Interactive Comment on "Can river temperature models be transferred between catchments?" by Faye L. Jackson et al.

We are pleased that the two anonymous reviewers considered that the paper was "technically sound" and would be "of interest to the readership of HESS". We thank them for useful and thoughtful comments and for identifying additional literature which we will incorporate into a revised manuscript. For the moment we have tried to respond to their comments below and to identify where we believe the manuscript should be adjusted in light of the comments received to date.

Referee #1

Overall, the paper is technically sound.

- We are pleased to see the reviewer agrees that the paper is technically sound

A short review of the literature on the use of GAM in water temperature modelling would be a welcome addition.

- We are unsure whether there is a desire for a discussion of the use of GAMs more generally, or the use of GAMs in combination with river network smoothers to account for spatial covariance. If the latter, then we are only aware of one previous paper doing this (Jackson et al., 2016b).
- We would be happy to add a brief summary of the use of GAMs in previous studies of river temperature to our discussion.

In my opinion, the main weakness of the manuscript is in the discussion. Two main points need to be further discussed:

1. The challenge of inter basin transferability using air temperature needs to be further addressed and potential next steps identified. For instance, the readers may ask the following question: Is it the seasonal component or the residual component of air temperature that make it so difficult to transfer? Could an alternative model be envisaged in which the parameters of the air temperature seasonal harmonic be estimated/transferred?

- In this paper we focus explicitly on the issue of spatial variability in Tw during summer (as do many other studies). In this context, the seasonal variability is something to be examined at another time, although recent studies have shown that seasonal variability in local Tw~Ta relationships can indeed be modelled (Li et al., 2014). In this study we are trying to understand and predict the spatial variability in Tw_{max}, the value of Tw at a single point in time (hottest 7 day period) from covariates that include the corresponding value of Ta, namely Ta_{max}. For Ta_{max} to be an accurate (precise and unbiased) predictor of the spatial variability in Tw_{max} requires the Tw_{max}~Ta_{max} relationships to be broadly consistent between catchments. For example, if the within-catchment relationships between Tw_{max} and Ta_{max} are linear, the slopes and intercepts must be similar across catchments to use a Tw_{max}~Ta_{max} relationship developed in one catchment to predict Tw_{max} from Ta_{max} in a different catchment. Our study finds that the relationships are not consistent across catchments and that models using Ta_{max} developed in one catchment might not transfer well to another.

- It is well documented that (temporal) Tw~Ta relationships can vary substantially between sites and catchments due to the effects of other controls including discharge, groundwater – surface water interactions, hydrogeology and landuse (e.g. Tague et al., 2007). If such controls differ between catchments (for example if one catchment has a higher groundwater component than another) then this would likely lead to spatial relationships between Tw_{max} and Ta_{max} that differ between catchments (particularly since differences are accentuated at high temperatures).
- Because Ta and Tw respond to similar physical drivers (Johnson, 2003), Ta is a good predictor of the temporal variability in Tw once that relationship has been established at a site. However, in terms of processes, Ta is not the dominant control on Tw (e.g. Hannah et al., 2008) and there are therefore major challenges in predicting how this relationship should vary spatially. It is this understanding that is required to make Ta a good predictor of the spatial variability in Tw. The finding that Tw~Ta relationships vary widely between catchments (both temporally and spatially as we demonstrate for the Tw_{max}~Ta_{max} relationship) with varying characteristics is not new; it is simply that we have shown this to be a problem for predicting spatial variability in Tw from Ta when models are transferred.
- One approach that could be considered for future work would be to model the Tw~Ta relationship and then allow this relationship to vary spatially depending on other covariates that are known to affect this relationship. For example, in the context of the current paper, the intercept and slope of a linear Tw_{max}~Ta_{max} relationship could be related to catchment-scale landscape or hydrological covariates. However, such an approach would require more than the four catchments considered here. It is also important to remember that, in many cases, managers do not have access to data from a wide range of catchments and are forced to make management decisions about one catchment based on models developed in another. In this context, our paper suggests that models using landscape covariates would be more robust than those also using some Ta metric.
- Given that this issue is raised by both reviewers we will revise the paper to further emphasise why we would not expect Ta alone to necessarily be a good predictor of the spatial variability in Tw at a fixed point in time, when models are transferred between catchments or regions.

