Interactive comment on “Using a data-driven statistical model to better evaluate surface turbulent heat fluxes in weather and climate numerical models: a demonstration study (https://doi.org/10.5194/egusphere-2024-568)”

Dear reviewer 1,

We thank the reviewer for his/her helpful suggestions, which led to significant improvements of our paper. Below we detailed how his/her comments are addressed in the revised version of the paper. The corrections are cited here in *italic*. *We refer to specific pages by “P” and lines by “L”. For example, “**P1, L1**” stand for page 1, line 1 in the revised version.*

General comment: Some of the grammar/language shall be revisited

We fully agree with this comment.

Specific comments:

Line 25: Mention also the first important source of bias.

The first source of bias in the numerical simulations is now mentioned. ***P02, L25-26***: “ *However, the representation of convection and surface processes are the two most important sources of systematic biases in numerical simulations (Zadra et al., 2018; Frassoni et al., 2023).”*

Line 74: In section 5

Correction done.

3.1:

Why not splitting the training datasets in each stability category (Stable, Neutral, Unstable) and different friction velocity ranges.

Splitting the training datasets based on surface stability and friction velocity is certainly a valid approach in some contexts. However, in our case, this suggests building separate statistical models for each category with not evenly distributing the data, with the risk of increasing the problem complexity (finding hidden layer topology) and imposing artificial discontinuities. Moreover, data-driven models trained on specific categories and ranges may struggle to generalize to unseen data, particularly when it falls near the boundaries of the defined categories. Therefore, we decide to build a single statistical model on the full range of conditions, using temperature and humidity vertical gradients (Δ_θ, Δ_q) and vertical wind shear (Δ_U) in the input parameters (Figure 5 and Table 2). In this way, the model learns to handle transitions between categories for better generalization in the fluxes' estimates.

Line 157: (ii) instead of (iii)

Correction done.

3.2:

Coarse time resolution (half-hour averages data may mask physical information for smaller time scales, at what time step the simulations were run?)

The simulation time step is of the order of a minute (90s), and output data have been archived at a coarser time resolution (3-hourly instantaneous or averages (depending on variables) and daily averages).

Lines 201-202:

The most conventional meteorological variables, such as T and RH at 2 m agl are also available. The lowest level as mentioned is within 20 m (pressure level akin to sigma coordinate which dynamically changes with time), so half eta-level is still higher than 2m. Any interpolation here?

Concerning numerical simulation, the atmospheric variables at conventional heights of observations (2 and 10 m agl) are diagnostic parameters. This means that they are output of the modeling system, calculated by interpolating between Earth's surface and lowest model level just above, using MOST stability functions. We avoid as much as possible the use of diagnostics that are based on a theory we want to assess.

Although, you mentioned this here Lines 303-305:

Moreover, the meteorological variables are directly taken at the first half-eta level (M=1, around 8 magl) instead of diagnostic variables as much as possible.

But 2m and 8m are quite far especially under stable conditions… Maybe, you need to comment on this..

The statement was modified in ***P15, L307-309***:

"*Diagnostic variables derived from numerical simulation (T2m, RH2m) are susceptible to contain bias due to the interpolation technique (based on MOST) and the inconsistency of terrain elevation. To avoid these uncertainties, the meteorological variables are taken directly at the first half-eta level (M=1, around 8 m agl) as much as possible.*"

Line 202:

The data are stored at a temporal resolution of 3 hours …

How did you compare the instantaneous 3rd hour snapshot with observations averaged in half-hour chunks?

Thank you for this comment which is very relevant because it points out a limitation of the model's evaluation (particularly for this method, but also in more classical approaches.) Indeed, simulated fluxes (turbulent and radiative) are not instantaneous but correspond to a time-centered mean over 3-hour, while temperature, humidity and wind data and vertical gradients are instantaneous. In the revised version, a section was added (Section 5.1) to clearly discuss the impact of the temporal resolution in both the comparison and the performance of the statistical model.

Line 204:

Meanwhile, the surface data, mostly provided by ORCHIDEE, consist of time-centred mean over a 3 hours window…

Specify what **surface data**…

The statement was completely revised (***P10, 205-212***).

