Point-by-Point response to Reviewer Comments
Biometrics MS #150926M
June 30, 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?

   We have assumed that the errors $v(\mathbf{s})$ have unit variance, and $z \sim N(0,1)$, so we require separate a and b so that a can control the shape/degrees of freedom while b controls the scale. Viewed this way, if $a = b$, then the degrees of freedom would determine the scale which is clearly too restrictive. To help clarify this, we have slightly modified the parameterization so that the degrees of freedom are a, and added additional clarification of the role of b in Section 2.1. We have also added Web Appendix E to demonstrate our more general model with $a \neq b$ is a skew-t distribution with a degrees of freedom.

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 $a \sim \text{Gamma}(0.1, 0.1)$ 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?

   In the original submission, the priors for a and b were inadvertently switched, so a should have had a discrete distribution over a mesh from 0.1 to 10 with spacing 0.1, and $b \sim \text{Gamma}(0.1, 0.1)$. With the reparameterization given in the response to Major Comment 1, this actually means that the prior for a is now a discrete distribution over a mesh from 0.2 to 20 with spacing 0.2. As we now clarify in the specification of these priors both below equation (1) of the main text as well as in Section 5.1 when describing the models we fit, a is the degrees of freedom, and so this prior should be uninformative in all applications. In contrast, b is a precision parameter, so its prior should be adapted to the scale of the response. In the real data analysis, we use the same discrete prior as in the simulation study, and find that for all settings in the cross validation, the posterior mean of a is less than 3.

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 λ parameter before adopting this reparametrization? It affected only λ?

   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.

   > We agree this was unnecessarily confusing. We have opted to use the parameterization in your suggestion where $\lambda \sim N(0, \sigma^2_\lambda)$.

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.

   > This is a key argument in extreme value analysis which we realize we inadvertently took for granted. This is a rendition of the classic bias variance trade-off. A low threshold reduces variance by increasing the amount of data used to estimate parameters but risks inducing bias if the model is misspecified. On the other hand, a high threshold mitigates potential bias because only the tails must be modeled. Therefore, by using a high threshold fewer precarious assumptions are required, but this increases variance by discarding data. For our fairly flexible skew-t process, we found that a moderate threshold optimized this balance for both simulated and ozone data. Section 3.1 now elaborates on this trade-off.

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.

   > Using simulated data, we have evidence to suggest that the model exhibits extremal dependence over time. We are unable to give a formal proof of this, but we have included simulated χ plots for temporal dependence in Web Appendix F.

Minor comments

1. Page 3, line 11: "... with computing on the order of Gaussian models "

   This has been corrected

2. Page 3, line 13: ... as well as TO make predictions ...

   This has been corrected

3. Page 3, line -2: ... compared TO Gaussian AND max-stable ...

   This has been corrected

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?

   As we now clarify in the final sentence of Section 3.3, by construction the entire process including the prior on the partition is the same as the space-only model of Section 3.2

5. Page 9, line -5: "the marginal distributions are Gaussian": which distributions, those of $Y(\mathbf{s})$?

   Yes, we now clarify that we are referring to the distribution 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.

   This typo has been corrected.

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?

   In a spatial analysis, the primary objectives are to make spatial predictions and to estimate extreme probabilities. Both of these objectives are met by marginalizing over partition-specific parameters and are thus unaffected by label switching. We have added a comment to this effect in the Section 5.1.

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.

   We have added several references to the introduction.

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 against these aims.

   As we now describe at the end of the first paragraph of the introduction, two objectives for our data analysis are spatial prediction and mapping the probability of an extreme event. In the extremes context, spatial prediction means using the values of monitored stations to determine whether an unmonitored site experienced an extreme event. Mapping the probability of extreme events could either be the marginal probability of an event, such as ozone exceeding 75ppb, or a map of an upper quantile, such as the map of the 99th percentile in Figure 7.

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.

   Yes, we can see how this could lead to confusion. In the revision, we now refer to the "threshold" as the value we select in the model-fitting process, i.e. T in Section 3.1. We also bolstered the description of the trade-off involved in selecting T (see AE comment 4) in this section. As mentioned in the response to the previous comment, we also want to compute probabilities of extreme events (e.g. ozone exceeding 75ppb). We now refer to this as the probability above a high *level*, L. That is, we might use the data above threshold $T = 50$ppb in order to estimate the probability of the response exceeding level $L = 75$ppb. Of course, this is only reasonable if $L \geq T$, and the user is free to select the T that gives the best estimate of this probability.

