Dear Referee,

Thank you for your time to review our manuscript and for all your constructive suggestions considering our study. It helped to improve the quality of the manuscript. We reply to your comments below. Our response to the comments appears in bold and revised text as *italic*.

Main comment:

All methods and comparisons are presented together, which overwhelms the narrative. I strongly suggest reorganizing the content into a more hierarchical or modular format. For instance, the manuscript could be structured as follows: a) Evaluation of the noise correction methods for C_{nn}, b) Performance assessment using remote sensing vs eddy covariance constraints, and c) Sensitivity to turbulence scaling assumptions (free vs full MOST).

This is just a suggested outline, and the author is welcome to be creative. However, a clearer and more modular structure would greatly improve the readability and impact of the paper.

We understand the reasoning of the reviewer. We would like to point out that point a) is actually covered in our previous paper: van der Valk et al. (2025), which is referred to in multiple places of the manuscript. Regarding the structure of the results and discussion, we have decided to add subsection headers to improve readability, also based on the comments of the other reviewers. In the results section, we added the following subsections: *4.1.1 Energy-balance method versus two-wavelength method, 4.1.2 C_{nn} Correction methods, 4.1.3 Free Convection scaling, 4.1.4 Comparison with alternative L_vE methods*

In the discussion section, we added the following subsections: *5.1 Energy-balance method versus two-wavelength method, 5.2 MOST L_vE estimates versus free-convection* and *5.3 Potential of CMLs to estimate L_vE*

Specific comments

- In my opinion, Figure 1 does not represent the core contribution of the paper and mainly provides background or contextual information. Therefore, I suggest moving it to the supplementary materials.
  As we think it is important for the reader to understand what are the implications of applying the free-convection scaling instead of the complete MOST scaling, we decided to explain this in the main text. Deriving this from the equations would be difficult.

- The manuscript states that the observation period spans from April 1 to October 1. However, the full time series (TS) over this period is not shown anywhere. While it is reasonable to highlight selected days to illustrate the diurnal cycle, an overview of the entire time series is important to assess the consistency and overall performance of the method.
  We agree with the reviewer that these two example time series are not illustrative of the performance over the entire data period. Therefore, we changed these time series to median diurnal cycles per month for all methods, including the uncertainty for the CML methods (25th and 75th percentiles). As a result, the text is changed as follows:
  A comparison of the reference methods for the full data period shows an overall comparable behaviour (Fig. 3), indicating negligible influence of differences in footprints between the two references. On average, however, the MWS-2λ method shows an overestimation in

comparison to EC. For the monthly median diurnal cycles, all CML methods with the spectral noise correction overestimate L_vE compared to the references (Fig. 4). This overestimation is roughly constant throughout our data period, with the exception of June, during which the overestimation in comparison to the reference methods is larger. The diurnal cycle is well captured by all CML methods, even when considering the interquartile range. Noteworthy, is the similarity between the CML-2λ and CML-EBM-OBS methods in the diurnal cycles, both in median values and uncertainty. The clearest differences between these methods occur before sunset, when the CML-2λ method overestimates L_vE, while the CML-EBM-OBS estimates are constrained by the measured net radiation.

