

July 02, 2026

L'Aquila, Italy

Referee responses

Manuscript: Daily maps of Boundary Layer Height combining radiosonde, satellite, and reanalysis over Europe

Authors: Carina Inés Argañaraz, Andreu Salcedo-Bosch, Simone Lolli, and Gabriele Curci

The authors would like to thank the reviewers for their very pertinent suggestions. We have considered each one carefully and have worked to improve the manuscript. The methodology section has been reorganised and explained in more detail, the results section expanded by adding new evaluation parameters and topographic covariate, and the discussion section expanded in line with the reviewers' suggestions.

Referee #1

This manuscript derives continuous boundary layer height (BLH) over Europe by combining radiosondes data and models fitted from spatiotemporally collocated meteorology and topography covariates. The main product presents BLH at 0 and 12 UTC at monthly resolution. A regression kriging (RK) method is implemented to model BLH, where different regression models are evaluated using cross-validation and spatiotemporal testing by independent soundings. The tests show the model outperforms ERA5 BLH. Daily BLHs are attempted over Germany in three 15-day periods. The results suit the scope of AMT, and the following comments helpfully help improving the manuscript before publication.

Reply: We would like to thank the reviewer for reading our manuscript and for their valuable suggestions, specially referred to reinforce methodology and procedures. We have considered all these suggestions in order to improve our work.

One main issue is that “daily” in the title does not well represent the core work, which is monthly, RK-interpolated BLH over Europe. The daily component over Germany appears to be a weaker attachment to the Europe study.

Reply: We agree with this observation. We have therefore changed the title to “Modelling monthly Boundary Layer Height maps combining radiosonde, satellite, and reanalysis over Europe” to better reflect the focus and scope of our study.

It seemingly applied the same modeling framework from the European study, just changing data to the three short periods in Germany, but without the independent testing. As acknowledged towards the end of the Discussion section, it is not possible to assert whether these daily-model-predicted BLHs are better or worse than plain ERA5. Since ERA5 BLH is already available hourly and globally for many decades, the significance of the results and analysis (qualitative association between modeled BLH and weather conditions) in the mesoscale case study is questionable. The clarity and overall quality of the corresponding sections are also lower than the European work. The authors are suggested to either enhance the rigor of this case study, or remove it as the manuscript is already on the long side.

Reply: The aim in presenting this case study was to apply the same methodology on a smaller spatial scale and evaluate its performance. We agree with the reviewer's observation

regarding the lack of independent testing stations for a proper validation of the results. Therefore, we have decided to remove this section from the manuscript.

The independent assessments using spatial and temporal radiosonde sites not in training are considered a strength of this work. It is suggested to follow more standard machine learning terminology, using “validation” as defined in section 2.3 to select models and tune hyperparameters, while using “testing” for the independent assessment of model performance. Then, only section 2.5.1 is validation as it is part of the training process, and sections 2.5.2 and 2.5.3 are spatial and temporal testing. It reads like the spatial sites in 2.5.2 are independent of the training process, but it is good to be explicitly confirmed. In addition, hyperparameter tuning appears to be missing this work, but random forest (RF) and gradient boosting (GB) are large families of models. Key hyperparameters will determine under- vs. over-fitting and should be optimized in cross-validation. It won't be fair to simply compare untuned RF or GB with linear regression.

Reply: We have reorganized the Methodology section in line with the reviewer's suggestions. Section 2.4 (Validation) now describes the cross-validation procedure (k-fold) and the diagnostic metrics used to evaluate model performance. Section 2.5 (Testing) describes the procedures applied for the independent temporal and spatial validation using the best-performing models obtained. Additionally, Section 2.3 now includes the description of the hyperparameter selection process for the two machine learning models.

This work may be further strengthened and advanced from prior works if the advantages of the interpolation (i.e., kriging) part can be quantified. It may be worth more details in methods and results if kriging makes significant contributions besides the optimized regression models.

Reply: We agree that it is important to distinguish between the contributions of the regression and kriging components. This provides a clearer assessment of their individual performance and allows us to discuss their relevance within the proposed methodology more effectively. In addition to the diagnostic parameters already considered, we have included the coefficient of determination (R^2) in Section 2.4.1 Diagnostic parameters.