2. The problem of the impossible air-water temperature relationship at Bladnoch needs to be further explained. This is very unusual. I suspect that it is caused in part by station locations on this basin and by the fact that the samples used in the model only include air temperatures ranging between 18.5 and 20.5 deg C (figure 4)?

- We agree that the physically implausible relationship could be caused by the limited range of Ta_{max} which was much smaller in this catchment (due to its physiographic characteristics) than the others, coupled with the way the landscape and hydrological controls vary across the catchment and in particular act at the station locations. We will add in some text to reflect this. We note that, whilst there is sufficient evidence of nonlinearity in the data to force a smooth effect of Ta_{max} in the single catchment model, there is not sufficient evidence to do so in the multi-catchment model (when the evidence of non-linearity in the Bladnoch relationship is swamped by the lack of evidence of non-linearity in the other catchments). The slope of the Bladnoch Tw_{max}~Ta_{max} relationship in the multi-catchment model is negative (but not significantly so) (Figure 8c). However, our main point is that care is required when transferring Tw~Ta relationships among catchments (or regions).

Referee #2

In this paper, the authors explore the transferability of statistical models to predict a metric of maximum summer stream temperature. They use data from four catchments in Scotland collected during one summer season. Consistent relations with landscape variables were found; however, the relation between stream temperature and air temperature was inconsistent among catchments, and was even physically implausible in one. The authors conclude that, overall, the ability to transfer statistical models among catchments is limited without further research to gain a better understanding of intercatchment differences.

Considering the high level of concern about rising stream temperature and the increasing number of papers focused on modeling stream temperature over the last decade or so, the topic is timely and would be of interest to the readership of HESS.

- We are pleased that the reviewers considers that this paper will be of value to the readers of HESS and believe that we can adequately address the reviewers concerns below

I have a number of concerns about this work in its present form.

1. It is difficult to judge the novelty and significance of this work because the authors have not effectively placed it into the context of previous research on the topic. It is unclear what specific knowledge gaps are being addressed, or what new knowledge has been generated. Although the authors do cite a number of relevant, related studies (e.g., Hrachowitz et al., 2010; Chang and Psaris, 2013), they do not adequately address how their results are similar to or differ from those in previous studies. In addition, a number of papers not cited have addressed landscape-scale modeling of stream temperature, including Isaak and Hubert (2001), Scott et al. (2002), Tague et al. (2007), Wehrly et al. (2009) and Moore et al. (2013). Some of the previous papers have focused on extensive regions and thus have implicitly demonstrated that models based on landscape variables can be applied consistently across multiple catchments.

- We agree that we have not cited all of the papers that model Tw as a function of landscape covariates and/or Ta; rather we cite a selection of papers that illustrate the various approaches and statistical methods that have been explored. We are happy to add additional references where this broadens the discussion. None of the additional studies that have been suggested explicitly demonstrate the consequences of transferring models developed in one catchment to another catchment and none of the suggested studies consider network structure in the underlying models.
- Isaak and Hubert (2001) develop a spatial model for a single river (catchment), but do not assess the ability of this model to predict within or between (transfer) river catchments or regions. Similarly Scott (2002) investigates the effects of landscape covariates on river temperature (and water quality) in a single large river catchment (series of sub-catchments) but again did not assess between-catchment transferability of the resulting models.
- Tague et al. (2007) investigate Tw~Ta relationships across sites in Western Oregon and observed substantial differences in Tw~Ta relationships, much of which could be explained by large scale

differences in hydro-geology. This paper adds to the existing references already in our paper that suggest Tw~Ta relationships can be highly variable between sites and catchments and provides further support for our suggestion that Ta models may not transfer well from one catchment to another. Indeed the abstract for Tague et al. (2007) states "In this study we show that, in regions where groundwater inputs are key controls and the degree of groundwater input varies in space, air temperature alone is unlikely to explain within-landscape stream temperature patterns". We are more than happy to cite this paper, providing further support for our findings. However, we note that again this paper does not explicitly demonstrate the consequences of transferring Tw~Ta relationships between catchments as we do in our paper.