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 μ 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.

Thank you for pointing out this important difference. Below (7), we now clarify that $\pi(h)$ goes to zero as h increases assuming that the knots follow a homogeneous Poisson process. We also state (end of Section 3.2) that in practice, the number of knots is fixed and selected by cross-validation.

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.

We have provided a more complete proof that $\lim_{h\to\infty} \pi(h) = 0$ in Web Appendix C.

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?

   Yes, as noted in the response to AE (major comment 5), we have evidence using simulated data that our temporal construction maintains extremal dependence. This now appears in Web Appendix F. Thank you both for the suggestion.

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).

   We have added clarification for these points in the description of the simulation study. The parameters in the marginal distribution for all methods were estimated in a fully-Bayesian manner using MCMC methods. The constant mean function was chosen somewhat arbitrarily, but permits a common threshold to apply to all sites. Finally, we have also added a 5th design to the simulation study using a Brown-Resnick process to generate data, and find that the skew-t method performs similarly to the other setting. For this new setting, we find that our model performs similarly to the case in setting (4) with some improvement over the max-stable method for lower quantiles (0.90 – 0.98), and the max-stable method tends to perform better at more extreme quantiles.

   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...

   We agree that maps of Brier scores would be informative. This information could conceptually inform future monitor sites. We have now added a map of Brier scores

(see Web Figure 3), which shows that performance is similar throughout the US with the most varied performance in California.

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).

   Ozone exhibits strong seasonal patterns, so analyzing a longer series would require additional parameters. Additional monitoring data would be easy enough to gather because it is available online. The real limitation in our case is the CMAQ data, which we only have for this time period.

Minor comments

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

   This typo has been corrected.

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

   This typo has been corrected.

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.

   We have replaced "use a copula" with "use a transformation of a Gaussian random variable" above (12) and (11).

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?

   As with any Bayesian mixture model, the labels of the partitions are arbitrary. That is, we could swap the labels of knots 1 and 2, and this would not affect the value of the posterior. It would, however, change the values of z_1 and z_2 making it difficult to study convergence for these parameters. However, since the quantities of interest (spatial predictions and probability of extreme events) are defined marginally over the partitioning, label-switching does not affect convergence for these quantities. We have bolstered this explanation and included coverage probabilities for λ and the degrees of freedom for design (2) and method (2) (skew-t with 1 knot) at the end of Section 5.1.

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

   These have been added

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 $\lambda \neq 0$.

   Yes, technically the mean of $Y(\mathbf{s})$ depends on the shape parameter. To match this equation with (1), we now simply specify that $\mathbf{X}_t(\mathbf{s}) = [1, \text{CMAQ}_t(\mathbf{s})]^\top$

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 $y(s_i), i = 1, 2, \ldots$ is an observation while $Y(s;\omega) \equiv Y(s)$ 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.

   > Conceptually we can imagine that $Y(\mathbf{s})$ is a random field that is only observed at n locations as described in and around (3). This notation could be problematic if the data consisted of multiple measurements at some sites that gave different values, but this is not our case and so we elect to retain this streamlined notation which is used, for example, throughout Gelfand et al. (2010).

   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 $h = ||\mathbf{s} - \mathbf{t}||$ to denote the distance between two generic locations. The latter (more general) should also be used in equation (3) to avoid confusion;

   > Thank you for pointing this out. We now use $\mathbf{s}_1$ and $\mathbf{s}_2$ in (2), (4), and (7) as opposed to $\mathbf{s}$ and $\mathbf{t}$ to avoid confusion between space and time.

   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:

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

   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 $s \in \mathcal{D}$ such that for a given w_k we have that $\min ||s - w_k|| \leq k$? In this simple example, is $\mathcal{D}$ formed by two disjoint sets?

   > Yes, P_k is not the set of points on the circle with center $\mathbf{w}_k$ and radius k, it is the set of points that are closer to $\mathbf{w}_k$ than any of the other knots. Therefore by definition $P_1, \ldots, P_k$ partition the spatial domain. This is clarified after (6).

