Review of "Regional, multi-decadal analysis reveals that stream temperature increases faster than air temperature"

Dear Authors, dear Editor,

The submitted paper presents a large-scale modelling effort of river discharge and temperature for the Loire River basin in France. The modelling is carried out in two steps, first for discharge and then for water temperature. The models are forced with reanalysis data provided by Météo-France and validated on a large number of measurement stations. The model results are then used to analyse and discuss the trends in discharge and water temperature, the factors influencing the modelled trends in water temperature, the spatial patterns of the resulting trends, and the role of riparian vegetation and Strahler order.

This article presents an important work and a valuable contribution to his field of research. In general, the paper is well written in terms of language and has clear and nice figures. However, I have some concerns which I detail below. The first concern is the clarity, completeness, and organisation of section 3 "Method and data", where, in my opinion, some crucial information is missing. The second concern is the robustness of some of the results, in particular the validation of the model and the strength of some of the analysis performed. Indeed, the results and discussion sections tend to analyse many different aspects and perhaps only the most robust ones could be kept. In addition, since quite simple models are used, I would be more careful for some interpretation of the results.

I also regret that no data or models are shared along the article. I imagine that sharing of forcing data may not be allowed, however I would strongly encourage the addition of a "Data and code availability" section to detail how this work can be replicated. If possible, sharing with the community the time series produced by the model would also be appreciated. In the same vein, the supplementary material could be improved to provide interested readers with a more complete overview of the large number of results obtained.

Despite these concerns, this paper has the potential to bring interesting new results. I think that a revised version would be suitable for publication in HESS. Water temperature is now a rapidly expanding area of research, and with climate change underway, further research efforts are needed. This article is certainly a valuable contribution.

Please do not hesitate to contact me for further discussions.

Best regards,

Adrien Michel (adrien.michel@epfl.ch)

Main comments:

1. Details of methods and data

1.1 Data description

My first recommendation is to create a "Data" section (with sub-sections) where all details on forcing, validation and geographical data are grouped. Currently, the information is spread over P3L83-86, P4L97-P5L104, P5L118-121, L123-124, L125-128, P7L165-167 (I may have forgotten some). Grouping them together would make for easier reading as many details now appear in the models' description sections.

In addition, a comprehensive table of all stations used for calibration and validation should be provided in the supplementary material (SM later), with coordinates and station names, and also indicating the data provider. The coordinates should be added in Table 1. For long-term stations, one could even consider numbering them in the table and then indicating the numbers on Figure 1. Maps similar to Figure 3 and S5, but showing the annual and seasonal average of Tw, Ta and Q over the whole catchment area could be a useful addition to the SM to capture the different local conditions in the catchment. A map similar to Figure 2, middle panel, with a colour indicating the Strahler order could be added in the SM. For non-French readers, a map showing the location of the Loire basin in France (or in Europe) should be added. It could be integrated as a new panel in figure 1.

In P5L103 it is said that the Q time series are "naturalised". However, no reference or details of the procedure used are provided. This information should be included because, as you note later in section 5.4, anthropogenic disturbances are of major importance. In addition, I did not find in the document the source of the time series for water temperature. Is it the same supplier as for the discharge? Is it also "naturalized"?

1.2 Description of the EROS model

Only a few details are given on how the model actually works. In addition, the two main references given for the model are in French. Although it is not necessary here to describe all the details of the models, the main points should be provided to the reader. In particular, details of the mass balance should be given. Is it precipitation - evaporation, or can some of the mass be lost through deep soil infiltration? How is the water transported through the soil? Does EROS use a reservoir model? Finally, is there any routing of water into the stream network carried out in EROS or is only the release of water at the sub-catchment scale simulated? P5L109-111 suggest that routing is done in T-NET, but this is not clearly stated, nor is it mentioned in section 3.1.2. On the other hand, section 3.2.1 suggests that Q is obtained directly from EROS, which means that the routing is done in EROS.

It is not mentioned which parameters are calibrated (this should be stated), which values are tested for calibration, and which values are finally chosen. This should be stated in the SM to allow reproducibility. P5L105-107 states that: *"The calibration aimed at maximizing the Nash-Sutcliffe efficiency criteria (Nash and Sutcliffe, 1970) on the square root of streamflow and minimizing the overall bias, in order to simulate correctly the whole range of Q values".* How are the two metrics combined (NSE and bias) to assess the quality of the calibration? Why is the square root used? How many calibrations are performed, how are the calibration values chosen (pure chance or more advanced algorithm)? Again, the only non-French peer-reviewed source provided is Thiéry (1988). This paper describes a water level model for an unconfined aquifer and I imagine that EROS is a model based on this work, but rather different from this 1988 version.