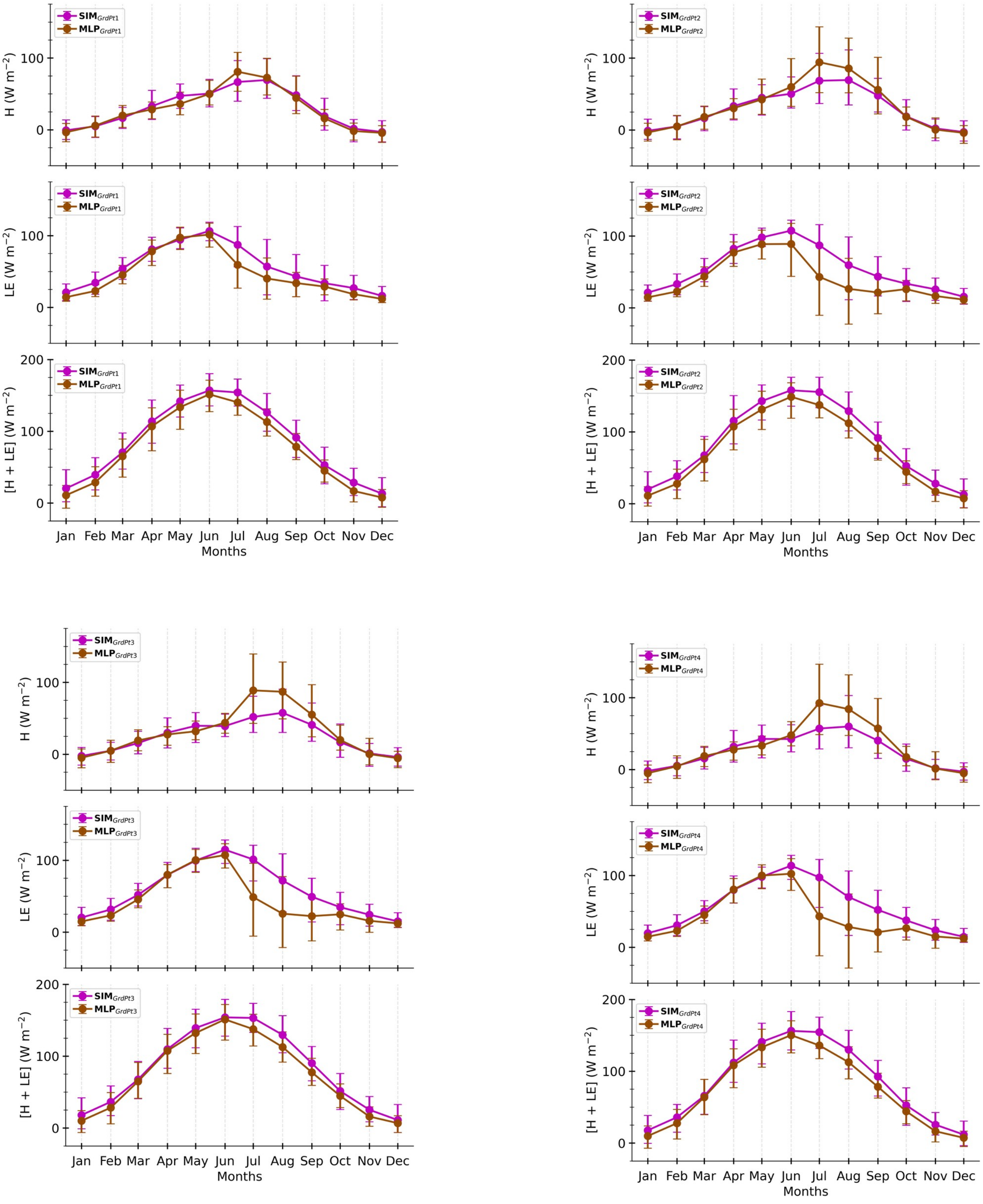
Lines 209-213:

Could you show in Fig. 3 the grid-layout that shows the real observation geographic coordinate and the 2 nearest grid cells considered in the analysis.

Figure 3 now displays the grid mesh surrounding the geographical location of the Meteopole site (***P11***).

In this context, given the locality of space (dx=dy=20 Km) and the point observations collected from a tower at specific coordinates, have you done sensitivity analysis on the effect of spatial averaging?

We thank the reviewer for his comments. Our method seeks to better identify the land surface errors by performing the comparison within environment from the numerical model. The surface turbulent fluxes of the RegIPSL model are already derived from the weighted average of the parameters (roughness, albedo, etc.) over the different types of land cover present in grid cell in the same atmospheric forcing (***P10, 198-200***). Averaging the fluxes across grid cells would have been difficult to discuss, since environmental forcing and fluxes from different landscape composition are mixed. Hence, for sensitivity analyses, comparisons were initially made at the four grid cells closest to the observation site (see figures below). Finally, the paper presents the results for the two closest grid cells, as the other two grid cells yield similar conclusions.



Composites monthly averages of simulated (lines in magenta) H, LE H+LE at the four nearest grid cells to the Météopole station, respectively, together with that of MLP-based estimates in simulated environment (lines in brown). The solid lines correspond to the means and the error bars represent the 10 and 90th percentiles, calculated by gathering the daily averages of 3-hourly data. All the selected diurnal cycles from 01 January 2012 to 31 December 2016 were considered here.

Also, any comments on the local dynamics forcings in the considered two grid cells compared to the observations? You considered the surface type aspect as a main criterion in selecting these 2 grid cells, but what about the topography effects and the local environmental dynamics (wind speed and directionality, temperature, …) when comparing outputs in these 2 grid cells to the observation’s dataset at that specific grid location?

Our method is designed to evaluate the simulated fluxes within the simulated environment. The data-driven statistical model learns on measurements to provide potentially observed heat fluxes in the simulated environment. The benefit is that it is not necessary to have identical time series of environmental forcing between the simulation and observation. Aside from soil moisture (SM), the input variables at the two nearest grid cells exhibit a fairly similar range of variability (see Figure A1). The soil is slightly drier at the second grid cell (GrdPt2, Figure 3), likely because of different landscape composition, but the difference is weak compared to the variability obtained in the learning and test datasets.

