In this manuscript the authors use a one way coupled drainage area to lake model setup to investigate future climate impact on spring processes including ice-off, 50% cumulative spring discharge, spring phytoplankton bloom and stratification onset. The bold and novel model setup include stream flow, nutrients and temperature (SWAT+, LOADEST, air2water) coupled to lake physics and biogeochemistry (GOTM, WET). The important findings of the authors show how the occurrence of important spring processes are occurring earlier in a future warmer climate. The manuscript is in a good order but would benefit from extra clarity, sliming down and expansion as my points hereunder show.

We thank the reviewer for the thorough assessment and the requested clarifications, which allowed us to improve the manuscript. We respond to each comment below, with our responses in text boxes and proposed additions to the text in red.

3A This manuscript continue and analyze deeper the effect of climate from the work done in Jiménez-Navarro et al. (2023). The reader needs to clearly understand what is the difference between the two works, both in regard to which questions are being addressed here as well as be given all relevant information for spring processes. This point runs throughout the rest of this review.

We thank the reviewer for this comment, and realised that there was indeed no statement at the start of the model framework description that stated this study as a continuation of Jiménez-Navarro et al. This will now be added. The previous paper described the setup of the models, the overall model performance, and the overall results of future climate projections. The present paper used the same model setup and simulations to look at spring events and how their timing might change under future climate conditions. We will add a line to clarify this.

L. 77: The present study builds upon a coupled catchment-lake model setup created by Jiménez-Navarro et al. (2023). This model setup was used to simulate catchment discharges, nutrient loads, and in-lake conditions under present and future conditions, and in the present study, we additionally assessed simulations of spring events.

We are aware that it can be challenging for readers to have information on the model in two separate places, and we strived to present (and if necessary, repeat) all necessary information for spring events in the current paper. The setup in two separate manuscripts was based on practical reasons, as we felt that a single paper with model descriptions, model performance statistics, future predictions, and additional analysis of spring events, would become too long and cluttered. We hope that our changes, to this comment and those below, will clarify the differences between the two studies.

3B The method description need to be expanded and put in line with Jiménez-Navarro et al. (2023). Among other things I cannot see how many parameters was used in air2stream, which is not a statistical model, it is a semi deterministic model (a hybrid process-based and data-driven model). Additionally, more detailed information regarding the GOTAM-WET model coupling is required. One of the things I miss is how transparency in the lake is modeled/treated. Do the biological model adjust lake transparency, and how do this affect spring bloom and stratification onset? And how do

the coupled model preform at deeper depth? The reader can now only see what happens at 3 m depth.

We used air2stream with 8 parameters. This information will be added.