- Moore et al. (2013) and Wehrly et al. (2009) both fit large scale Tw models using landscape covariates and air temperature as predictors, the former focussed on maximum mean weekly temperatures and the latter on mean July temperatures. Wehrly et al. investigate a range of statistical models, including linear mixed models that account for (Euclidean) spatial structure, and find that across the large spatial scales investigated, the models that considered spatial structure performed best (but not by a large margin). Both of these papers examined the performance of predictions (within the dataset) using random selections of sites. This provides a measure of the performance of the models to **interpolate** across the model space. Wehrly et al. include terms to account for spatial structure and this allows Tw predictions to vary spatially (independent of the covariates); however it is unclear how well the model would perform when predicting Tw in new regions (e.g. into adjacent states). Moore et al. acknowledge that there could be some negative spatial bias in their model in the north east, indicating that the covariates may not be completely transferable or that there is additional spatial variability that should be accounted for. Indeed Figure 7 in Moore suggests that the Tw~Ta relationship is not constant across sites and, consequently, using a 'global' Tw~Ta relationship (i.e. fitting an average relationship across all sites) could introduce biases into predictions depending on the spatial distribution of the within-site response gradients. To determine the ability of the Moore model to extrapolate rather interpolate would have required the model to be fitted regionally and then the predictions tested in other regions (for example, using cross-validation with one sub-region excluded at a time). The reviewer notes that "some of the previous papers [presumably the Moore and Wehrly papers] have focused on extensive regions and thus have implicitly demonstrated that models based on landscape variables can be applied consistently across multiple catchments". This is true up to a point. However, because the models are fitted at such a large scale, it does not follow that the models will be good predictors of local (e.g. within-catchment) spatial variability in Tw. For managers operating at a more local (catchment) scale, areas e.g. at risk due to sustained high temperatures might not be well identified because the local Tw~Ta relationship is different to the global Tw~Ta relationship.
- Our paper investigated the consequences of **extrapolation** rather than **interpolation**. We investigated the consequences of transferring model predictions between catchments (i.e. to areas outside of the model domain). This is often required when widespread data are not available, for example when financial and logistical considerations have focussed data collection around a few critical rivers. The consequences of **extrapolation** have not been investigated previously as far as we are aware. In particular, we show that for our catchments, models that only use landscape covariates provide more

reliable predictions that those using a Ta metric. Further, because there is also increasing interest in the use of spatial statistical river network models (and recognition of the need to use these models where data collection is focussed on particular rivers) we think it important to quantify the relative accuracy of predictions for new sites within a river catchment for which there are already data, with those for catchments for which there are no data (and only landscape covariates / Ta can provide predictions). This has not been explored in previous papers. Finally, given the widespread use of Ta in large scale models we find it useful to highlight that Tw~Ta relationships are not universally applicable and that this can result in catchment or regional biases. Although Tague et al. suggest this could be an issue, we explicitly demonstrate some of the problems.

What was the sampling design? Were sites selected randomly within some predefined strata (e.g., based on catchment area)? This point is important, because a carefully designed sampling scheme can minimize issues with multi-collinearity and enhance model identifiability.

- The sampling design for the sites is described in detail by Jackson et al. (2016a) which is cited in the paper. In short, we tried to cover the environmental range of all the predictor covariates across space (in each of the catchments). It is a strategically designed network with careful quality control. Unfortunately it is not possible to entirely exclude collinearity given the spatial structure of the rivers involved and their physiographic / hydro-climatological context, but we did exclude strongly correlated covariates from the analysis (see methods section)

The authors note that the relation with air temperature is inconsistent. In discussing this point, they draw upon the results of studies of the temporal relation between stream and air temperature. However, it is not valid to draw inferences about spatial patterns from temporal relations. See Luce et al. (2014) for a discussion of stream thermal sensitivity. The authors should focus on relations between stream and air temperature in a spatial context. The cited paper by Fellman et al. (2014) did try to include air temperature as a spatial covariate but did not find a significant relation. However, their sample size was only 9 and thus their analysis had limited power. A number of studies have found significant relations between stream and air temperature in a spatial context (e.g., Tague et al., 2007; Wehrly et al., 2009; Moore et al., 2013).