Specific comments

Last sentence of the abstract: It is unclear by reading the main text how the BLH product can be incorporated in atmospheric model simulation.

Reply: This statement has been removed from the abstract, because it was specifically related to the case study that have also been removed.

Section 2.2.3: The importance of topography is highlighted in the manuscript, but only the altitude is considered. Are higher order topography terms worth consideration? The same altitude may be on a flat plateau or mountain ranges.

Lines 686-689: Not sure if altitude alone can present the flatness. Consider more specific topographical indices.

Reply: Based on these two observations, we decided to include the Terrain Ruggedness Index (TRI) as an additional covariate (Lines 260-265), which showed good performance at 12:00 UTC. This allowed us to better assess and discuss the relevance of topographic effects in our results.

Tables: Tables B1 and B2 are referred to as “1B” and “2B” in the text. The tables are better placed within the corresponding sections. It is unclear what “a” and “b” means in Table 1.

Reply: We have reorganised the tables to better align with the structure and sequence of the text. The subscripts (a and b) have been removed, as they were considered unnecessary.

Section 2.3: Consider streamlining and clarifying the definition of K-fold cross-validation. It is mentioned before formal definition.

Reply: The section was reorganised and K-fold was redefined in *Section 2.4. Validation*.

Line 289: Please provide more information about how the loss function is constructed via GLS, especially the construction of observation error covariance matrix (if that is what GLS implies) and the rationales/assumptions.

Reply: We have added further details regarding the regression kriging methodology, providing additional context and a clearer explanation of the approach for the readers.

Lines 332 and 347: It is unclear what a, b, and c means.

Reply: This section has been reorganised to make it clearer and to avoid the use of irrelevant references.

Line 336: ERA5 BLH is referred to as both upper and lower cases, BLH_ERA5 and blh_ERA5. Is the lower case reserved for testing and the upper case as a covariate in regression? Please clarify.

Reply: We understand that this point may have caused some confusion, as the present study uses two different BLH data sources: radiosonde observations (BLH_{RS}) and BLH data extracted from ERA5. The ERA5-derived BLH data are used for two different purposes: (1) as a covariate in the regression kriging (RK) models (BLH_{ERA5}), and (2) as an independent dataset for comparison with the RK model outputs in the Testing section. To facilitate differentiation between these two uses, we renamed the ERA5 BLH data used for comparison as ERA5_{blh} in the Testing section. We have clarified these distinctions throughout the manuscript.

Line 343 and 441: It is unclear how GLMM is used here, what it accomplishes, and how it relates to/differs from the regression models (RF, GB, LR).

Reply: The GLMM was used to identify significant differences between the BLH values obtained from the RK models, ERA5, and the observations, allowing us to statistically assess whether the models provided a better or poorer agreement with the observed values and available products. This section has been revised and clarified in *Section 2.5.2 Temporal testing*.

Section 3.1: This appear to be a comparison of training loss, which is not as important for model selection as validation loss in 3.2.1. Reconsider allocating two long tables for this.

Respuesta: This section has been reorganised to make it clearer and consistent with the text.

Line 615: 12 UTC does not necessarily means convective BL.

Response: We agree with the reviewer that making a direct reference to CBL at this hour was not entirely appropriate, so this sentence has been rewritten.

Line 642: The regression part in RK is still machine learning models. The kriging contribution is better to be justified as mentioned above.

Reply: We have expanded this methodological discussion section by providing a more detailed analysis of how kriging contributes to the regression component. This section has

been rewritten to highlight its advantages and limitations more clearly. Please see *Section 2.3 Interpolation Method*

Referee #2

General Comments

This manuscript presents a methodology for generating gridded boundary layer height (BLH) maps over Europe by combining radiosonde observations, ERA5 reanalysis, topographic information, and satellite-derived variables through regression kriging and machine learning approaches. The study addresses an important topic, as accurate BLH estimates are relevant for weather forecasting, air quality applications, and climate modeling. The manuscript is generally well organized and presents a comprehensive methodological framework. The comparison between different regression approaches and the independent validation exercises are valuable aspects of the study. The results indicate that combining observational and reanalysis datasets can improve BLH estimates compared to ERA5 alone, particularly during daytime conditions. However, several aspects require clarification and further discussion before publication. In particular, I believe that the manuscript would benefit from a deeper physical interpretation of the results, a more critical discussion of ERA5 limitations (especially during stable nighttime conditions), and a clearer justification of some methodological choices.

