1 Executive Editor's comments

Dear Heße,

We are pleased to accept your paper "exPrior: An R Package for the Formulation of Ex-Situ Priors" subject to the major revisions described below.

We would appreciate a revised version and point-by-point response to the reviewers comments within 2 months. Remember, that when responding to the reviewer's comments, your job is to persuade me, the editor, that either you've dealt with the issue, or that it's not relevant. To this end, please produce a single document that includes all the reviewers comments mingled with your responses. I particularly recommend the strategy described at http://matt.might.net/articles/peer-review-rebuttals/.

Regards,

Catherine Hurley

We appreciate the chance to produce a revised version of the manuscript. Below, we address each of the points raised by both reviewers and explain how we revised the manuscript to address them. Unfortunately, we were not able to produce a manuscript with the changes being highlighted.

Reviewer #1

The authors present exPrior: An R package for the Formulation of Ex-Situ Priors. The paper is well written is easy to follow and the explanations of models in the paper are clear. The approach used in the prior estimation based on the ex-situ data is reasonable and relatively straightforward. At it's simplest the authors implement a standard hierarchical modelling structure where site-specific parameters are informed by a global parameter allowing for cross-site variation. There are further modelling options for taking spatial autocorrelations into account, which is important for potential applications. In addition, the functionality outlined with Example 4, related to assimilating data from bounds and moments, is a useful feature. Overall, I think this could be a useful package. However, I have some concerns in relation to a users expectation of what the package can produce vs what the package actually produces. Therefore, I have some comments/questions that I would like the authors to address. Therefore, my recommendation is to Accept with Revisions. My comments (denoted C1-C9) are given below.

We appreciate the feedback of the reviewer. His/her criticism has helped to identify current problems of the manuscript, which we were consequently able to improve on. In the following, we go through each of the points raised by the reviewer and explain how we revised the manuscript accordingly. Unfortunately, we were not able to produce a manuscript with the changes being highlighted. Instead, we point towards the portion of the manuscript, where the changes were made.

- C1: In the abstract the authors mention that the model can be used to "assimilate multiple types of data" but I think this is a bit misleading. The authors are referring to the fact that the model can assimilate data coming from measurements, bounds or moments which becomes clear in Example 4. However, when I read this in the abstract without that context I assumed that 'multiple types of data' referred to categorical, discrete or continuous. It appears to me that the model can only work with continuous data. Perhaps saying that "the model can be used to assimilate various summary measurements associated with continuous data" would be more appropriate?
  We agree that our model can only assimilate continuous data. However, for our field of target application, ie. geostatistics in hydrology, hydrogeology, soil sciences etc, we do not consider this to be a limiting choice. In this field, the relevant modeling paradigm for uncertain quantities is the Gaussian process model, which models the unknown quantity as a continuous variable. In the revised version of the manuscript, we explain this issue now in the introduction to avoid this misunderstanding in the future.

- C2: In the introduction it says that the package produces prior distributions for geostatistical parameters. I think the term "geostatistcial parameters" needs to be more clearly defined for reasons outlined in comments

below.
We appreciate this feedback and changed this to geostatistical quantities. How to derive geoststistical parameters for, say, a Gaussian process model, will be explained in our answer to the next comment.

- C3: On page 2 (Ex Situ Priors section) the authors write: "Yet this would leave out any spatial correlations, so most geostatistical analyses try to account for them by using spatial random field models, say a Gaussian process (Rasmussen and Williams, 2006; Gelfand and Schliep, 2016). Since such models are fully defined by their parameter vector the aim of Bayesian inference is to use available, in-situ data y_{in} and derive the posterior distribution over these parameters $P(\cdot|y_{in})$." The authors then go on to discuss how a prior for $\cdot$ can be generated using ex-situ data. The package produces $P(\cdot|y_{ex})$. However, $\cdot$ describes the model implied outcome for the observed data y_{ex}. So, I am not currently seeing how $P(\cdot|y_{ex})$, obtained using exPrior can provide prior distributions for the parameter vector characterising a GP, for example. I think the intention is for users to just use this posterior predictive distribution as the foundation for informing priors for specific parameters in potential models for y_{in} (e.g., a GP model). I think this should be made more explicit. Based on this, I think readers would benefit from a clear link being made between the posterior predictive distribution $P(\cdot|y_{ex})$ produced by exPrior and how priors for the parameter vector of a GP (or some other model example) could be obtained.
  It is true that the examples given in the manuscript only refer to the distribution of the expected quantity itself, eg. porosity. This does however not mean that the parameters of a GP cannot be inferred. To explain how, one must distinguish between two kinds of parameters; one-point statistics like mean and variance and two-point statistics like characteristic length scale and anisotropy. As regards one-point statistics, we can already use the code developed in the package to get the predictive prior for the mean and variance of the GP. Since μ and σ are nodes in the hierarchical model, an MCMC is computed for these parameters and the predictive prior can be computed by evaluating the corresponding statistical distribution. With the current formulation of the hierarchical model and the current implementation: (a) σ is a hyperparameter so its predictive prior distribution can be directly found under the `d_hyperPar` value in the output of the `genExPrior` function (b) μ is not a hyperparameter, it is defined by its mean α and its standard deviation τ. α and τ however are hyperparameters so their posterior predictive distributions are computed and available in the `d_hyperPar` value in the output of the `genExPrior` function. It is possible for a user to sample from these distributions and then sample from distributions parameterized with sampled α and τ in order to calculate the posterior predictive prior for μ. As regards two-point statistics, generally speaking they do not need a hierarchical model to be inferred. This is due to the fact that in the vast majority of applications these parameters are

