Review #2

This manuscript develops polynomial emulators of a stability-corrected energy-balance model for crop canopy temperature and embeds them in LPJmL without adding too much runtime. I thinkk the problem is well motivated, the method is openly archived, and the honest reporting that canopy temperature does not beat air temperature for daily means is a strength. The core idea is sound and within GMD's scope. But the current version has several methodological and evaluation weaknesses that I think are addressable within a revision and that, if addressed, would substantially raise the paper's value to the modeling community.

Thank you very much for the positive evaluation of our manuscript. We highly appreciate the feedback and provide responses to the individual points and suggested changes for improvement of the manuscript below.

My three main concerns are listed below:

1. The fitted air-temperature coefficient of the upper-limit emulator is essentially one (Table A4: 0.959 for the mean setup, 0.997 for the maximum). This shows the polynomial is spending its leading term reconstructing the identity $T_{c,U} \approx T_a$ + small correction, which Eq. (4) already gives in closed form. Maybe you can emulate only the expensive unknown, the stability-corrected aerodynamic resistance, and compute the limits from the physical equations. Because the stability correction depends on the surface temperature, this requires two resistances (r_a for the cool lower-limit case and for the hot upper-limit case), or one neutral r_a plus stability factors, not a single object. The benefit is that the additive dependence on T_a and the multiplicative dependence on R_n are then exact rather than approximated, so the two artifact guards in Table 1 (the $T_{c,L} > T_{c,U}$ check and the +/-10 K clamp) become unnecessary. I want to be precise that this would not remove the physically motivated thresholds (LAI < 1.5, low radiation, and especially low wind, where the EBSC itself is unreliable and r_a is intrinsically hard to emulate).

Thank you very much for this profound and reasoned comment. As you point out, the fitted parameter values corresponding to the air-temperature emulator terms of the daily mean and maximum upper canopy temperature limits are very close to one. We consider this a strength as it indicates that the upper-limit emulators perform as intended, namely emulating equation 4. We think that your suggestion to emulate only the stability-corrected aerodynamic resistance is very reasonable and could reduce the error. We are already working on a follow-up development in LPJmL where we consider to realize that. For this study, we do not see a substantial benefit in changing the whole model approach. Particularly as we assume that weather conditions outside the training range could likewise produce artifacts, e.g. unusually high values for the aerodynamic resistance, resulting again in the need of artifact guards. Instead, we suggest to test the robustness of the emulator's results more extensively. We will evaluate, whether and under which conditions the used thresholds are surpassed.

2. Also I strongly recommand to validate the water-stress scaling factor K_{WS}. Currently the $K_{WS} = \min(1, g_{c,LPJmL} / g_{c,opt})$ divides an LPJmL photosynthesis-based canopy conductance (computed with air temperature, by your own statement in Sect. 2.5) by an optimum conductance defined as $1/r_{c,L}$, where $r_{c,L}$ derives from a grass reference

stomatal resistance (r_1 ≈ 100 s/m, Allen and Wright, 1997). These two conductances are not guaranteed to share a definition or scale, so their ratio can be biased even when each is individually reasonable. Since the emulator reproduces the EBSC to about 1 K while the site RMSE is 3.6 K, the scaling step is one of the candidate sources of the residual error and is currently the least validated link in the chain.

Thanks a lot for pointing this out. We agree that the water-stress scaling factor K_WS needs further evaluation and we will conduct a respective analysis. We will also explain K_WS and its applicability in more detail in the manuscript.

3. The Discussion right now lists four qualitative causes of the site mismatch but quantifies none, so the reader cannot separate emulation error (about 1 K) from the two-limit linear-scaling approximation (Eq. 5 interpolates a quantity that is nonlinear in r_c) from the K_WS proxy. A clean decomposition is feasible with objects you already have, evaluated on the validation cell-days: (i) full EBSC with actual r_c, (ii) the two-limit linear scaling with the true transpiration ratio, (iii) the same with emulated limits, (iv) the LPJmL chain with K_WS. The successive differences attribute the error to each step and would tell future users which part to improve.

An evaluation of the whole model chain would indeed be very insightful, especially to disentangle the sources of error. However, the stand-alone EBSC model unfortunately does not provide a basis for that. That is, because both the stand-alone model and the emulators need to be embedded in a crop model to have information about actual canopy resistances or transpiration rates to compute canopy temperatures with actual r_c or via the two-limit linear scaling. Unfortunately, the experimental data used for the site-specific validation do not include the necessary information. Thus, data on r_c, transpiration ratios or K_WS are not available as inputs for the stand-alone models. To estimate the additional error caused by the model chain we will discuss the relation of the final error to the ones of the EBSC models reported in Webber et al. (2017) more thoroughly. Furthermore, we will evaluate the errors of the single components more comprehensively. Particularly, we will conduct a sensitivity analysis to challenge the assumptions we use to compute daily maximum canopy temperatures with the EBSC model (first component, see also response to reviewer 1, point 4.) and we will evaluate the water scaling factor K_WS in LPJmL (third component, see also point 2.).

4. The current cross-validation draws the validation set from the same distribution as the training set, so it demonstrates interpolation skill, whereas the paper's motivation is future, hotter, higher-VPD conditions, which is extrapolation. Please add an explicit out-of-domain test: hold out the upper decile of temperature or VPD, or hold out the entire future SSP5-8.5 subset, refit on the remainder, and report emulator-versus-EBSC error in the held-out tail. Relatedly, the training LAI comes from a standard air-temperature-driven LPJmL run, but the emulator is applied inside a canopy-temperature-driven run where phenology and LAI will differ. This is harmless only if runtime inputs stay inside the training envelope, and it is precisely the heat-stressed, water-limited cells that are most likely to drift outside it.

