



A dropsonde-based machine-learning framework for estimating marine atmospheric boundary layer height around the Korean Peninsula

Min-Seong Kim and Tae-Young Goo

Observation Research Department, National Institute of Meteorological Sciences, 33, Seohobuk-ro,
Seogwipo 63568, Jeju-do, Republic of Korea

Correspondence to: Min-Seong Kim (givemeasong@korea.kr)

Abstract. This study presents an observation-based machine-learning framework for estimating marine atmospheric boundary layer (MABL) height around the Korean Peninsula using aircraft dropsondes collected during 2020–2024. Of 1193 profiles, 1179 were collocated with ERA5 and 666 had well-defined MABL tops assigned through expert interpretation of thermodynamic and moisture profiles; 642 complete cases were used for machine learning. Dropsonde observations and ERA5 sea surface temperature were combined to estimate air–sea heat fluxes using a COARE 3.0-based formulation. Selected profile-based diagnostics and ERA5 planetary boundary layer height (PBLH) were evaluated against the expert-interpreted reference. Five machine-learning algorithms were trained using all thermodynamic regimes together and evaluated by five-fold GroupKFold cross-validation grouped by observation date. For the same 642 profiles, the machine-learning models reduced mean absolute error by 38.3–41.0 % relative to ERA5 PBLH. Random Forest, XGBoost, and a deep neural network yielded mean absolute errors of 155.5–155.9 m; the neural network had the lowest root-mean-square error (209.5 m) and the highest correlation (0.815). Random Forest permutation importance identified near-surface relative humidity and relative humidity at 925 hPa as the leading predictors. Moisture-related variables remained important across three thermodynamic regimes defined using 850 hPa temperature and lower-tropospheric stability. The proposed framework thus reproduced the expert-interpreted MABL height reference more closely than ERA5 PBLH for well-defined cases. The expert heights are physically informed references, not independent direct measurements of turbulent boundary layer depth.

1 Introduction

The marine atmospheric boundary layer (MABL) is a region in which heat, moisture, and momentum are actively exchanged between the sea surface and lower atmosphere. It plays an important role in understanding air–sea interactions and the development of marine weather phenomena. In particular, MABL height is a key variable representing the depth of turbulent mixing and the vertical extent of dispersion in the lower atmosphere. Therefore, it is essential for marine weather prediction, air quality and dispersion analyses, and the evaluation of boundary layer simulations in numerical models.



30 However, continuous, direct observations of MABL height over the ocean remain challenging. In situ observing platforms, such as aircraft, ships, buoys, and radiosondes, are constrained by high operating costs and limited spatiotemporal representativeness. Previous aircraft-based dropsonde studies have mainly focused on specific high-impact weather events or intensive observation campaigns, such as tropical cyclones, atmospheric rivers, and winter storms (Intrieri et al., 2014; Feng and Wang, 2019; Zheng et al., 2024). These studies provide important information for understanding the vertical structure and forecasting sensitivity of individual phenomena; however, they may have limitations in representing the general marine boundary layer structure under diverse seasonal and regional oceanic conditions.

To complement these observational gaps, reanalysis datasets such as ERA5 have been widely used. In addition, established profile-based diagnostics, such as the Heffter inversion criterion, the bulk Richardson number method, and temperature-increment threshold approaches, have been applied to estimate boundary layer height (Heffter, 1980; Arya, 2001; Liu and Liang, 2010). These methods are physically interpretable but may identify different layers depending on the chosen diagnostic definition, vertical structure, and atmospheric stability (Seidel et al., 2010; Chen et al., 2023).

Based on reanalysis data and traditional diagnostic methods, recent studies have indicated that boundary layer height estimates may have substantial uncertainties under complex lower-atmospheric conditions. A boundary layer height evaluation using aircraft observations and ERA5 profiles showed uncertainties in ERA5-based interpretations of boundary layer structure (Carneiro et al., 2025). Additionally, an ocean-observation-based study showed that ERA5 planetary boundary layer height (PBLH) can be sensitive to errors in the representation of vertical thermal structure and wind profiles (Xi et al., 2024). Furthermore, because traditional diagnostic methods rely on thresholds based on potential temperature gradients, the Richardson number, or atmospheric stability, they may incorrectly identify an upper transition layer or residual layer as the MABL top in marine boundary layers with multiple inversion layers, residual layers, or cloud-induced decoupling (Ralph et al., 2024). These uncertainties underscore the need to establish a reliable MABL reference dataset based on the physical structure of the observed profiles, and then evaluate how well existing diagnostic methods and reanalysis PBLH agree with it.