There is also no mention of the other input parameters of the model. The authors describe three different HydroEco regions (HER), but there is no information on how the region influences the model parameters (and thus the model outputs).

Overall, more details about the model are needed to enable the reader to understand how the EORS model works. The main assumptions and principle of the model should be stated, as well as details on the calibration procedure and parameters (a part can be added to the SM). I have found many applications of this model in the literature; these could also be cited as application examples.

1.3 Description of the T-NET model

Here too, important pieces of information are missing. This is even more problematic than for the EROS model as no references are given for this model. P5L115-116 says: "*To simulate Tw, the equilibrium temperature (Te) is first computed, the temperature at which the net heat flux across the surfaces of the stream is null*". Then details of how some of the energy fluxes are calculated are given, but the above sentence is never followed by a "then" or "next". So, if I understand correctly, the temperature of the water being modelled is the equilibrium temperature? Or are you using a formulation similar to Bustillo's (2014) eq. (6). A crucial piece of information is missing here. And if only Te is used, this critical assumption and its consequences need to be discussed.

In section 3.1.1 it says that EROS is used to calculate discharge over 368 sub-catchments, whereas in section 3.1.2 52'278 reaches are mentioned. So, there are several reaches per sub-catchment and within each sub-catchment, the water supply simulated by EROS is distributed to the reaches using the drainage area? But again, how and in which model is the routing calculated? P6L156-157 clearly states that "*Q is the daily mean streamflow provided by the EROS model*". But how can this be done if there is no routing in EROS? P7L160-161 explains that the travel time is calculated in T-NET, but this information is not used for the calculation of the water temperature (at least that is my understanding), but would be mandatory information for routing.

No details are provided on how sensible, latent and groundwater heat fluxes are calculated. Equations or references should be provided. I understand that no calibration is done for T-NET, is this correct? Are there any other parameters used in the model? For example, as with EROS, how are the properties of HERs taken into account in the model? The results are discussed for the different HERs, but this discussion only makes sense if the HERs are somehow parameterised in the model.

Finally, in addition to my questions about routing, there is no information about reach-to-reach heat advection. Is there reach-to-reach heat transfer? If not, this would considerably weaken the analysis carried out in terms of the Strahler order and of the whole model in general.

To conclude this first part of the discussion, I would really encourage the authors to add a few paragraphs better detailing the data sources, the model workflow, and the main assumptions of the models. As I say below, it would also be important to indicate the limitations of the models.

2. Model validation and robustness of the results

2.1 Calibration and validation of EROS and T-NET

Firstly, details of the calibration procedure and parameters should be provided (see above). Secondly, why is no validation period used to infer the quality of the calibration? Indeed, time series are usually divided into a calibration period and a validation period. By using only a calibration period without validation, we have no information on the potential overfitting of the model at the calibration station during the calibration period. Depending on the modelling effort required, I would strongly recommend recalibrating the model over a shorter period and using a few years for validation. Furthermore, looking at Figure 3 of Thiéry (1988), we see a clear decrease in the performance of the model during the validation period.

In section 3.2.1, NSE on Q, ln(Q) and sqrt(Q) are mentioned, does this correspond to the "the mean NSE criteria for low, medium and high flows" mentioned in P9L221? Little detail is given on the quality of the calibration of Q. Indeed, only the mean NSE is given (no details on the variance), a graph of the distribution of values should be added in SM, as well as a graph showing the simulated and measured time series for some stations. In Figure S2, the bias is only shown for the 44 NHN stations, why not show all 352 stations (using a shorter time period)?

In Section 4.1, it is shown that most of the simulated summer trends are not significant and that the simulated summer trends are poorly correlated with the observed trends. However, these summer trends are used extensively later in the document (e.g. Figure 5, Table 2). Given the uncertainty around these trends over the calibration period, I doubt that they are robust enough to perform such an analysis.

For Tw, only the performance in terms of trends and biases is presented. A presentation of performance in terms of mean square error would also be informative in assessing the performance of the model as this metric is commonly used in the literature. The bias is discussed for Q and Tw. However, as the subsequent analysis is mainly trend based, and bias has no impact on trend, I am not sure of the relevance of this metric here.

Section 5.1 does not add much to section 4.1. I would like to see in these sections more discussion of where the models are underperforming and therefore the limitation of the dataset obtained. Indeed, the models used, like all models, are not perfect. Identifying the limitations and thus focusing the analysis and discussion on the part that proved to be within the radius of validity of the models is a mandatory step in modelling studies.