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

   > Yes, we now use your phrasing.

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.

> Setting aside the skewness parameter, the joint distribution of observations in different partition sets is that of $Y_1 = \sigma_1 e_1$ and $Y_2 = \sigma_2 e_2$ where σ_1^2 and σ_2^2 are independent inverse gamma and (e_1, e_2) is bivariate normal. If σ_1 and σ_2 are fixed, then (Y_1, Y_2) is bivariate normal, and if $\sigma_1 = \sigma_2$ is random then (Y_1, Y_2) is bivariate t marginal over σ_1. However, if σ_1 and σ_2 are different and unknown, then $\mathbf{Y}$ is neither Gaussian nor t. From our perspective (as we now mention above (7)), however, the important feature of this distribution is that it has asymptotic independence, which is sufficient to break the long-range dependence of the usual spatial-t process.

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 $IG(\nu/2, \nu/2)$, 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;

   > As discussed in the response to the Associate Editor's Major Comment 1, given the confusion shared by all reviewers, we have added clarification about the role of b as a scale parameter and provided Web Appendix E which shows the connection between our parameterization and the more standard parameterization.

   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?

   > $\mathbf{X}(\mathbf{s})$ is the usual covariate vector as in multiple linear regression. This is now made explicit below (1). Then the coefficient has the linear regression interpretation that an increase of one in the jth element of $\mathbf{X}$ shifts the distribution of Y by the jth element of $\boldsymbol{\beta}$.

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?

We have heeded your advice and swapped the transformation equations (now equations (10) – (12)) and the latent time series definitions (now equations (13) – (15)).

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.

   In simulation design (2), we generate the data using a skew-t distribution with a single knot, and in method (2), we fit the data using a skew-t distribution with a single knot. We have provided additional clarification in Section 5.1 that in this case, we do appear to correctly estimate the skewness parameters and degrees of freedom. We have also added clarification that the difficulty in assessing convergence is only relevant for models with multiple partitions (i.e. models 4 and 5) specifically for the parameters that change across partitions. Because the parameters in our model do appear to properly converge when the design matches the method, we did not consider a centered parameterization as in Arellano-Valle and Azzalini (2008).

Minor comments

1. Page 4, first paragraph of Section 2. The references Beranger et al. (2016), Azzalini and Capitanio (2014) page 129 and Azzalini and Capitanio (2003), should be added for a clearer explanation of the additive representation;

   We have added these references to clarify the 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;

   The revision uses $^\top$ for transposes and n_t for the number of time points.

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 $\lambda \in (-1, 1)$?;

   Below (1) we clarify that $X(\mathbf{s})$ and β are p-vectors. Yes, we use the same λ as given in equation (5.20) of Azzalini and Capitanio (2014, p. 129), so $\lambda \in \mathcal{R}$.

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 $\rho(h)$;

   At the end of the sentence following (1), we now define correlation function $\rho(\mathbf{s}_1, \mathbf{s}_2)$ and indicate the ρ we use to populate the correlation matrix in (2).

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;

   We have given the sizes of the vectors and matrices below equation (3) (previously equation (2)). Based on the Minor Comment 4, we have now moved the $\rho(h)$ specification to (2), and we now indicate after (3) that $\mathbf{\Sigma}$ is obtained from $\rho(h)$.

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

   As we now clarify below (2), this is the Euclidean distance between locations $\mathbf{s}_1$ and $\mathbf{s}_2$.

7. Page 5, equation (4). Since you are specifically referring to the skew-t process, why don't 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;

   Here we provide a general definition which allows for a bounding density. We now immediately clarify below (4) that for the skew-t distribution, $c^* = \infty$.

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

   This has been fixed

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.

   Yes, $v_t(\mathbf{s})$ is a stochastic process, and simulating data from the model requires generating v at the data points from a multivariate normal distribution.

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.

    We now say throughout the proof in Web Appendix C that the point process is a homogeneous Poisson process with intensity λ_{PP}.

References

Arellano-Valle, R. B. and Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis* **99,** 1362–1382.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 367–389.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Beranger, B., Padoan, S. A., and Sisson, S. A. (2016). Models for extremal dependence derived from skew-symmetric families. *ArXiv e-prints* .

Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.