To overcome these limitations, increasing efforts have been made in atmospheric science to estimate or automatically diagnose the boundary layer height using machine learning (ML) (Rieutord et al., 2021; Peng et al., 2023). Machine learning can learn nonlinear relationships by simultaneously considering multiple meteorological variables. In addition, explainable artificial intelligence (XAI) or variable-importance analysis can improve the physical interpretability of model predictions by identifying key variables that contribute to the predictions (Silva and Keller, 2024). XAI techniques have also been widely used to interpret the major predictors in climate and environmental datasets (Li et al., 2024; Ozupek et al., 2025). Therefore, using expert-interpreted MABL heights, derived from a comprehensive examination of the thermodynamic structure of vertical profiles, as target values for supervised learning provides a practical approach for representing complex MABL structures that cannot be captured using a single diagnostic equation.

However, the performance of the ML models strongly depends on the quality of the reference data used for training. If the outputs from traditional diagnostic equations or a single automatic detection method are used directly as training targets, the



threshold dependence, possible misdetections, and structural limitations of these methods may be inherited by the ML model.

65 Therefore, the construction of a consistent reference dataset based on the physical structure of the observed profiles should precede the development of MABL height estimation models (Krishnamurthy et al., 2021; Guo et al., 2024). From this perspective, rather than aiming to automatically assign a single MABL height to all marine boundary layer cases, including complex multilayer structures, this study focused on well-defined cases for which a single MABL top could be clearly identified. An expert-interpreted MABL-height reference dataset was constructed for these cases and used as a common
70 reference to evaluate traditional diagnostic methods, ERA5 PBLH, and ML models.

Even when expert-interpreted data are used, uncertainty associated with expert judgment may remain. Therefore, the internal consistency of the reference dataset must be quantitatively evaluated through repeated interpretations, such as blind re-reading.

The validation strategy for ML models is also important. Multiple dropsonde profiles collected on the same observation date
75 may share similar synoptic and oceanic conditions as well as spatial characteristics. If such profiles are included simultaneously in the training and validation datasets through random cross-validation, the resulting model performance may be overly optimistic and may not accurately reflect the model's actual generalization performance. Therefore, for ML analyses using multiyear aircraft observations, a validation strategy that accounts for data dependence by observation date or flight unit is required, rather than the random splitting of individual profiles.

80 MABL height can also vary depending on the thermodynamic environment in the lower atmosphere. In particular, the 850 hPa temperature (T850) is commonly used as an indicator of lower-atmospheric air-mass characteristics, while lower-tropospheric stability (LTS) represents the stability difference between the near-surface air and the lower free troposphere. LTS is generally defined as the difference between the potential temperature at 700 hPa and the near-surface potential temperature, and has been widely used in studies on marine low clouds and lower-atmospheric stability (Klein and Hartmann,
85 1993; Wood and Bretherton, 2006). In addition, thermodynamic regime classification using T850 and LTS can be used to interpret how the boundary layer structure, cloud, precipitation, and turbulence characteristics vary across different lower-atmospheric environments (Giangrande et al., 2020; Blount et al., 2023).

Against this background, this study developed a framework for MABL height estimation using dropsonde observations collected by a meteorological aircraft over the seas around the Korean Peninsula during 2020–2024. The study regions
90 include the Yellow Sea, northern East China Sea, and the East Sea. These seas are characterized by diverse seasonal air–sea interactions, including cold air outbreaks in winter, warm and humid maritime air masses in summer, and moisture supplies associated with the Changma rainy season, Korean summer monsoon, and typhoons. Given this, the dropsonde dataset used in this study spans multiple years and is not limited to a single weather phenomenon or intensive observation campaign, enabling the evaluation of MABL structures under various seasonal and regional oceanic conditions.