assumed to be the same over the whole site. This means that in order to get a proper prior distribution for, say, the characteristic length scale, a histogram or a KDE is sufficient. Such a non-parametric estimation procedure may possibly be weighted by similarity measures to further reduce the uncertainty (more on this below). While this assumption of a stationary covariance model for a whole site is certainly a simplification, it is a standard modeling choice. Furthermore, any modeling scheme that would try to infer, say, a series of two-point statistics for a given site would need an overwhelming amount of data. Not only are these data not available, it is even difficult to get enough data for a simple histogram. We are currently in the process of adding data on such two-point statistics to the database to perform at least such simple inferences. Getting, however, enough data to infer more complicated models than that, will take a long time.

- C4: On page 3 (Formulation of the hierarchical model section) the authors write: "This general formulation allows to flexibly choose parametric models used for all distributions. Since $P(y|\cdot)$ represents the data, this distribution should fit the empirically observed frequencies of y. Depending on the geostatistical parameter of interest, a user can, e.g. use the normal, log-normal, multivariate normal or truncated normal distributions to model parameter behavior." I agree that there is some flexibility here but this wouldn't be suitable, for example, if you have discrete data? Perhaps limitations could also be mentioned. Related to flexibility, can the model be extended to include covariates other than spatial coordinates?
  As already mentioned above, the package is tailored toward the needs of practitioners in the fields of hydrology, hydrogeology soil sciences, etc. Within this context the relevance of discrete data are comparably low, which is why they are omitted from consideration. The second point, regarding covariates is, however, more interesting. exPrior itself does not feature this possibility. However, there is an R package called siteSimilarity (which is developed jointly with exPrior) that helps to do exactly that. In this package, we use the available covariates for every site in the databank and use them to determine a similarity index between the ex-situ sites and the in-situ site under consideration. Then, only those sites similar to the in-situ site can be used for the computation of the prior, therefore potentially decreasing the resulting uncertainty. Our current results show only modest improvements, when doing a leave-one-out validation, which shows that the amount of data is currently too modest for large improvements. We strongly revised the manuscript by adding a complete new section to the manuscript. In it we detail the connection between the exPrior and the siteSimilarity package and how they work together.

- C5: Page 4 (The Generation of the ex-situ prior distribution section) is the first (and only?) time the $p(y_{in}|y_{ex})$ notation is used. Previously we have only seen this as $p(\cdot|y_{ex})$ (e.g., eq 1). Is there a need to switch notation

here?
This was indeed confusingly formulated. We removed this in the revised version of the manuscript to avoid further confusion.

- C6: On page 8 (Example1: Using exPrior with synthetic data section) the authors write: "The second command plotExPrior shows the ex-situ data from the three sites jointly with the predicted prior distribution for the new site S0 (see Figure 5 right panel)." The right panel plot does not show the data from the three sites as indicated in the text. I assume it should be similar to Figure 7?
That observation is correct. The figure initially showed the data along with the prior distribution as in Figure 7. In order to avoid presenting too many information at once, we took it out, but without adapting the text. To make both consistent again, we changed the caption of the figure accordingly.