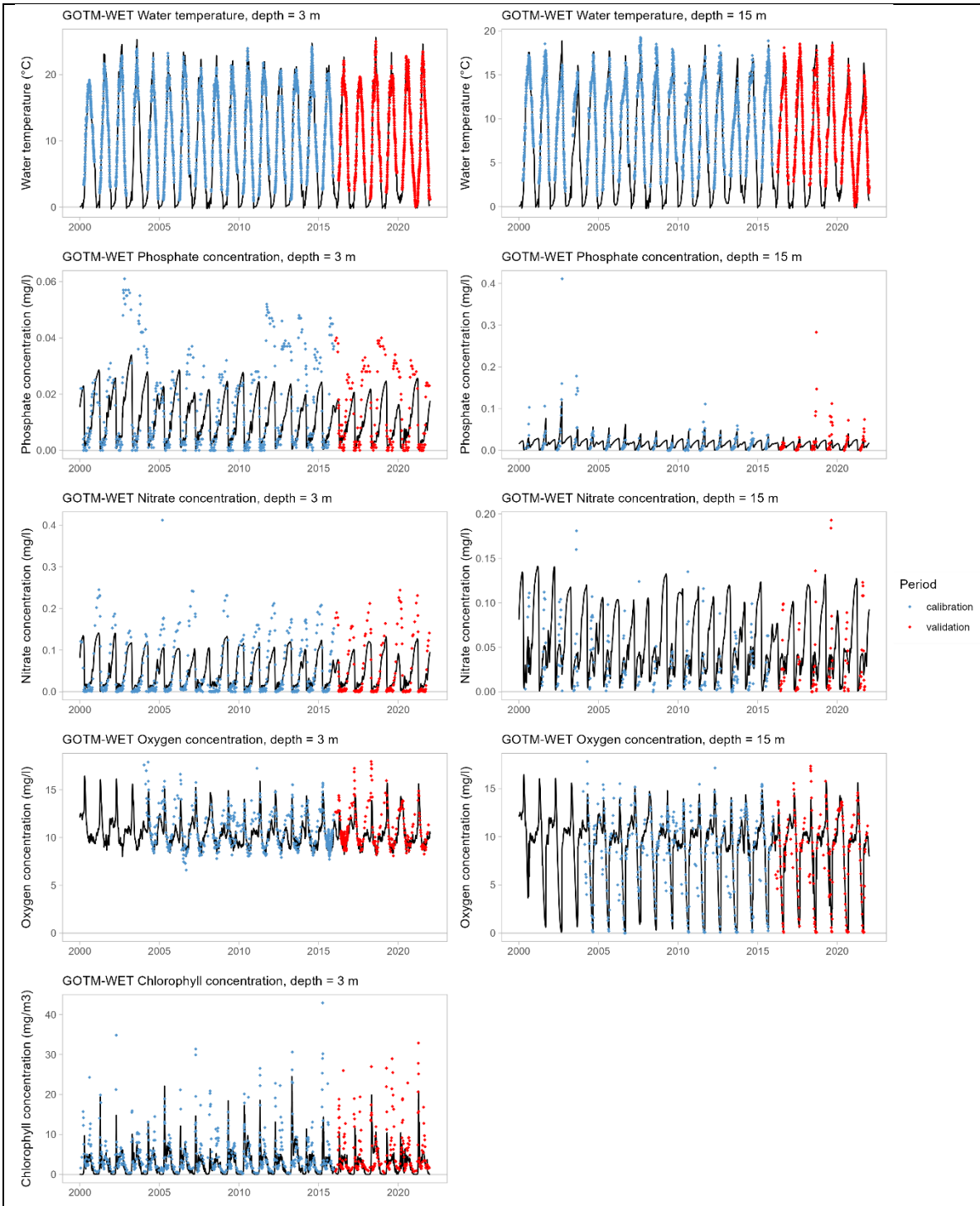
L. 81: … and air2stream (8-parameter version, Toffolon and Piccolroaz, 2015; Piccolroaz et al., 2018),…

Yes, model components from the WET model (inorganic matter, particulate organic matter, and phytoplankton biomass) contribute to the turbidity. We will add this information. We did not run tests with and without biological feedback regarding transparency. However, the final value for the transparency parameter "g2" was very high (5.62 m, giving an extinction coefficient of only 0.18 m^{-1}, this information can be found in the Supplement of Jiménez-Navarro et al., 2023), therefore a very clear water column without considering the WET components. In the final model, the biological components therefore contributed a lot to the turbidity of the water column (Secchi depth varied between 6 m and 1 m), which is in line with observations in Erken (clear water – Secchi depth > 4 m - outside of the growing season, more turbid during the spring and summer blooms, when Secchi depths of 1.5 – 3 m are common).

L. 85: Light absorption by components in the WET model (inorganic matter, particulate organic matter, and phytoplankton biomass) feeds back to the physical model.

Performance at deeper depth is also reported in the Supplementary Material of Jiménez-Navarro et al., (2023). We will add plots of the simulated model variables at 15 m depth to the Supplement (section S1, Figure S1). However, we decided to not give further information about performance at deeper depth in the text. Despite its importance for lake dynamics, our focus is on spring events, when the lake is ice-covered, fully mixed, or starting to stratify. In this period, profiles are mostly homogeneous, with the only clear exception being the onset of stratification, and this is the latest event studied here. As such, we felt that reporting on the performance at deeper depth in the main text would not be in line with the focus of the present study.

New Figure S1:



Caption Figure S1. Time series of modelled (black line) and observed (blue dots for calibration, red dots for validation period) GOTM-WET variables.

3C the authors struggle with model correctness, needing to use a surface temperature threshold for ice-off despite having an ice module and need to explain discrepancies in stratification onset and chlorophyll spring peak. I ask myself how this can be and have some points here which might enlighten the manuscript. First do the grid resolution compared to measurement resolution affect the results? From Figure S5 timing of stratification it looks like the vertical lines denoting stratification onset do not match the

data and should in fact be earlier for the measurements (red line crossing green threshold before timing of stratification). Is this due to a too short window for continuing stratification, is there an error in the script, or do the resolution play a role? As for data. The one way coupled catchment and lake model setup was calibrated from 2000 to 2015 for the lake part and from 2007 to 2015 for the river part. Is the difference in calibration period affecting the results? Looking at Figure 2 for Ice-off this looks to be the case. And how do you deal with the 2000 to 2006 period in regard to river input into the lake model? Building on this, can the less than ideal model correctness be explained by the location of measurements in and above the lake? Lake measurements come from a station at the deepest point in the lake ca 400 m from the eastern shore. This distance might be far enough away for near-shore processes to play a role, but are the location representative for the overall lake physics covering the central parts of the lake? Additionally but not required for this manuscript, it would have improved the results if the complete time frame of available data was considered for calibration, validation (if deemed necessary) could have been carried out in the start and not the end of available measurements see ex. Shen, H., Tolson, B. A. & Mai, J. Time to Update the Split-Sample Approach in Hydrological Model Calibration. Water Resour. Res. 58, (2022). https://doi.org/10.1029/2021WR031523.