Finally, the part concerning riparian vegetation seems to be a new addition to the model in this article. My concern is that no validation is shown. The approach should be validated using a Tw measurement station in a shaded area and compare the model performance with and without shading. In the absence of validation and evaluation of the effectiveness and limitations of the approach, I find it difficult to proceed with the analysis.

2.2 Link between Ta, Tw, and Q

The comparison of Ta and Tw trends and the potential impact of Q are widely discussed. I appreciate this effort and think that there is still much to understand about the interaction between Tw and Q.

First of all, great importance is attached to the finding that Tw increases faster than Ta for most seasons and on an annual basis. This result has already been found in some regions (see e.g., Webb and Nobilis, 2007; and Arora et al., 2016), but is in contradiction with other studies (see e.g., Moatar and Gailhard, 2006; Orr et al., 2015; and Michel et al., 2020). I believe this result would merit further discussion to assess its strength.

The main factor used to explain why the trends in Tw are more important than those in Ta is the discharge. However, trends in other forcing variables should be shown. Indeed, in winter, Figures 3 and S5 show that the variable explaining why TW trends > TA is probably not discharge (since no significant trend in discharge is found). Similarly, summer and spring show marked negative trends in discharge in some parts of the catchment, whereas these are the seasons where the trends in Ta and Tw are most similar. The addition of the seasonal discharge trend in the boxplot in Figure 4 would facilitate the analysis.

Figure 5 is used to support the hypothesis that Q is the main driver, see P11L259-263: "*Overall, Tw trends were more spatially variable than Ta trends, suggesting the conditional influence of Q trends (Fig. 4). Indeed, where Tw trends exceeded Ta trends, decreasing Q trends occurred coincidentally at the majority of reaches for all seasons – with the exception of winter – (43-72 %, depending on season; Fig. 5. Of these specific reaches where all factors converged (trend in Tw higher than trend in Ta, and decreasing trend in Q)*" (please note that there are some grammatical problems in the second sentences). There are many shortcuts here. Firstly, the larger scatter suggests that factors other than Ta have an impact on water temperature, but this does not in itself show that Q is responsible, it could be any other forcing variable. Secondly, and more importantly, Figure 5 shows the percentage of reaches where "all factors converge", which I interpret as "all trends are significant". Please clarify this point, as when comparing Figures 5 and 3, I rather understand that all trends are included in Figure 5. Including non-significant trends would be a significant bias here, as many non-significant trends are just above or just below zero and the figures are based on trend signs.

For Figure 5 to be complete, the distribution between $Q>0$ and $Q<0$ should also appear in the blue part. Indeed, the figure now shows that $Tw>Ta$ in most cases when $Q<0$, however Figure 4 shows that $Tw>Ta$ in most cases anyway. This figure could be used to show the impact of Q if, and only if, we can see that the proportion of $Ta>Tw$ vs $Ta<Tw$ changes if $Q>0$ and $Q<0$. Also, the number of catchments used probably differs significantly between seasons (at least if only significant trends are shown). Here I see that there may be something irreconcilable in this figure:

- Either all reaches are kept, including those with insignificant trends, but then the figure itself will lose much of its meaning for the reason mentioned above.

- Either only significant trends are retained, but only a small minority of reaches show significant Q trends in certain seasons, and the figure would then only show a small subset of reaches.

This question of the relationship between Tw and Q trends is not straightforward (you can look at the introduction of Arora et al. (2016) for a good review of the literature available by then) and, it is important to note, seeing a correlation between Tw trends and Q trends does not imply any causality. This is what we see in the subsection "Synchrony of annual anomalies" and in Figure 8: low Q summers correspond to high Ta summers, so it is difficult to assess which of the two, or the combination, leads to an increase in Tw. Furthermore, Figures S10 and S11 suggest that the negative discharge trend is caused by an increase in ET rather than a decrease in P. The regions concerned (see Figure 3), are those where the increase in Ta is most significant. Thus, the causal chain here appears to be Ta increasing $\rightarrow$ ET increasing $\rightarrow$ Q decreasing. So, even if Q is shown to be a factor influencing Tw, it originates (mostly) in the increase in Ta. Thus, it might be misleading to say that a decrease in Q is a contributing factor, as I think that for most readers a decrease in Q would mean a decrease in precipitation, whereas here no significant decrease in precipitation is found, and thus no impact of the precipitation regime on Tw can be assessed. The real impact of ET can be assessed by comparing the measured trends of discharge to the trends on measured precipitation in the catchment, in order to confirm if the increase in ET modelled is correct.