- C7: Related to $p(\cdot|y_{ex})$ produced by exPrior on page 9 (Example 2: Using exPrior with real-world data section) the authors write: "Using this prior therefore provides a practitioner with a sound foundation for the geostatistical inference of the in-situ porosity." I think this sentence is important related to the points i've made above. It should be stressed that results provide a foundation for inference with the in-situ data and this is indeed very useful but more work is required to produce priors for specific in-situ data model parameters. Perhaps this section would be a good place to provide the reader with more context about how they could take the prior produced by exPrior and use it practically to inform prior distributions for parameters that characterise a model for the in-situ data.
We are not sure whether we agree with the assessment of the reviewer. It seems clear to us that the predictive prior distribution is the prior distribution for the in-situ data. Why would it not be? Perhaps there is a misunderstanding here that needs clarification?

- C8: In the introduction the authors write: "Yet there is no package to date, which would provide such tools with the necessary foundation, i.e. prior distributions for the geostatistical parameters." I think in this specific context this is true. However I do think the rstanarm package could potentially produce something very similar to what the authors have presented here, at least potentially for Example 1 and Example 2. Specifically rstanarm::stan_glmer could be used to fit a model with group-specific intercepts similar to that described in Eqs 2a-2c. Then rstanarm::posterior_predict could give you draws from the posterior predictive distribution. It might be worth comparing the results for Example 2 with results from using these rstanarm functions.
We agree that it is possible to produce similar results with the rstanarm package. But the same is also true about any other R package that facilitates the user to do hierarchical Bayesian modeling, like BUGS, JAGS or the nimble package that our package is based on. This is, however, not a

drawback since we specifically see our exPrior package as an application package, ie. we target practitioners of geostatistics without a strong background in Bayesian hierarchical modeling. This is the very reason why we labeled this manuscript as an application paper. To make this point more clear, we revised the introduction of the manuscript accordingly.

- C9: I think it would be useful if exPrior could provide the user with summary statistics for · (mean, median, sd etc). Access to the posterior samples themselves might be also be useful, similar to what is produced using rstanarm::posterior_predict.
  These are two important comments, so let's address them consecutively. As regards the first point, we revised the exPrior package to provide the user with summary statistics, as the reviewer recommends. Since we regard them as the most useful, we provide now the user with a quick access to mean, median, mode and standard deviation. Say, you have a fully inferred prior distribution in the exprior object. Then, these statistics can be accessed via exprior$exPrior_summary Currently, this patch is only available in the `summary-stats` branch of the `https://github.com/GeoStat-Bayesian/exPrior` repository. If these changes meet the reviewer's request, we will merge this into the main branch and create a new release, both on GitHub and CRAN, jointly with this manuscript. As regards the second point, exPrior – in its current form – does already provide the user with samples from the prior distribution by virtue of the nimble package that exPrior is build upon. These can be accessed through the MCMC attribute of the exprior object.

Reviewer #3

Review of "exPrior: An R Package for the Formulation of Ex-Situ Priors" by Falk Heße, Karina Cucchi, Nura Kawa, and Yoram Rubin

The overall premise of informing priors using external information is obviously a key attribute of Bayesian inference. With a few exceptions noted below, the authors are right that most existing geostatistical R packages, e.g., spTimer and spBayes, are tailored to specific modeling tasks and don't allow for hierarchical structures on the covariance parameters (aside from routine priors, e.g., inverse-gamma, uniform, etc.). While there is nothing stopping users from analyzing external data to inform hyperparameter selection for the covarinace parameters' priors (e.g., it is routine to do EDA to set the nugget, partial sill, and range priors in geostatistical models), specifying explicit models for these parameters are not available unless you move to a more general coding environment, e.g., INLA, BUGS, etc.

We appreciate the feedback of the reviewer. His/her criticism has helped to identify current problems of the manuscript, which we were consequently able to improve upon. In the following, we go through each of the points raised by the reviewer and explain how we revised the manuscript accordingly. Unfortunately, we were not able to produce a manuscript with the changes being highlighted. Instead, we point towards the portion of the manuscript, where the changes were made.

Below is a list of general comments and questions:

- 1) It is not clear from the manuscript, why someone would use the exPrior package over coding up the corresponding hierarchical model in openBUGS, JAGS, or NIMBLE. What is the value added by this package? Please take this question as a constructive suggestion for framing the revision.
  This is an important feedback, since the reviewer is right that our package is based on nimble and its functionality can therefore be achieved with this or similar packages. To clarify, our package is supposed to provide a ready-to-use software tool for assimilating ex-situ data into ex-situ priors. The intent is to encourage the use of informative distributions for practitioners in geosciences that might not be experts in Bayesian hierarchical models. This is the very reason that the manuscript was labeled as an application paper. To make this point more clear, we strongly revised the introduction of the manuscript.