We will treat the points raised by the reviewer separately.

We attribute the issues with the ice module to the lack of snow parameterisation in GOTM (L. 108-111). Although onset of ice is not affected by this, the offset (as predicted by the ice module) is likely to occur too early due to the lack of insulation that the snow provides. This is unfortunate, and an ice module including snow (such as Simstrat's) would have been better, but the choice for GOTM enabled the use of the WET model and its elaborate description of biogeochemical processes. We used the temperature-threshold to compensate for this issue, although it is not optimal (as explained further below as well).

The discrepancies in stratification onset had indeed to do with the time windows and density thresholds. Although the model simulates bottom-top density difference accurately most of the years (as seen in Supplement figure S5), the observed data is noisier than the modelled data, and the threshold approach occasionally defined different periods as the onset. We tested multiple thresholds (both regarding the time window and density difference), but there were some consistent mismatches regardless of the choice of threshold value (despite such methods being well-established in literature). Still, our method is rather well-established in literature and these issues are unlikely to have an effect for the climate projections, as the degree of noise in the signal will be constant in the model. Also in reply to Reviewer 2's comment 2B, we will clarify this further in the text.

L. 157: As such, we concluded that it was noise in water temperature observations that caused the threshold method to occasionally fail, rather than an inability of the model to simulate the state of the lake.

The shorter calibration period for the discharge was due to lack of measured discharge data before 2007. We do not understand the comment about Figure 2 in connection to the shorter calibration period for discharge, as the discharge did not seem to have a worse fit during the validation period. Regarding ice-off, we do indeed see a worse performance in 2020 and 2021. Rather than attributing this to the validation period, we expect this had to do with the exceptionally (though in coming

decades perhaps normal) short period of ice cover in those years. The temperature-threshold approach seemed to perform less well in such years. We will add this information to the manuscript. Moreover, since years with short or no ice cover will become more frequent, we will also add a line to the Discussion how this may impact our future projections.

L. 155: Particularly, this occurred in years with short ice cover duration, in which the 2 °C threshold may estimate ice-off to occur too late.

Add after L. 205: The method to estimate ice-off from the model results (a 2 °C threshold) tended to simulate ice-off too late in years with low ice cover. Therefore, our study is likely underestimating the advancement rate of ice-off date, and ice may be disappearing even faster than the rates predicted here.

Change to Table S1: indicate as annotation for years 2008, 2014, 2020, and 2021, for ice-off, that these years had the lowest recorded ice duration in the study period.

Although the shoreline is not too far away from the monitoring location, there are no major inflows anywhere near, as the largest part of the watershed is to the west of the lake. The measurement location is at the edge of the main basin of the lake, and for example seiche movements are occasionally visible in the high-frequency data (though they have a frequency around one day, so disappear with daily averaging). Regarding the processes under study here, we do not foresee a major effect of the location where the measurements were chosen, though we acknowledge that measurements from a single location are only moderately representative of the whole lake. Ice cover is likely longer in secluded bays compared to the main basin, stratification can form in shallow areas first (thermal bars), and blooms may occur in bays while the main basin is less affected, but overall, we expect our measurements to be moderately representative, due to the open connection to the main basin.

We appreciate the reviewer's comment regarding using the full period for calibration, with potential validation at the start. Although the reviewer did not request particular changes to be made to the manuscript, we will take the opportunity to elaborate on this issue, perhaps for no other reason than that we find it interesting as well! We wanted to use the same model setup as Jimenez-Navarro et al. for reasons of clarity, but in general, this is an interesting proposition and it could be considered whether the model would have been more accurate if we had considered all data for calibration. In our view, the degree to which models have been established and are prone to overfitting, plays a large role here. Hydrological and hydrodynamical models are based on purely physical equations that have been widely applied (even if the models themselves have not) and are usually not heavily calibrated, so that overfitting, or other issues related to model stability, are less of a big risk. In biogeochemical models, however - at least the rather complex type that we used here -, many parameters are calibrated, many different equations are in use that describe the same process, overfitting is a real risk, and pools of N or P running dry can easily lead to unrealistic projections. A separate validation period may help to partially countermeasure these issues. Likewise, a validation period at the start could have downsides if there is still an effect of initial conditions, as biogeochemical models may need a longer spin-up than physical models. So, in this sense, we wonder if the recommendations of Shen et al. could/should be extended to biogeochemical models. Yet at the same time, data availability (both in frequency and period of coverage) is more pressing for biogeochemical variables, so being able to use the full period for calibration would have additional benefits as well. In short, we consider this topic