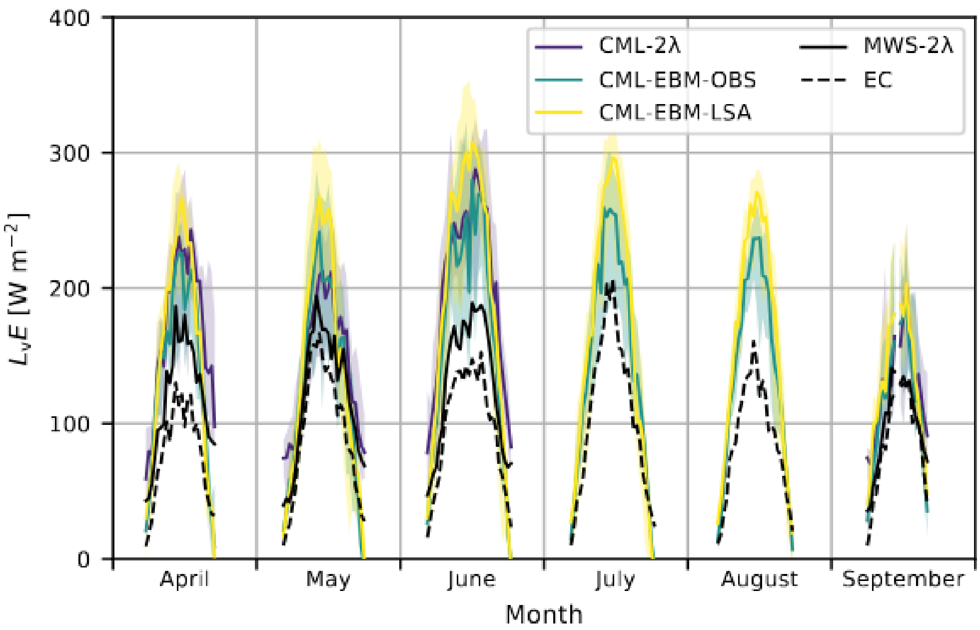

Figure 4. Median diurnal cycles and interquartile ranges (shading) of 30-min L_vE estimates of the methods for the Nokia CML in comparison with the reference instruments. CML-2λ uses the two-wavelength scintillation method with the CML and LAS, CML-EBM-OBS uses the measured energy balance method as constraint to infer the turbulent heat fluxes and CML-EBM-LSA uses the estimated net radiation by LSA SAF instead of the measured net radiation. Shown estimates are obtained using the spectral noise method for the CML estimates and complete Monin-Obukhov scaling. We removed timestamps with less than 10 available days in order to obtain representative diurnal cycles. We refer the reader to Appendix A for a complete overview of the used abbreviations.

- For the Monin–Obukhov Similarity Theory (MOST) scaling, two additional parameters— z0m (roughness length for momentum) and d (displacement height)—are required. These parameters depend on the characteristics of the roughness elements, vegetation height, and atmospheric conditions. However, the manuscript does not provide any information on how these values were determined or set.

  We indeed did not mention these in the methodology section. Our apologies for this omission. For the roughness length at the location of our experimental setup (Cabauw, the Netherlands), we use the values reported in the MSc thesis from Robert Moonen (2021), who determined the roughness lengths based on EC data, depending on wind direction: 40° – 100°: 0.03 m

100° – 160°: 0.01 m
160° – 280°: 0.005 m
280° – 40°: 0.01 m
These values are in the same order as found by Verkaik and Holtslag (2007).

We relate the displacement height to the roughness length based on Brutsaert (1982):
$d_0 = 16/3\ z_0$

We added this into Section 2 as follows:
Thus, next to the scintillometer measurements, the two-wavelength method requires temperature, humidity and horizontal wind speed measurements. Additionally, it requires more complex measurements of the correlation between temperature and humidity fluctuations, the roughness length for momentum and the displacement height. For the roughness length z_0, we make use of values reported by Moonen (2021), who determined the roughness length for various wind sectors using EC data (Table 1) and found similar values to Verkaik and Holtslag (2007). For the displacement height d, we use the relationship $d = 16/3\ z_0$ from Brutsaert (1982).
In order to obtain the turbulent heat fluxes, Eqs. (4 - 7) have to be solved iteratively. In figures, we refer to the two-wavelength method using the suffix 2λ.

Table 1. Lower and upper bound of wind direction classes and roughness length z_0 values as determined by Moonen (2021).

Lower bound [°]	Upper bound [°]	z_0 [m]
40	100	0.03
100	160	0.01
160	280	0.005
280	40	0.01

- The current version of Figure 6 is not very clear or easily readable, especially regarding the size of the symbols in relation to the scale of the graph.
  We agree that identifying individual symbols for some methods is hard, given the selected limits of the y-axis. However, we selected these scales on purpose, to emphasize the largest differences between methods, and not focus on minor differences in statistical metrics between methods.

- Line 241-242: The statement "When considering the H estimates as well, the overall performance of both EBM versions reduces" is unclear. Does this mean that when metrics are computed jointly for both H (sensible heat flux) and LvE (latent heat flux), the overall performance of the Energy Balance Methods (EBM) decreases? If so, how was the joint evaluation done?
  We refer here to the fact that for the sensible heat flux, the two-wavelength method performs better than the EBM. We agree that this was not clearly formulated. We replaced this sentence with:
  The H estimates of both EBM versions perform less well than the estimates of the two-wavelength method. This is in line with our expectations, because the LAS signal dominates these estimates.

- Line 245: "For the two-wavelength method, this overestimation is fully attributed to the LvE, since H is constrained by the LAS, while for the EBM, this overestimation can be distributed among LvE and H." while for the EBM, this overestimation can be distributed among LvE and H. The reasoning that overestimation is "fully attributed to the LvE" in the two-wavelength method, due to LAS constraining H, is understood. However, this relies on the implicit assumption that H is accurately captured by the LAS. Could the authors clarify whether this assumption holds true in this context?
  We are not sure what the reviewer means here. The overestimation we are referring to here is in comparison to the two-wavelength method with the MWS and LAS (and EC). So, the LAS estimates for the reference and the CML are exactly the same. This implies that all differences between the reference and the CML can be attributed to the L_vE estimates. For clarification, we added as follows:
  For the two-wavelength method, this overestimation is fully attributed to the L_vE, since H is constrained by the same LAS estimates for the CML-2λ and the MWS-2λ methods, while for the EBM methods this overestimation can be distributed among L_vE and H.