95 Previous ML-based studies on boundary layer height have often focused on lidar, ground-based remote sensing, or reanalysis-based PBLH. Studies that use expert-interpreted MABL heights from multiyear aircraft dropsonde vertical profiles as target values remain limited. To address this gap, this study presents a unified framework that combines the



construction of an observation-based MABL reference dataset, the evaluation of traditional diagnostic methods and ERA5
PBLH, and observation-date-based ML validation using multiyear dropsonde observations collected around the Korean
100 Peninsula.

The specific objectives of this study are as follows: First, we constructed an expert-interpreted MABL height reference
dataset using five years of aircraft dropsonde observations and evaluated the internal consistency of the interpretation
through blind re-reading. Second, we compared dropsonde-input surface heat-flux estimates calculated using a COARE 3.0-
based bulk formulation and near-surface meteorological variables with their ERA5 counterparts, and quantitatively evaluated
105 how well traditional MABL diagnostic methods and ERA5 PBLH agreed with the expert-interpreted MABL height reference.
Third, we applied observation-date-based GroupKFold cross-validation to evaluate how well the out-of-fold predictions of a
pooled global model trained on the combined profiles from all three thermodynamic regimes agreed with the expert-
interpreted MABL reference, and compared their performance with that of ERA5 PBLH for the same cases. Fourth, we used
T850–LTS-based thermodynamic regimes to interpret the differences in MABL structure and the regime-dependent
110 importance of predictors in the global model. Finally, we examined the key predictors used by the model for MABL height
estimation and assessed the limitations of its physical interpretation using Random Forest permutation importance analysis.

2 Data and Methods

2.1 Meteorological aircraft dropsonde observations

This study used dropsonde observations released by the Korea Meteorological Administration/National Institute of
115 Meteorological Sciences (KMA/NIMS) Atmospheric Research Aircraft (NARA) over the seas surrounding the Korean
Peninsula during 2020–2024. The aircraft was based at Gimpo Airport and conducted observations within the Korea Air
Defense Identification Zone (KADIZ), mainly over the Yellow Sea, the northern East China Sea, and the East Sea. As noted
in Sect. 1, these seas exhibit diverse seasonal air–sea interactions. The operational characteristics of the KMA/NIMS
Atmospheric Research Aircraft and airborne observations around the Korean Peninsula have been described in previous
120 studies (Lee et al., 2011; Kim et al., 2023a).

The Vaisala RD-94 dropsonde used in this study is a high-resolution in situ instrument that measures air temperature,
humidity, pressure, GPS-based position, and vertical profiles of wind direction and speed. After being released from the
aircraft, the dropsonde descends toward the sea surface under a parachute and continuously observes the thermodynamic and
kinematic vertical structures of the atmosphere. The data received during the descent were collected using the Airborne
125 Vertical Atmospheric Profiling System (AVAPS; Hock and Franklin, 1999) and subsequently processed through standard
quality control and correction procedures using the Atmospheric Sounding Processing Environment (ASPEN, version 3.3-
668).

During the analysis period, 1193 dropsondes were released. Based on these raw dropsonde data, this study performed
spatiotemporal collocation with ERA5 reanalysis data, examined the completeness of the lowest-level observations,



130 constructed an expert-interpreted MABL height reference dataset, and prepared an input dataset for machine-learning analysis. The data screening procedure and sample sizes for each analysis are summarized in Sect. 2.2 and Table 1.

2.2 Data preprocessing and spatiotemporal collocation with ERA5 reanalysis

135 The dropsonde vertical profiles underwent a stepwise quality control procedure to minimize the effects of communication errors, sensor malfunctions, missing-value flags, and physically unrealistic observations. First, observations flagged as missing values with -999.0 were removed, and physically unrealistic extremes in the lower layer used for MABL analysis were excluded. In this study, observations were considered abnormal when wind speed exceeded 60 m s^{-1} , air temperature exceeded $50 \text{ }^{\circ}\text{C}$ or was lower than $-30 \text{ }^{\circ}\text{C}$, or relative humidity exceeded 100% . The lowest-level observations were
140 screened to determine whether valid observations were available within the 2–30 m layer for surface flux calculations and the evaluation of lower-atmospheric variables.

ERA5 reanalysis data were used after spatiotemporal collocation with dropsonde observations (Hersbach et al., 2020). Hourly ERA5 data, with a horizontal resolution of $0.25^{\circ} \times 0.25^{\circ}$, were used. The representative location and time of each dropsonde profile were defined by the release latitude, longitude, and time. For each profile, the ERA5 grid point and hourly
145 field corresponding to the release location and time were extracted. This nearest-neighbor matching was applied not to reconstruct ERA5 fields along the dropsonde descent path but to compare the ERA5 grid-scale lower-atmospheric state, surface fluxes, and boundary layer diagnostics at the time of dropsonde release.

