if so, how should the product with the vector β be interpreted?

3. The section on the spatio-temporal extension is hard to follow and the technical details are not immediate to verify, in the present form. The presentation should be simplified. Is it not much easier that you define first the AR(1) time series and then you define appropriate transformations that link these to the functional parameters of the skew-t process and you clearly show that the finite dimensional distribution of the resulting process including time and space is skew-t?

4. Section 5 Simulation study. The current session should be shortened. Furthermore the section should be divided into two parts. Before showing the performances of your method in predicting exceedances, you should show by means of a simulation study the ability of your approach when estimating the model parameters. Since that you are considering also replications over time it should be possible to estimate reasonably well the slant parameters and the degrees of freedom. It may be the case that you have to consider larger samples. Have you tried to consider alternative parameterizations? See e.g. Arellano-Valle and Azzalini (2008). You should investigate more seriously this aspect.

Minor comments

1. Page 4, first paragraph of Section 2. The references Beranger et al. (2015), Azzalini (2013) page 129 and Azzalini and Capitanio (2003), should be added for a clearer explanation of the additive representation;

2. Page 4, equation (1). You used the symbol "T" at least 3 times, e.g. in order to denote: transposed operation, a threshold (at page 6), a time index (at page 8). Different symbols should be used to avoid confusion;

3. Page 4 below equation (1). Specify what the size of $\mathbf{X}(s)^\top$ and of β are. Are you sure that $\lambda \in \mathbb{R}$? Is it actually $\lambda \in (1, 1)$?;

4. Page 4 below equation (1). A symbol for the correlation function of the process $v(s)$ should be used, e.g $\rho(h)$, in order to justify the presence of the matrix Σ which is obtained from $\rho(h)$;

5. Page 4 below equation (2). The size of the vector and matrices should be specified. Furthermore, specify that matrix Σ is obtained from $\rho(h)$ and this part should be jointed to equation (3) for brevity. Finally Ω is the dispersion matrix of Y;

6. Page 5, equation (3). What does $||\cdot||$ mean?

7. Page 5, equation (4). Since you are specifically referring to the skew-t process, why dont you simply say that $c = +\infty$? You should also confirm once more that for all given $\mathbf{s} \in \mathcal{D}$, $Y(\mathbf{s})$ is a skew-t random variable, to avoid confusion;

8. Page 8, two rows below equation (10). The elements of the sequence $w_{t1}, \ldots, w_{tK}$ should be bolded.

9. Page 11, Section 5, 8 lines from the bottom. You wrote "the $v_t(s)$ terms were generated using...". This is incorrect, $v_t(s)$ is not an observation. According to (10), $v_t(s)$ is a stochastic process.

10. Appendix C, you use $\mu(\cdot)$ to denote the intensity measure of the Poisson point process but then μ is also a constant in the definition of the intensity measure. You should use different symbols.