With all this discussion, I really question this sentence in the abstract: *"Importantly, air temperature and streamflow exerted joint influence on stream temperature trends, where the greatest stream temperature increases were accompanied by similar trends in air temperature (up to +0.71 °C/decade) and the greatest decreases in streamflow (up to -16 %/decade)".* Indeed, as discussed above the discharge decrease might just be a "side-effect" of the increase in Ta through increased ET, but since it is also the region with the highest Ta increase

(P16L229), I don't think that we can conclude that "*air temperature and streamflow exerted joint influence on stream temperature"*.

Continuing with the subsection "Synchrony of annual anomalies", I do not really see the added value of the change point analysis. Furthermore, as it seems to illustrate an abrupt change in the late 1980s, I would think that a trend analysis is then not appropriate for these time series.

Going back to the Tw>Ta question more generally (sorry, all the variables are related and I had trouble organising my comments), it would be interesting to calculate trends at long-term water temperature stations to see if the same trend is observed, to reinforce the results. In addition, a map like the one in Figure 3, but showing the difference between Ta and Tw, would be very informative about the spatial distribution of catchments where Ta trends are more important than Tw and help the comparison with discharge trends.

This result receives a lot of attention in the article (see for example the title). The details (Figure 4) show that this is mainly due to the winter and autumn seasons (although the difference is statistically significant in spring, it is still really small). However, Figure S4 and the indicated R coefficient shows that the simulated trends in autumn have the lowest correlation with the measured trends (this is also stated in P9L232), so autumn is the season where we have the lowest confidence in the results. In any case, I would condition the general statement of Tw>Ta on the seasonal aspect since it is not general (and I would add this information in the title) and I would really stress the uncertainties about it. This result is interesting, and certainly deserves attention, but in my opinion not robust enough to be asserted in the way it is in the title of the paper (at least with what is currently shown). The discussion around the cause can also be enhanced.

2.3 HER, Strahler order and riparian vegetation

The study is complemented by an assessment of the influence of HERs, Strahler order and riparian vegetation on Tw. I have already raised some concerns about HERs (how they are accounted for in the models), and about riparian vegetation (the model is not validated for this). These two points need to be addressed in order to present the analysis. In general, all these topics are interesting, but I have the impression that they are treated only superficially and not with the necessary rigour. Moreover, there are already a lot of results presented and I think the article would still be interesting, and perhaps easier to read, if these parts were removed.

Regarding the Strahler order analysis, this brings back to the issue of the reach-to-reach heat advection mechanism that is missing in the model description. Also, showing a Strahler order map for each reach in the SM would be really informative. As mentioned in in P22L397-398, the correlation between Strahler order and Tw may be due to riparian vegetation (and not a concentration of the warming when going downstream). However, this should be analysed in more details, e.g., by looking at the correlation between Strahler order and Tw separately for different shading factor values (e.g., the categories of Figure 10).

For riparian vegetation, do you have a mechanism for why it would change the trend in Tw? I understand that adding or removing vegetation would change the absolute value of temperature, but by what mechanism would the trends be affected? If the riparian temperature sections were to be retained, this should be addressed.

To conclude this second part, I would recommend that the authors revise and strengthen the analysis of Ta vs Tw trends and the impact of Q. With the results presented, such strong statements as those in the title and abstract do not entirely hold water. I would also recommend mentioning in the title the location where the study was conducted (Loire, France). Perhaps some of my reservations stem from a lack of understanding of exactly how the models work and a better description might help to alleviate these reservations. Along with the emphasis on the main message that I recommend, perhaps some of the "secondary analyses" could be removed from the document. I think that a more thorough discussion of the limitations of the methods and results should be provided. I have not commented in detail on all the discussion, summary and conclusion sections, but certainly some of them are relevant to my comments above.

A final comment concerns the real added value of such a major modelling effort. A significant part of France is modelled; however, I feel that some of the results could be obtained by analysing only past measurements (and getting rid of all the modelling uncertainty). The added value could come from the whole analysis of the riparian vegetation for example, but as mentioned these analyses need to be strengthened. One solution (but which would involve a lot of extra work), could be to first publish just the data set in a journal like ESSE (https://www.earth-system-science-data.net/), where all the modelling and validation aspects are discussed in detail, and then have an article in HESS focused on the data analysis only allowing for an in-depth analysis. This would also allow for elaboration on some aspects that are not yet discussed (*e.g.*, elevation, impact of snow if relevant for this catchment, more detailed spatial analysis). I think this option would really increase the potential impact of the significant modelling effort that has been made. But I would understand if the authoring team do not want to go through this extra work.