Reply: We would like to thank the reviewer for reading our manuscript and for their valuable suggestions, specially referred to reinforce discussion and concepts. We have considered all these suggestions to improve our work.

Overall, I recommend

Major Revisions.

Major Comments

1. Limitations of ERA5 during stable boundary layer conditions: Throughout the manuscript, ERA5 is presented as a generally reliable predictor of BLH. While this statement is supported by several studies, the discussion remains largely focused on overall performance metrics. A substantial body of literature has shown that ERA5 exhibits significant difficulties in reproducing stable boundary layers, particularly during nighttime conditions. This issue is especially relevant because the manuscript consistently reports poorer performance at 00:00 UTC. The discussion should explicitly address the known limitations of ERA5 under stable conditions and explain how these limitations may propagate into the proposed interpolation framework.

Reply: We have considered this suggestion and have expanded the discussion regarding the limitations of ERA5 at night and the results obtained, along with the resulting limitation in the methodology applied. Lines 680-710.

2. Representativeness of ERA5 grid points: The manuscript states that ERA5 values were extracted from the nearest grid point to each radiosonde station. However, it is unclear how situations were handled when multiple radiosonde stations fall within the same ERA5 grid cell. Given the relatively coarse spatial resolution (~25 km), this issue could potentially affect the independence of the predictor dataset. The authors should clarify: i) How frequently this situation

occurred; ii) Whether multiple stations shared the same ERA5 pixel, iii) Whether this could introduce artificial correlations between predictors and observations.

Reply: In our study, no data overlap was present, as the stations were separated by a minimum distance of 100 km (Figures 1b and 1c). We have included additional details regarding the minimum and average distances between stations for each UTC time. These details are now provided in *Section 2.1*.

3. Justification for using MODIS land surface temperature; The manuscript employs MODIS land surface temperature (LST) as an additional predictor. However, ERA5 already includes surface temperature information and a complete surface energy balance representation. Therefore, it is not immediately obvious why satellite-derived LST should provide independent information beyond what is already contained within ERA5. The authors should better justify why MODIS LST was selected instead of ERA5 surface temperature and what additional information MODIS contributes.

Reply: We selected MODIS LST as a covariate instead of ERA5 surface temperature based on the spatial approach and methodology of our study. In a spatial modelling framework, the accuracy with which environmental variables represent local conditions at each location (in this case, meteorological stations) is an important source of uncertainty. Since the observations are associated with specific spatial coordinates (latitude and longitude), a finer spatial resolution increases the likelihood that the extracted values represent the local conditions more accurately.

The LST values extracted for each station have a spatial resolution of 1 km², compared with the approximately 25 km² resolution of ERA5 data. Therefore, these values provide a more detailed representation of local surface conditions and are expected to have a stronger relationship with the radiosonde observations used for model training. Additionally, although secondary in importance, another consideration was the need to include covariates derived from as independent sources as possible, considering their methodological and/or instrumental origins. This approach helps to minimise potential error propagation associated with the limitations of a single data source.

We added (briefly) these ideas in section 2.2 *Covariates*. Additionally, we discuss the relevance of the spatial resolution in our work, lines 800-810.

4. Necessity of auxiliary datasets when radiosonde observations are available: The manuscript introduces several auxiliary variables to improve BLH estimation. However, from the reader's perspective, it is not entirely clear why some of these variables are necessary when direct radiosonde-derived BLH observations are already available at the training locations. A more explicit conceptual explanation of the interpolation strategy would help the reader understand the role of each predictor and its physical contribution to the final product.