outside the scope of our present study, but absolutely see the importance of looking further into this. We believe that the aquatic modelling community would benefit from an open discussion on this topic and indeed numerical testing of various methods, to find the advantages and limitations of different calibration and validation strategies.

3D Lake processes are heavily dependent on local atmospheric conditions, so to for the drainage area processes. The authors used five GCM models which by their global nature are course resolved. The GCMs are bias correction toward local measurements in Jiménez-Navarro et al. (2023), but if I understand Supplementary table E 4th column (RMSE) this bias correction is almost nonexistent. Taking the difference between GCM INM-CM5-0 and measurements as an example, mean air temperature RMSE (Root Mean Square Error) drops from the unbiased comparison of 5.712 °K to 5.687 °K after bias correction and for Wind Speed from 4.283 to 2.588 m/s, and improvement with <1% and ~40 % respectively. The bias correction of precipitation, a key input to the drainage area model, looks to have failed. Now I might misunderstand how the Bias correction results are shown, but this illustrate my first point. Can we trust that the calibration is still valid using the climate models as input? Additionally, the reader needs to know why these climate models and scenarios were selected. I suspect because they cover the extreme ranges of for example temperature, precipitation, wind speed etc.. Furthermore since the setup is used for projecting climate effects, is the time frame (for drainage area and lake) long enough so that the models capture the climate trend (which is small compared to seasonal variations)? It would help the reader to see how the trends during the setup/calibration period are in the model compared to measurements.

We would argue that our bias-correction succeeded: the quantile mapping method that was used only aims to decrease the bias. The Supplementary table in Jiménez-Navarro et al. (2023), referred to by the reviewer, indeed shows that Bias significantly decreased and, in many cases, also RMSE, though there were cases in which RMSE slightly increased. With regard to precipitation, the bias correction did not fail, but we admit that we should have used scientific notation to show the RMSE and Bias values, which now appear to be 0 due to the unit (kg/m2/s, or mm/s). For example, for GCM INM-CM5-0, the precipitation bias before correction was $-2.7 \cdot 10^{-6}$ mm/s, and after bias correction it was $-8.5 \cdot 10^{-7}$ mm/s. In some combinations of GCM and variable, the bias correction had indeed little effect, but only because the GCM prediction already was relatively unbiased compared to observations.

A high NSE and low RMSE cannot be expected, as a hindcasted GCM is not intended to simulate the same weather events as were observed (e.g. a storm may pass at a different time than observed), but it should rather reflect the observed weather over a longer temporal scale (as opposed to a reanalysis dataset, which does intend to match observations as close as possible). A biased GCM, however, would be an indication of a biased prediction, and it is this that the quantile mapping mitigated. We therefore don't consider this study to be less accurate in terms of its future projections than other studies, though of course these projections present a large degree of uncertainty.

The five GCMs were selected because they represented a wider range of predictions compared to a single projection, but another main reason was that these GCMs provided all the necessary forcing needed to run both SWAT+ and GOTM-WET. Some other GCMs, for example EC-Earth-Veg and GFDL-ESM4 that were included in an earlier study using SWAT+ in Lake Erken (Jiménez-Navarro et al. 2021, doi:

10.3390/f12121803), missed some variables that were needed to run GOTM-WET (at least without using a different approach compared to the other GCMs).

We will now state in the manuscript that these GCMs provided the required forcing and that they were bias-corrected.

L. 92: Each GCM provided the meteorological forcing required to run both SWAT+ and GOTM-WET and the projections were bias-corrected to locally observed meteorological data using quantile mapping (see Jiménez-Navarro et al., 2023).