- 2) The manuscript provides very little information about what the package actually does. It is not clear that the external data is used in a full MCMC hierarchical model (i.e., one-for-one sampling across the model's hierarchical levels) or summarized as point estimates to be used as hyperparameters in priors (i.e., there is that language at the bottom of page 4 about providing data to genExPrior() as moments).
  We thank the reviewer for bringing this to our attention. We revised the

manuscript to better explaining the inference performed by exPrior and by pointing to the mechanics of the underlying nimble package.

- 3) The manuscript opens with very specific language and discussion about geostatistical models, but most of the examples don't use spatial data. The one spatial data example doesn't provide much detail about the effectiveness of using external data to inform covariance parameters. A clear demonstration on what is gained by informing the partial sill and range parameter would be useful. It is just not clear how the external data is used to inform the spatial covariance parameters' priors. I recognize the value in informing priors using external information, but there are just not enough details provided here to understand how this is actually implemented. The authors might also explore or note, some of the dangers in using external data to inform priors in models (spatial and otherwise). What happens if the covariance parameters estimated from the external data are quite different from those in the target population? What are the implications to prediction with getting the priors wrong?
  There are at least two different points here raise by the reviewer. Let us start with the first one about inferring the parameters of the covaricance model. This point is also raised by the other reviewer, which demonstrates its importance. Repeating the above reply, we do not consider the inference of specific covarianve parameters (like length scale or anisotropy) to be a necessary feature of the package for the simple reasons that they are, usually, not hierarchical. This is to say that the clear majority of current practical applications assumes a single covariance model to be valid for the whole site. To get informed prior distributions for, say, the length scale (alternative called correlation length/range/integral scale ect.) a simple histogram of relevant ex-situ data would be necessary. This is e.g. the way we currently perform Bayesian inference with the parameters of a Gaussian process model. This is to say that we determine the prior distributions for mean and variance with exPrior, whereas prior distributions for length scale and anisotropy are calibrated against literature data. While more complicated models are known in the literature (as the reviewer for example cites), these models do, to the best of our knowledge, lack both from applications in practice and the necessary amount of data to calibrate them. In particular the last point is a big problem, since any model that considers more than a single, stationary covariance model per site, would need a large amount of data to provide informative prior distributions. In our estimation, we are still quite far from a place where such data would become available. As regards the second point, we respectfully disagree. The whole point of Bayesian inference is not to get the right model parameters but to provide the best estimates of them given all available data. In that sense, those estimates cannot be wrong. What can be, and often does go, wrong is the choice of data being used for the inference. Yet, there is no royal road to avoiding this. Having access to good data sets of relevant measurements combined with years of

experience and a good sense of judgement are the only safeguards against this. We strongly revised the introduction of the manuscript to better convey these ideas and avoid future misunderstandings.

- 4) Geostatistics is so much broader than estimation of subsurface features. I would recommend a more general presentation of point-referenced spatial data.
  We agree with the reviewer that the field of geostatistics is much broader. However, our focus is on fields from earth science such as geology, hydrogeology, hydrology or soil science, since this is where our expertise is the strongest. Consequently, we are less certain about fields from the life sciences like ecology or econometric, although our package may find useful applications there, too. We now motivate this focus better in the introduction of the revised version of the manuscript.

- 5) The phrase "...the ongoing disregard of this source therefore further exacerbates the already pressing problem of subsurface uncertainty. The reasons for this disregard are hard to explain, since the use of informative prior distributions in this field is easy to motivate" is not accurate. I wouldn't say the authors of most spatial model methodology or associated software willfully disregard the value of the prior, but rather make a choice to allow users to define priors, hyperparameters, etc. In fact, there is quite an active literature on providing flexible and informed spatial and spatio-temporal covariance structures, see, e.g., Risser and Calder (2015) and associated software and other works referenced there in Risser and Turek (2020).
  This is a relevant point made by the reviewer concerning the overly generalized nature of our statement. It is true that most software packages on Bayesian inference allow the user to put in their own prior distributions. If this were not the case, the benefits of our exPrior package would be very difficult to make use of. To better reflect this notion, we revised this portion of the manuscript. As regards the covariance structure, we refer to our answer to the previous comment.

Risser, M. D. and Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. ., Environmetrics, 26(4):284–297.

Turek, D. and Risser, M. (2019). BayesNSGP:Bayesian Analysis of Non-Stationary Gaussian Process Models. R package version 0.1.1.

Risser, M. D. and Turek, D. (2020). Bayesian inference for high-dimensional nonstationary gaussian processes. https://arxiv.org/pdf/1910.14101.pdf