Reply: We have provided a more detailed description of the regression kriging methodology, including the role of spatial data, the regression variables, the required assumptions, and other relevant methodological aspects in *Section 2.3 Interpolation Method*

5. Physical Interpretation of Predictor Importance: While the manuscript provides a thorough statistical evaluation of model performance, the physical interpretation of the selected predictors remains limited. Beyond identifying which variables improve the interpolation skill, it would be valuable to discuss the underlying atmospheric processes that may explain these relationships.

For example:

- Why does land surface temperature (LST) emerge as one of the most important predictors for daytime boundary layer height?
- Why does elevation (DEM) appear to contribute more strongly to nighttime BLH estimates?
- Why do wind-related variables provide relatively limited improvements despite their known influence on boundary layer development and turbulent mixing?

A deeper discussion of the physical mechanisms linking these predictors to BLH variability would considerably strengthen the manuscript and help readers understand whether the reported relationships are physically meaningful or primarily statistical. Such an analysis would also improve confidence in the applicability of the proposed framework beyond the specific training dataset.

Reply: We appreciate this suggestion and have revised the covariates section. Specifically, we expanded the description and discussion of the variables that were found to be most relevant in our results. In addition, we incorporated a new section addressing the potential limitations associated with the data sources used and their implications for the applicability of the proposed methodology, *Section 4.2.5 Covariates limitation*. For example, specifically for the examples listed by the reviewer:

- Why does land surface temperature (LST) emerge as one of the most important predictors for daytime boundary layer height?
See discussion around lines 755-785 of the revised manuscript
- Why does elevation (DEM) appear to contribute more strongly to nighttime BLH estimates?
See discussion around lines 715-740 of the revised manuscript
- Why do wind-related variables provide relatively limited improvements despite their known influence on boundary layer development and turbulent mixing?
See discussion around lines 780-810 of the revised manuscript

6. Applicability for atmospheric modeling: One of the motivations presented by the authors is the potential incorporation of these BLH maps into atmospheric models. However, the proposed product is based on only two synoptic times (00:00 and 12:00 UTC) and monthly climatological averages. This raises important questions regarding its direct applicability to mesoscale or operational atmospheric models, which require much higher temporal resolution. The authors should better discuss the limitations of the dataset for operational and modeling applications and clarify the intended use of the resulting product.

Reply: We agree that, in the first version, the approach to implementing the results obtained in a climate model was not made sufficiently clear. This proposal has been omitted from the new version of the paper, as we have decided to remove the case study. Unfortunately, we do not currently hold any new data with which to test the results. However, this idea will be revisited in future projects.

Minor Comments

1) Keywords - The keyword list could be improved by including: Boundary Layer Height; Machine Learning

Reply: We have added ‘Machine Learning’ to the keywords. Since ‘Boundary Layer Height’ is already included in the title.

2) Section 2.3 (Cross-validation): The choice of $k = 5$ and $k = 3$ appears reasonable, but additional justification would be useful. It would also be helpful to report the sensitivity of the results to the selected cross-validation configuration.

Reply: We performed a sensitivity analysis on the choice of k , which is reported in *Section 2.4 Validation*. Based on this, we selected $k=5$ and $k=4$, for the different sample size considered in this study.

3) Discussion Section - The manuscript would benefit from a broader discussion comparing the proposed methodology with recent machine-learning-based BLH products developed for other regions.

Reply: We compared our result with recent that studies using ML methods to modelling BLH in other regions, such as: Ayazpour et al. (2023) (EEUU), Guo et al. (2024) (worldwide), Zhang et al. (2025) (EEUU), and Stapleton et al. (2025) (Amazonia). Lines 610-625.

4) The objectives could be rephrased to explicitly highlight the novelty: "...to generate observation-constrained gridded BLH maps for Europe using a hybrid regression-kriging framework."

Reply: This suggestion was considered and added in the objective paragraph



Dra. Carina I. Argañaraz

Dipartimento di Scienze Fisiche e Chimiche
Centro di Eccellenza CETEMPS
Universita' degli Studi dell'Aquila
Via Vetoio, 67100 Coppito - L'Aquila ITALY