- We agree that, ideally, we should focus on relations between stream and air temperature in a spatial context. However, whilst other studies have found significant spatial relationships between stream and air temperature (as we did in the Tweed and Dee and, admittedly implausibly, in the Bladnoch), as far as we are aware ours is the first study to investigate differences in spatial relationships between catchments or regions. If the referees are aware of studies, we would be happy to incorporate these in our revision. To explain why such differences might exist, we feel that it is valid to consider the widely reported variation in temporal Tw~Ta relationships both between sites and catchments. Clearly, the more temporal Tw~Ta relationships vary between sites, the harder it will be to establish significant spatial Tw~Ta relationships, and the less precise predictions based on such spatial relationships will be. Equally, just because there is variation in the temporal relationships between sites, it doesn't necessarily mean that a spatial relationship developed in one catchment will differ from that in another. However, if the temporal

relationships vary between catchments, for example because of differing ground water influences, then there is no reason to expect consistent spatial relationships across catchments, particularly when considering maximum summer temperatures, when differences between catchments will be accentuated. One of the papers suggested by the reviewer also uses these arguments. Tague et al. illustrate that the temporal Tw~Ta response varies depending on geological setting and state that "In this study we show that, in regions where groundwater inputs are key controls and the degree of groundwater input varies in space, air temperature alone is unlikely to explain within-landscape stream temperature patterns". We will revise our discussion so that we are clearer about the possible effects of between-site and between-catchment variation in the temporal Tw~Ta relationship.

- We are happy to add the additional Luce et al. (2014) reference which also shows considerable variability in Tw~Ta relationships (or climate sensitivity).
- We acknowledge that other papers have found a relationship between Tw and Ta in large scale spatial models. We agree that average relationships can be observed over large spatial scales. Indeed, we could have forced our multi-catchment model to have a common Tw_{max}~Ta_{max} relationship across catchments. However, instead we demonstrate that the Tw_{max}~Ta_{max} relationships are not consistent across catchments, and consequently that predictions from a model which assumes a common response would be biased.

The authors should consider more thoroughly the reasons for the "physically implausible" relation between stream and air temperature for one catchment. Presumably it reflects a confounding effect of some variable not included as a covariate. For example, Hrachowitz et al. (2010, p. 3383) found that stream temperature tended to increase with elevation, which they attributed to the fact that upper elevations were not forested. Apart from this within-catchment scale, was the among-catchment variation in stream temperature consistent with the spatial pattern of air temperature? Perhaps air temperature can be effectively used at some spatial scales and not others? This could be an interesting point to address with reference to the broader literature.

- We believe that this response is most likely due to systematic spatial variability in covariates that affect the gradient of the temporal Tw~Ta relationship between sites (e.g. geology, hydrogeology, landuse), coupled with a limited range of observed values of Ta_{max} (see response to referee 1). If the gradient of the temporal Tw~Ta relationship varies among sites in a spatially systematic way, then the spatial Tw_{max}~Ta_{max} response could be more complex as seen here. We agree that this could be because of some covariate that are not included in the model, or because of some interaction between the covariates that are included. However the exact cause of the pattern in the Bladnoch is supposition. We will improve our discussion of this issue in our revision.
- We are a little unclear about precisely what is being asked. However, the estimated catchment effects in the multi-catchment LS model are not obviously related to the average Ta_{max} in each catchment (although with only four catchments, there is little power to pick up such a relationship).