As in many climate projection studies, the time frame with measurements is indeed comparatively small to detect climatic trends, and the projected simulations are longer than the period with measurements itself. To test whether our model detected climatic trends during the calibration and validation period, we selected several model output variables that were predicted to show a trend with warming in Jiménez-Navarro et al. (2023): discharge, water temperature, oxygen, and NH_4 concentration. Both 3- and 15-m depths were assessed and annual averages were taken, and for simplicity, gaps in observations were linearly interpolated in order to fit a Mann-Kendall test. At a 0.05 significance level, according to the Mann-Kendall test, only corresponding trends in 3- and 15-m simulated oxygen concentration could be found, and in the observations, none of the variables showed a similar trend as in the climate simulations. In short, the reviewer's question could therefore be answered with "No, the calibration/validation time frame is not long enough to capture a climate trend". It should be noted that over longer periods, climatic trends in historical data of Erken do become visible for physical parameters at least (see Moras et al., 2019, cited in manuscript). Still, we do not consider this a restriction for this study. Considering biogeochemical data, Lake Erken has a comprehensive dataset, covering a longer period than most other sites (even longer data is available, but less regular and less variables, which is why we do not model even further back). As such, for this type of studies, it presents an optimal site to do this, and a lack of climatic trends over a comparatively short period in both observations and model, does not invalidate the use of the model itself. Since these findings are more in line with the "general" model performance, we did not see a convenient place in the manuscript to add this information, as this would rather be added to the manuscript by Jiménez-Navarro et al. (2023). In the current paper, readers can see both observed and simulated spring event timing over the period with observed data in Figure 2. Nevertheless, we hope that this information satisfies the reviewer.

3E Through the analysis of trends from the climate simulations, the authors treat the climate scenarios as constant change over time ex. Fig 3. This is not correct, in fact the gradient for each scenario change over time, especially for SSP 245. I suggest dividing the model output into 30 year chunks while conducting the analysis, or look at the amount of change from a reference to a far future period.

Although the air temperature change is indeed not linear (at least for SSP 2-45), we chose a linear model because it fitted the response well (Figure 3). We retained the use of a linear model for ease of communication and to facilitate a comparison with previous studies, which report phenological trends often as well as "x days per decade". From a practical point of view, the Mann-Kendall analysis allowed us to also assess relative changes. However, we agree that reporting the output as suggested by the reviewer has benefits, and it could further facilitate comparison with other

studies and future meta-analyses. We now additionally report the values for the chunks 1985-2014, 2040-2069, and 2070-2099 in the Supplement and refer to these values in the Results and Discussion. It can indeed be seen there that some future trends do not seem to behave linearly, such as the chlorophyll peak date under SSP 2-45 (although it should be noted that the mid-century period is not in the middle of the other two periods). Additionally, we will shortly discuss linearity and how the slopes should be interpreted, in the Discussion.

Add after L. 224, and after the changes made in response to comments 1O and 2A of the other reviewers: Although the predicted changes in event timing are reported as linear trends, it should be noted that we do not assume that these changes are entirely linear. Especially in SSP 2-45, the development of air temperature through the simulation period is not linear, and the timing of events will not follow a linear trend over time either. In Supplement section S3, averages in separate time periods are reported, and for instance the advance in timing of the spring chlorophyll peak gives an indication of slowing down or stopping in the second half of the century under SSP 2-45. Reported linear changes should therefore be seen as the average change over the period 1985-2100, and we did not investigate the shape of the trend during this period.

New section in the Supplement:

S3. Future projections - Time periods 1985-2014, 2040-2069, and 2070-2099

Table S2. Average values for time periods 1985-2014, 2040-2069, and 2070-2099 under the SSP 2-45 and 5-85 scenarios.

			SSP 2-45		SSP 5-85	
Variable	**Unit**	**1985-2014**	**2040-2069**	**2070-2099**	**2040-2069**	**2070-2099**
Chlorophyll peak date	DOY	108.31	87.46	89.66	86.01	77.64
Peak spring chlorophyll concentration	mg/m^3	14.53	11.71	11.22	11.81	10.84
50% spring discharge date	DOY	78.37	60.43	59.13	57.52	55.31
Cumulative spring discharge	m^3	$8.92 \cdot 10^6$	$1.10 \cdot 10^7$	$1.16 \cdot 10^7$	$1.20 \cdot 10^7$	$1.29 \cdot 10^7$
Ice-off date	DOY	101.93	90.50	83.83	80.96	68.15
Ice-on date	DOY	3.79	28.01	35.03	34.21	45.24
Number of days with ice	days	72.02	31.31	24.16	21.06	7.04
Average ice thickness	m	0.155	0.061	0.048	0.039	0.014
Stratification onset	DOY	140.71	132.26	130.78	128.61	125.65
End of stratification	DOY	261.49	267.89	267.81	271.13	273.29
Number of stratified days	days	122.11	136.56	138.56	143.41	149.11
Average Schmidt stability during stratification	J/m^2	177.71	221.87	232.13	236.61	266.61
Average mixed layer depth during stratification	m	6.53	6.35	6.32	6.00	6.01