From ERA5 single-level data, sea surface temperature (SST), 2 m air temperature (t2m), 2 m dew-point temperature (d2m), 10 m zonal and meridional wind components (u10 and v10), surface pressure (sp), mean surface sensible heat flux (msshf),
150 mean surface latent heat flux (mslhf), and boundary layer height (blh) were extracted. The t2m and d2m variables were used to calculate the ERA5 2 m relative humidity, and u10 and v10 were used to calculate the ERA5 10 m wind speed. The derived ERA5 2 m relative humidity and 10 m wind speed were compared with lower-level dropsonde observations, and the ERA5 SST was compared with the buoy SST. To assess the influence of the lower-atmospheric input data source, the same COARE 3.0-based iterative bulk-flux formulation described in Sect. 2.3 was applied to ERA5 meteorological variables.
155 ERA5 SST, surface pressure, 10 m wind speed derived from u10 and v10, 2 m air temperature, and 2 m relative humidity derived from t2m and d2m were used in this calculation. Wind speed, air temperature, and relative humidity were applied at their ERA5 reference heights of 10, 2, and 2 m, respectively ($z_u=10 \text{ m}$, $z_t=2 \text{ m}$, and $z_q=2 \text{ m}$), without vertical interpolation or extrapolation. The ERA5-provided msshf and mslhf were time-mean surface fluxes in W m^{-2} , and were therefore used without additional conversion from time-accumulated quantities. The native ERA5 fluxes were directly compared with the
160 dropsonde-input bulk-flux estimates. Because the native ERA5 vertical fluxes follow a downward-positive sign convention, their signs were reversed when they were compared with the dropsonde-input bulk-flux estimates, such that upward heat flux from the ocean to the lower atmosphere was positive. In this study, the blh variable was defined as the ERA5 planetary



boundary layer height (PBLH). It was compared with the expert-interpreted MABL height. ERA5 PBLH was not included as an input variable in the machine-learning models.

165 To evaluate the representation of ERA5 SST, marine meteorological buoy observations from the KMA Open MET Data Portal were used independently of the dropsonde profile dataset. These buoy SST data served as independent auxiliary validation data for the SST comparison (as shown in Fig. 4) and were not directly paired with the number of dropsonde profiles. The buoy SST dataset consisted of hourly SST observations from buoy stations surrounding the Korean Peninsula during 2020–2024. To be consistent with the seasonal and diurnal comparison (as shown in Fig. 4), only data collected
170 between 07:00 and 15:00 Korea Standard Time (KST), corresponding to the dropsonde observation period, were used. For each dropsonde observation time window, hourly SST observations were matched with the SST from the nearest ERA5 grid point to construct buoy–ERA5 SST comparison pairs. After pairwise validity screening and the contextual SST quality-control procedure, a total of 5412 valid buoy–ERA5 SST comparison pairs remained. These 5412 pairs do not correspond directly to the number of dropsonde profiles, but represent the station-hourly buoy–ERA5 SST validation dataset used only
175 for the SST comparison in Fig. 4. Buoy SST data were not used for the MABL height interpretation, evaluation of traditional diagnostic methods, or machine-learning model training and validation.

Among the 1193 raw dropsonde profiles selected during the analysis period, 14 were excluded because spatiotemporal collocation with ERA5 was not possible due to missing time information or GPS coordinates. As a result, 1179 profiles were matched one-to-one with ERA5. Subsequently, profiles without observations in the 2–30 m layer, mainly because of
180 communication loss immediately before reaching the sea surface, were excluded to ensure reliable near-surface atmospheric information for evaluating ERA5 surface fluxes and lower-atmospheric thermodynamic variables. Accordingly, 1138 lower-level-complete profiles were retained for evaluation.