  Moreover, as can be seen in Fig. C1, the differences between the *H* estimates of the EC and the two-wavelength method with the MWS and LAS show roughly the same statistical metrics as the L_vE comparison.

- The discussion and conclusion sections feel somewhat verbose. Please streamline the content to avoid redundancy.
  Based on your comments and the comments of other reviewers, we have adapted the discussion and conclusion sections, and we think the content in these sections is now less redundant.

- The manuscript lacks a discussion on how differences in footprints, sampling frequencies, and spatial/temporal representativeness among the various systems, such as eddy covariance (EC), microwave scintillometer (MWS-LAS), and commercial microwave links (CMLs), might affect the flux estimates. While a detailed intercomparison may be beyond the scope of this study, a short paragraph acknowledging these differences and citing relevant literature would greatly help readers interpret discrepancies in the time series and understand the strengths and limitations of each measurement system.
  We understand the comment of the reviewer. However, originally the site of Cabauw was selected given its relatively homogeneous terrain, especially in the dominant south-westerly wind direction. Also, a comparison by our reference instruments illustrates the similar behaviour between the EC and the MWS-LAS (Fig. 3). Moreover, the footprints of the CML and MWS are nearly identical, given their collocated installation. To emphasize this, we added in Section 3.2 the following:
  As reference instruments, we use an optical-microwave scintillometer setup and an eddy-covariance system (EC). The scintillometer setup consists of a Radiometer Physics RPG-MWSC-160 microwave scintillometer, transmitting at 160.8 GHz (i.e., a wavelength of 1.86 mm), and a Kipp & Zonen LAS Mk-II, transmitting at 352.7 THz (i.e., a wavelength of 850 nm), which are both sampled at 1 kHz. Hereafter we refer to this system as MWS-2λ. The path of this scintillometer setup is nearly identical to the path of the CML (e.g., Fig. 2), so that differences in footprints can be neglected. It should be noted that this setup has a major data gap between 10 July and 6 September 2024. The EC system consists of a sonic anemometer (Gill-R50) and an open-path H_2O/CO_2 sensor (LICOR-7500), sampled at 10 Hz, and installed at

3 meter above the ground (Bosveld et al., 2020). The fluxes obtained with the EC can be directly obtained from the KNMI Data Platform, next to other more common meteorological measurements. The EC data is available during the entire study period. During the dominant south-westerly wind direction, the footprints of the EC, scintillometer setup and CML all predominantly consist of grass fields.

And in Section 4.1:

A comparison of the reference methods for the full data period shows an overall comparable behaviour (Fig. 3), indicating negligible influence of differences in footprints between the two references. Even though on average, the MWS-2λ shows an overestimation in comparison to the EC. In general, the uncertainty of the CML methods observed in two example timeseries does not deviate largely from the intercomparison between reference methods (Fig. 4).

- Additional minor thought: While comparing to the EC tower's available energy (R_n – G), it would be very helpful if the authors could plot the sum of H + LvE derived from the CML data vs EC tower H+LvE. Although the footprints and instruments are not perfectly collocated, such a comparison would provide valuable insight into the total energy captured by the CML-based approach. This could also be benchmarked against typical energy closure rates observed at flux towers (approximately 80% on clear days), offering an overview of the method's overall performance.

  We understand the reasoning of the reviewer, and agree that it is indeed relevant to know for the reader how well the energy balance closes, to get an idea of the performance of the EBMs. However, we do think that the proposed analysis is out of scope for this paper, also because there are multiple studies performed to this (e.g., Kroon, 2004, which is in Dutch, but the main findings are reported also in this triennial report by KNMI in English). We therefore added as follows in Section 2.2 (also based on other reviewer's comments):

  For the in-situ method, we use the measured ground heat flux, while for the remotely sensed data product we assume 10 % of the net radiation is used for the ground heat flux. Note that the measured energy balance hardly ever closes, especially in more complex measurement environments, e.g., forests or cities (Mauder et al., 2020). For the field site used in this study, Cabauw in the Netherlands, typically an imbalance during day-time is found between 10% (afternoons) to 40% (mornings) (Kroon, 2004). For our data period we find similar values using EC data (not shown).