Minor comments:

P1L38 [Q] should be (Q)

P1L52-52 "*in the face of a changing climate*", maybe just say "to climate change"

P3L79-80 Maybe define exactly what HERs are, or give a reference

Figure 1 Add HERs "borders" also in left panel. In the manuscript, this figure is not vectorized and small text are really pixelized when zooming to reads them. Maybe provide a vectorized figure or a higher resolution bitmap figure. Add in the figure or caption the source for the maps shown

P5L101 "*bottom*" should be "right"

P5EQ(5) Can't it be written in the more compact form $max(SF_{Left},SF_{Right})$?

P8L200 Why using *log(Q)* in this analysis?

P9L224 What does IQR mean?

P9L226 "r" is used here, while "R" is used in figures

Figure 2 There is a "a" on the right below the colour legend

P11L267-269 Please add a reference here to support this statement

P16L306-308 How negative Q trends, just by themselves, suggest an effect on Tw? "suggested" should be "suggesting".

P20L335-338 First, despite it is clearly stated on the abstract of the paper Michel (2020) that: "*The mean trends for the last 20 years are + 0.37 ± 0.11 ◦ C per decade for water temperature, resulting from the joint effects of trends in air temperature (+0.39 ± 0.14 ◦C per decade), discharge (−10.1 ± 4.6 % per decade), and precipitation (−9.3±3.4% per decade)"*, I think now that this paper does not show a real impact of Q on Tw, but rather a correlation between Q and Tw in summer. If I had to rewrite this paper today, I would not be so categorical (and this why I also question it in your paper). Second, when you say: "In *contrast with our results, they found Tw trends lower than Ta trends due to influence of snow melt and glacier melt",* this is not totally exact. Indeed, trends found in Alpine catchments are lower due to the mentioned effects. For the low-altitude catchments where snow plays no role, trends are indeed closer, but Tw trends remain slightly slower than Ta trends (compare Figures S17 and S18). However, on an annual basis, we are talking about a few tenths of a degree less in my article and a few tens of a degree more in yours, so taking into account all the uncertainties involved, I see no contradiction. In addition, different regions are studied.

Introduction and Section 5.5 Nuclear plants cooling is never mentioned in the paper. This might not be relevant for the Loire (but Bustillo et al. (2014) mention some plants in the catchment), but in general in France the question of cooling nuclear plants in the future with increasing air and water temperature will be a real challenge and I think it is worth mentioning it (see e.g. Bourqui et al., 2011).

P22L398 Shouldn't it be "small rivers"?

References

Arora, R., Tockner, K., and Venohr, M.: Changing River temperatures in northern Germany: trends and drivers of change, Hydrol. Process., 30, 3084–3096, DOI: 10.1002/hyp.10849, 2016.

Bourqui, M., Hendrickx, F., and Le Moine, N.: Long-term forecasting of flow and water temperature for cooling systems: Case study of the Rhone River, France, AHS Publ, 348, 135–142, 2011.

Bustillo, V., Moatar, F., Ducharne, A., Thiéry, D. and Poirel, A.: A multimodel comparison for assessing water temperatures under changing climate conditions via the equilibrium temperature concept: case study of the Middle Loire River, France. Hydrol. Process., 28: 1507-1524. DOI: 10.1002/hyp.9683, 2014.

Michel, A., Brauchli, T., Lehning, M., Schaefli, B., and Huwald, H.: Stream temperature and discharge evolution in Switzerland over the last 50 years: annual and seasonal behaviour., Hydrology & Earth System Sciences, 24, DOI: 10.5194/hess-24-115-2020, 2020.

Moatar, F. and Gailhard, J.: Water temperature behaviour in the River Loire since 1976 and 1881, Comptes Rendus Geoscience, 338, 319–328, DOI: 10.1016/j.crte.2006.02.011, 2006.

Orr, H. G., Simpson, G. L., Clers, S., Watts, G., Hughes, M., Han- naford, J., Dunbar, M. J., Laizé, C. L. R., Wilby, R. L., Bat- tarbee, R. W., and Evans, R.: Detecting changing river temperatures in England and Wales, Hydrol. Process., 29, 752–766, DOI: 10.1002/hyp.10181, 2015.

Thiéry, D.: Forecast of changes in piezometric levels by a lumped hydrological model, Journal of Hydrology, 97, 129–148, DOI: 10.1016/0022-1694(88)90070-4, 1988.

Webb, B. W. and Nobilis, F.: Long-term changes in river tempera- ture and the influence of climatic and hydrological factors, Hy- drolog. Sci. J., 52, 74–85, DOI: 10.1623/hysj.52.1.74, 2007.