MABL height interpretation was performed for 1179 profiles that were spatiotemporally collocated with ERA5. Based on a comprehensive examination of the vertical thermodynamic structure and auxiliary variables, expert-interpreted MABL
185 heights were assigned to 666 profiles, for which a single MABL top could be clearly defined. This dataset was used to evaluate the error characteristics of traditional MABL diagnostic methods and the ERA5 PBLH. In contrast, cases for which a single MABL top could not be robustly identified because of multiple inversion layers, strong transition layers, decoupling, or inconsistencies among the major thermodynamic variables were excluded from the expert-interpreted MABL height reference dataset. Therefore, the subsequent evaluation of traditional diagnostic methods and machine-learning models
190 should be interpreted as applying only to cases in which a single MABL top was relatively well-defined, rather than to all dropsonde cases. Among these profiles, 642 profiles for which both expert-interpreted MABL height and all 11 input variables were available were used for the K-means-based thermodynamic regime classification, observation-date-based machine-learning model training and validation, and variable-importance analysis. The data subsets and final sample sizes for each analysis are summarized in Table 1.

195 **Table 1. Data subsets used in this study and their main applications.**



Dataset	Selection criterion	N	Main use
Raw dropsonde profiles	Dropsonde profiles collected around the Korean Peninsula during 2020–2024	1,193	Initial dataset
ERA5-collocated profiles	Profiles with valid time stamps and GPS coordinates that could be collocated with ERA5	1,179	Base profile dataset
Lower-level-complete profiles	Profiles with valid near-surface observations within 2–30 m above sea level	1,138	ERA5 surface flux and lower atmospheric variable evaluation
Expert-interpreted MABL profiles	Profiles for which an expert-interpreted MABL height could be assigned	666	Evaluation of traditional MABL diagnostics and ERA5 PBLH
ML-complete profiles	Profiles with both expert-interpreted MABL height and all 11 input variables	642	Thermodynamic-regime classification, date-grouped global ML training and validation, and XAI analysis
Blind re-reading subset	Selected from expert-interpreted profiles with valid T850 and LTS, using stratified random sampling of 30 profiles from each T850–LTS-based thermodynamic regime	90	Intra-analyst repeatability assessment of expert-interpreted MABL heights

Abbreviations: N, number of profiles; ERA5, fifth-generation ECMWF atmospheric reanalysis; GPS, Global Positioning System; MABL, marine atmospheric boundary layer; PBLH, planetary boundary layer height; ML, machine learning; T850, temperature at 850 hPa; LTS, lower-tropospheric stability; XAI, explainable artificial intelligence.

200

2.3 Observation-based surface flux calculation

The sensible heat flux (SHF) and latent heat flux (LHF) used in this study were not directly measured using the eddy covariance method. Instead, they were observation-based reference fluxes calculated by combining the lowest-level dropsonde observations with ERA5 SST data. Because direct measurements of high-resolution turbulent fluxes over the ocean are practically limited, this study defined the near-surface atmospheric state using dropsonde profiles with valid lowest-level observations within the 2–30 m layer, and combined these observations with ERA5 SST to calculate surface fluxes. This dropsonde-based surface flux calculation approach is consistent with previous studies that estimated air–sea heat fluxes using bulk methods by combining GPS dropsonde- or aircraft-based atmospheric observations with SST data (Zhang et al., 2017; Jaimes de la Cruz et al., 2021; Kim et al., 2023b).

Surface turbulent heat fluxes were estimated using an iterative bulk-flux formulation based on the core surface-layer framework of the Coupled Ocean–Atmosphere Response Experiment (COARE) 3.0 and Monin–Obukhov similarity theory (MOST; Fairall et al., 2003). The input variables were the lowest-level air temperature, humidity, and wind speed observed by the dropsonde, together with the ERA5 SST. In the bulk-flux calculation, SHF, LHF, friction velocity, and Monin–



Obukhov length were iteratively calculated by considering air–sea temperature and humidity differences, wind speed,
215 surface roughness, and stability-dependent transfer coefficients.

The SHF and LHF values calculated in this study were used for two purposes. First, they were used as reference data to
evaluate the representation of the surface sensible and latent heat fluxes provided by ERA5. Second, they were used as input
variables in the MABL height estimation machine-learning models, representing the thermal and moisture forcing from the
sea surface to the lower atmosphere. Therefore, the surface fluxes presented in this study should be interpreted not as directly
220 observed turbulent fluxes, but as observation-based reference values calculated from lowest-level dropsonde observations
and ERA5 SST.