Thank you very much for this comment. We agree that using training and validation datasets built from separate sources is preferable. However, if parts of the future SSP5-8.5 scenario are held out, the production of artifacts becomes more likely and the emulators become less robust and suitable for future projections. That is why we suggest to stick with the existing approach but additionally cross-validate the emulators with a different SSP scenario, namely SSP3-7.0, and with climate data from a different climate model.

We also agree that a systematic one-directional bias in LAI could be problematic as it might shift the training data set towards higher LAI values, while smaller LAI values might be missing. To address this, we conducted an analysis, which we will include in the revised manuscript. Our analysis does not indicate any shift, which we attribute to the quantity of data and the underlying variety of growth conditions. The LAI dataset shows full coverage between the crop specific minimum and maximum LAI values across all value ranges of the other regressors, most importantly 2 m air temperatures.

5. For the maximum setup you substitute T_max, VPD_max, and a 3 pm radiation value (Eqs. 8 to 10), but wind speed appears to be left at its daily mean even though afternoon wind and atmospheric stability differ substantially and r_a is sensitive to both. Please state explicitly how wind is handled in the maximum setup. Separately, the soil heat flux G appears in Eqs. (1), (2), (4) but is not one of the six regressors and its computation is never described (set to zero for daily means, a fixed fraction of R_n, and how it is treated at 3 pm where it is not negligible).

Indeed, different studies indicate that wind speeds can have a diurnal pattern with a maximum around noon (e.g. Peterson and Parton (1983), and Ephrat et al. (1996)). However, the publications we found that provide simple methods to compute afternoon wind are based on empirical approaches for specific locations that are not necessarily applicable globally. That is why, in our study, we use mean wind speeds for both cases, i.e. the computation of both mean and maximum daily canopy temperatures. In order to test whether this assumption is valid, we will conduct a sensitivity analysis including offsets between daily mean and afternoon wind speed derived from the empirical relationships mentioned.

We approximate the soil heat flux G to be 10% of R_n as was done in Webber et al. (2016) and described in Allen et al. (1998) for both cases (daily mean and daily maximum). We will include a description and the corresponding literature in the revised manuscript.

6. Also, ERA5 daily-maximum skin temperature is not a direct measure of crop canopy temperature (or how you assess the gaps between skin temperature of the models vs the canopy temperature?) and carries two opposing biases of unknown net size: a warm representativeness bias, because the 0.5° value mixes in hot midday bare soil and sparse cover, and a cold model bias, because ERA5 daytime land-surface temperature is documented to be underestimated where vegetation cover is overestimated (Johannsen et al., 2019, Iberian Peninsula, correlation about -0.45). Since the sign of the net mismatch is unknown, the quantitative agreement for daily-maximum Tc (bias, slope, MSD partition in Figs. 5 and 6) should not be read as a clean measure of skill; only the qualitative ranking is robust, because air temperature underestimates the surface maximum under either bias. We recommend stating the result at the ranking level and, for a cleaner test, comparing against satellite land-surface temperature at a known overpass time (e.g. MODIS Aqua near 13:30, aggregated from native 1 km and restricted to high crop-fraction cells), which removes the reanalysis bias and the constructed daily maximum. But at the same time, the remote sensing derived skin temperature will also have bias with the real canopy temperature, but it shall be good to see the differences between models and remote sensing products.

Thank you very much for this interesting comment and the suggestion. We agree that a comparison of canopy temperatures against measured satellite LST data with higher resolution would be very insightful. However, unlike the modelled ERA5 data, satellite data does not include 2 m air temperatures, which is a crucial variable in our analysis for both computing canopy temperatures and evaluating their accuracy compared to the standard 2 m air temperature input. For our specific application, the ERA5 data has a valuable benefit as it provides a consistent set of skin and 2 m air temperature data. At the same time, we acknowledge the drawbacks you have pointed out and we will discuss them in the revised manuscript. Moreover, we will refine the interpretation of the ERA5 comparison (see also response to reviewer 1, point 1).

Bibliography

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., et al.: FAO Irrigation and drainage paper No. 56, Rome: food and agriculture organization of the United Nations, 56, 1998.

Ephrath, J. E., Goudriaan, J., and Marani, A.: Modelling diurnal patterns of air temperature, radiation wind speed and relative humidity by equations from daily characteristics, Agricultural systems, 51, 377-393, 1996.

Peterson, T. C., and Parton, W. J.: Diurnal variations of wind speeds at a shortgrass prairie site—a model, Agricultural Meteorology, 28, 365-374, 1983.

Webber, H., Ewert, F., Kimball, B., Siebert, S., White, J. W., Wall, G., Ottman, M., Trawally, D., and Gaiser, T.: Simulating canopy temperature for modelling heat stress in cereals, Environmental Modelling & Software, 77, 143–155, 2016.

Webber, H., Martre, P., Asseng, S., Kimball, B., White, J., Ottman, M., Wall, G. W., De Sanctis, G., Doltra, J., Grant, R., et al.: Canopy temperature for simulation of heat stress in irrigated wheat in a semi-arid environment: A multi-model comparison, Field Crops Research, 202, 21–35, 2017.