The study is fundamentally constrained by the limited sampling in both time and space. Although the authors acknowledge some implications of the small sample size, including the inability to include interaction terms, they do not fully address how the small sample size has constrained their ability to draw inferences. Two key points follow.

a. The authors do not provide sample sizes, but inspection of Figure 1 suggests about 20 to 30 per catchment. These are not large sample sizes, especially for the application of multiple regression. One guideline is that roughly 10 samples are required to support each predictor variable. Hence, the authors are fundamentally unable to incorporate potentially important predictors or interactions among predictors. Studies with greater sample sizes have been able to incorporate more predictors, leading to broader insights into landscape-level controls on stream temperature (e.g., Isaak and Hubert, 2001; Scott et al., 2002; Wehrly et al., 2009).

- The sample sizes (which are stated in line 15, page 3) are 59 Dee, 34 Tweed, 25 Spey, 19 Bladnoch. We accept that the sample sizes are modest (especially in the case of the smaller Bladnoch catchment), but we consider them adequate considering the strategic nature of the network which covers the environmental range of the covariates. Further, the sample sizes were sufficient to fit 'full' models for each catchment which included all the main explanatory variables. We agree that there is a danger of over-fitting when there are many covariates and small sample sizes; hence our use of AICc / BIC to penalise more complex models. We were not seeking to develop complex models, but rather simpler models that would be more likely to be transferable (e.g. Millidine et al., 2016). Although it was not the primary focus of the study, we note that our multi-catchment models, which were fitted to 137 sites, contained a limited number of covariates.

b. The study only covers one season, in which temperatures were low and substantial rain fell. It is therefore unclear whether the results are specific to this one period. Perhaps in a warmer, drier summer, there would be greater spatial variability and perhaps different predictor variables would dominate.

- It is true that the dataset is constrained to one year. We believe that greater spatial variability could occur in a hotter, drier year, but we also believe that the ranking of inter-site differences in temperature would remain similar (we have seen this in our longer term data). We therefore believe that in other years the LS models would be capable of predicting hotter and cooler areas of the catchment, but not of adequately characterising between site variation or absolute temperatures. For those catchments with a feasible LS_Ta model, it is possible that better predictions of absolute temperatures could be obtained, but inter-site variability could well be greater. This would need further investigation. We expect that the same covariates would remain important even in other years, but of course cannot be certain; this would again require investigation at a future date. We can discuss these limitations in the revised submission.

The authors mention the effect of continentality on stream temperature, but the causal mechanism is unclear. I could imagine that the effect arises through the effect of continentality on air temperature, yet this seems inconsistent with the findings related to air temperature. Alternatively, could it reflect variations in precipitation and thus streamflow?

- Yes, the effects of continentality could relate to both Ta and rainfall (with greater rainfall over the mountainous areas). Continentality might have a greater effect in the larger east coast catchments and indeed there was a positive relationship between Tw_{max} and Ta_{max} in the Tweed and Dee catchments.

In the conclusion, the authors suggest that further work should investigate the modelling of among-catchment variability. It might be useful for the authors to take a first run at this by examining whether the among-catchment variability is correlated with some catchment-scale measure of air temperature (or some other relevant variable).

- This is an interesting proposition for future work when we are able to bring in more catchments (and indeed this work is underway) but seems of limited value here where we have considered just four catchments.

In Tables 2 and 4, the authors should include the standard error of estimate or the root-mean-square error from validation. It would be interesting to see a comparison of the precision of their models with that found in previous studies using similar temperature metrics.

- The RMSE, standard deviation and bias of the models in Table 2 are presented in Table 3. E.g. for the Dee you would look at the Dee donor catchment, applied to the Dee.
- We do not include performance metrics for the models in Table 4 because they are not transferable (they include catchment effects and interactions that cannot be transferred more widely). We only report these models to illustrate that the responses differ between catchments.
- We are happy to extend the discussion to compare the performance of our catchment specific models with the models developed in other studies. However, we note that there are very few studies available that model T_{max} (many models focus on mean temperatures that appear to be more spatially predictable) and not all of these provide comparable performance metrics.

The authors should provide some more information about the RNS model, which is not as commonly applied as network models based on spatial covariance functions. Are there limitations related to sample size? For example, for network models based on spatial covariance functions, a general guideline is that one needs at least 50 samples.