2.4 Vertical-profile-based MABL height interpretation and reference dataset construction

Because marine atmospheric boundary layers frequently exhibit multiple inversion layers, residual layers, and cloud-induced
225 decoupling, it is difficult to robustly identify the MABL top using only a single-threshold-based automatic diagnostic
method. Therefore, in this study, MABL heights were manually interpreted by an expert analyst for the 1179 dropsonde
profiles that were spatiotemporally collocated with ERA5, based on a comprehensive examination of the thermodynamic
structure of the vertical profiles and auxiliary variables.

MABL heights were interpreted primarily from coherent transitions in the virtual potential temperature (θ_v) and specific
230 humidity (q) profiles. In particular, the analyst examined where a relatively uniform lower-layer θ_v profile began to increase
and where q decreased across the transition to the overlying atmosphere. Relative humidity (RH) was used to characterize
the corresponding moisture transition, while wind speed and direction and surface-flux information were examined as
supplementary context; no fixed wind or wind-shear threshold was imposed. The analyst reviewed the combined profiles
using an in-house graphical interface rather than applying a single automatic threshold. Figure 1 illustrates an example from
235 13 November 2024, for which an expert-interpreted MABL height of 599 m was assigned. The example shows a virtual-
potential-temperature increase and a pronounced drying transition near the interpreted height; it is intended to illustrate the
interpretation procedure, not to establish an independently measured turbulent mixing height.

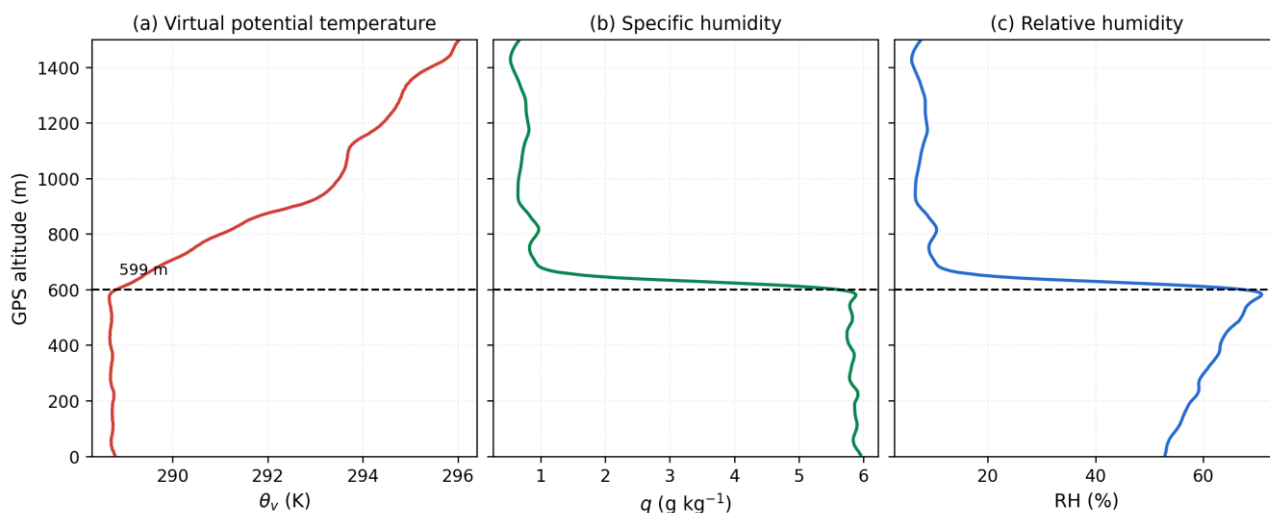


Figure 1. Example of an expert-interpreted marine atmospheric boundary layer (MABL) height from a dropsonde released on 13 November 2024. Vertical profiles of (a) virtual potential temperature (θ_v), (b) specific humidity (q), and (c) relative humidity (RH) are shown using the observed GPS altitude. The dashed line marks the expert-interpreted MABL height (599 m), identified primarily from the coincident thermodynamic and moisture transition. The height was visually interpreted and is not an independent turbulence-based measurement.