3.3.1:

What’s the criteria for selecting the training dataset?

The learning set was selected to adequately cover the inter-annual variability of turbulent fluxes with as many samples as possible.

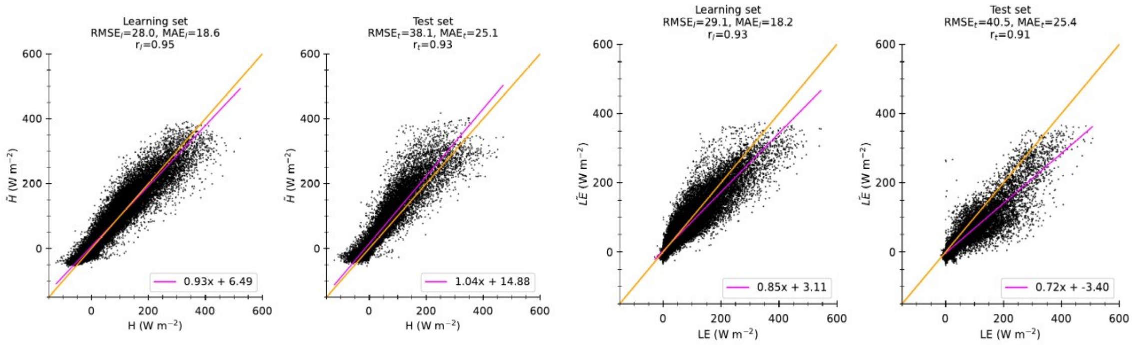
Is it season independent?

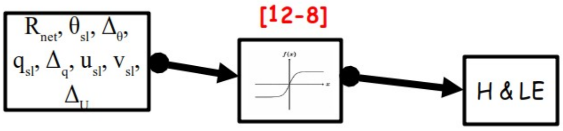
Fortunately, learning data examples are well distributed across the four typical seasons (Figure 2b, ***P09***).

Have you done any sensitivity analysis on choosing different dataset i.e., dataset convergence in terms of variability in the final weights and biases of MLP?

Yes, sensitivity analyses specifically focused on the choice of different datasets (including inputs of statistical model) and hyperparameters of MLP have been performed. The paper presents the experiment with the best results in terms of RMSE and correlation. The figures below show results of some sensitivity experiments performed for choosing input variables.

➔ Tests on input variables





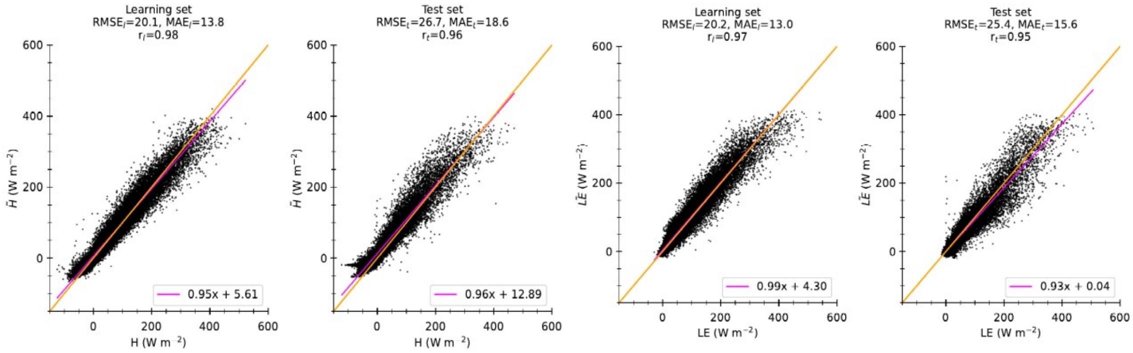
R_{net}, θ_{sl}, Δ_θ, q_{sl}, Δ_q, u_{sl}, v_{sl}, Δ_U

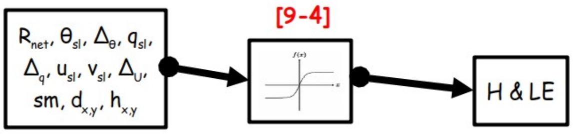
Learning = {2015, 2016, 2017, 2019, 2020}
Test = {2012, 2013, 2014, 2018, 2021}

Correlation > 0.9

These 8 input variables do not generalize correctly to unknown observed conditions.

➔ Tests on input variables





R_{net}, θ_{sl}, Δ_θ, q_{sl}, Δ_q, u_{sl}, v_{sl}, Δ_U, sm, $d_{x,y}$, $h_{x,y}$

Apprentissage = {2015, 2016, 2017, 2019, 2020}
Test = {2012, 2013, 2014, 2018, 2021}

Generalization is substantially enhanced with these 13 variables.

3.3.2:

Lines 298-300: Define these variable acronyms especially:

Eventually, 4 trigonometric temporal coordinates are added for the description of seasonal (dx, dy) and diurnal (hx, hy) cycles.

All MLP inputs are defined in Table 2. The paragraph was modified, reference to Table is now at the beginning.