- A detailed description of the RNS is provided in O'Donnell et al., (2014), with application to temperature models described by Jackson et al., (2016b). We do not want to repeat these papers here. The limitations on sample size come from the combination of the RNS and the covariates. Previous exploratory work suggested that the RNS required up to 6 d.f. for models of Tw_{max} in the river Spey

(Jackson et al., 2016b). In this paper we therefore allowed up to 7 degrees of freedom; the Bladnoch did not contain a RNS, and the Tweed, Spey and Dee used 3.91, 4.2 and 6.8 d.f. respectively in the LS models.

References:

Hannah, D. M., Malcolm, I. A., Soulsby, C. and Youngson, A. F.: A comparison of forest and moorland stream microclimate, heat exchanges and thermal dynamics, Hydrol. Process., 22(7), 919–940, doi:10.1002/hyp, 2008.

Isaak, D. and Hubert, W.: A hypothesis about factors that affect maximum summer stream temperature across montane landscapes, J. Am. Water Resour. Assoc., 37(2), 351–366, 2001.

Jackson, F. L., Malcolm, I. A. and Hannah, D. M.: A novel approach for designing large-scale river temperature monitoring networks, Hydrol. Res., 47(3), 569–590, doi:10.2166/nh.2015.106, 2016a.

Jackson, F. L., Hannah, D. M., Fryer, R. J., Millar, C. P. and Malcolm, I. A.: Development of spatial regression models for predicting summer river temperatures from landscape characteristics: implications for land and fisheries management, Hydrol. Process., doi:10.1002/hyp.11087, 2016b.

Johnson, S. L.: Stream temperature: scaling of observations and issues for modelling, Hydrol. Process., 17(2), 497–499, doi:10.1002/hyp.5091, 2003.

Li, H., Deng, X., Kim, D. and Smith, E. P.: Modelling Maximum Daily Temperature Using a Varying Coefficient Regression Model, Water Resour. Res., 2014.

Luce, C., Staab, B., Kramer, M., Wenger, S., Isaak, D. and McConnell, C.: Sensitivity of Summer Stream Temperatures to Climate Variability in the Pacific Northwest, Water Resour. Res., 50(4), 3428–3443, doi:10.1002/2013/WR014329, 2014.

Millidine, K. J., Malcolm, I. A. and Fryer, R. J.: Assessing the transferability of hydraulic habitat models for juvenile Atlantic salmon, Ecol. Indic., 69, 434–445, doi:10.1016/j.ecolind.2016.05.012, 2016.

Moore, R. D., Nelitz, M. and Parkinson, E.: Empirical modelling of maximum weekly average stream temperature in British Columbia, Canada, to support assessment of fish habitat suitability, Can. Water Resour. J., 38(2), 135–147, doi:10.1080/07011784.2013.794992, 2013.

O'Donnell, D., Rushworth, A., Bowman, A. W., Scott, M. E. and Hallard, M.: Flexible regression models over river networks, J. R. Stat. Soc. Ser. C (Applied Stat., 63(1), 47–63, doi:10.1111/rssc.12024, 2014.

Scott, M. C., Helfman, G. S., McTammany, M. E., Benfield, E. F. and Bolstad, P. V: Multiscale influences on physical and chemical stream conditions across blue ridge landscapes, J. Am. Water Resour. Assoc., 38(5), 1379–1392, doi:10.1111/j.1752-1688.2002.tb04353.x, 2002.

Tague, C., Farrell, M., Grant, G., Lewis, S. and Rey, S.: Hydrogeologic controls on summer stream temperatures in the McKenzie River basin, Oregon, Hydrol. Process., 21, 3288–3300, doi:10.1002/hyp, 2007.

Wehrly, K. E., Brenden, T. O. and Wang, L.: A Comparison of Statistical Approaches for Predicting Stream Temperatures Across Heterogeneous Landscapes, JAWRA J. Am. Water Resour. Assoc., 45(4), 986–997, doi:10.1111/j.1752-1688.2009.00341.x, 2009.