Among the 1179 profiles spatiotemporally collocated with ERA5, cases exhibiting transitional structures or for which a single MABL top could not be robustly defined were excluded from the expert-interpreted MABL height dataset. These cases included profiles with distinct multiple inversion layers, in which the lower MABL could not be clearly separated from an upper residual layer; profiles with cloud-induced decoupling, which made a single boundary layer height difficult to define; and profiles in which the interpretation criteria based on major thermodynamic variables (such as θ_v , q , and RH) were inconsistent. Following this review process, a reference dataset comprising 666 profiles with valid expert-interpreted MABL heights was constructed.

Subsequently, the key variables required to calculate the machine-learning input variables were screened. Among the 666 profiles, 642 profiles, for which all 11 input variables were available, were defined as the ML-complete dataset. Therefore, 666 expert-interpreted MABL profiles were used to evaluate the error characteristics of traditional MABL diagnostic methods, including the Heffter method, the bulk Richardson number method, and the θ_v threshold (0.5 K) method, as well as ERA5 PBLH. Among these, 642 ML-complete profiles were used for the K-means-based thermodynamic regime classification, observation-date-based global machine-learning model training and validation, and variable-importance analysis.

For comparison with the expert-interpreted reference, we selected three complementary profile-based diagnostic baselines representing distinct physical criteria: potential-temperature inversion structure (Heffter), the balance between buoyancy and vertical wind shear (bulk Richardson number), and a fixed virtual-potential-temperature increment (θ_v threshold, 0.5 K). The



Heffter-type approach identified the base of the first inversion layer satisfying a potential-temperature gradient of at least 5 K km⁻¹ and a cumulative potential-temperature increase of at least 2 K. The bulk Richardson number was calculated from the observed virtual potential temperature and zonal and meridional wind components relative to the lowest jointly valid reference level. Its diagnosed height was the first level where the bulk Richardson number reached 0.25.

265 The third baseline, termed the θ_v threshold (0.5 K) method, identified the first altitude z at which $\theta_v(z) - \theta_v(z_0) \geq 0.5$ K, with z_0 denoting the lowest valid reference level. The temperature-increment screening concept and the 0.5 K threshold are related to Liu and Liang (2010), but that study used the 0.5 K threshold for land and 0.1 K over oceans and included stability-dependent and additional gradient-based procedures. Our implementation instead used virtual potential temperature and a single fixed threshold, without reproducing the full Liu and Liang algorithm. The threshold was retained here as a simple
270 diagnostic benchmark, not as a validated ocean-specific optimum.

Other established approaches include parcel-based methods and maximum or minimum gradients of temperature, potential temperature, relative humidity, and specific humidity (Seidel et al., 2010; Chen et al., 2023). Such methods need not identify the same physical boundary layer feature, and their estimates can depend on stability, vertical resolution, and the profile's multilayer structure. We therefore use the selected diagnostics as interpretable benchmarks rather than claiming an
275 exhaustive intercomparison of PBL-height retrieval methods.

To evaluate the internal consistency of the expert-interpreted MABL heights, blind re-reading was performed approximately three months after the initial interpretation. Ninety profiles for rereading were selected from the 666 expert-interpreted MABL profiles with valid T850 and LTS values, for which T850–LTS-based K-means regime labels could be assigned. To prevent the rereading subset from being biased toward a particular thermodynamic environment, 30 profiles were randomly
280 sampled from each of the warm-stable, transitional, and cold/weakly stable regimes using a random seed of 42. Regime information was used only as a stratification criterion for selecting the rereading subset and was not provided to the analyst during the rereading process.

During the rereading process, the analyst did not have access to the initial expert-interpreted MABL heights, diagnostic estimates from the Heffter method, bulk Richardson number method, the θ_v threshold (0.5 K) method, ERA5 PBLH, thermodynamic regime information, T850, LTS, or other auxiliary diagnostic information. In addition, no MABL-height
285 lines or numerical labels were shown in the vertical-profile plots prepared for rereading. The same expert analyst independently reinterpreted the MABL height by re-examining the vertical profiles of θ_v , q , RH, wind speed, and wind direction. The re-read values were compared with the initial expert-interpreted values using the correlation coefficient, mean bias, root-mean-square error (RMSE), mean absolute error (MAE), median absolute difference, and agreement rates within
290 50, 100, and 150 m. This procedure was designed to evaluate the repeatability of valid MABL height values and did not include an assessment of the repeatability of classifying profiles for which a single MABL top could not be assigned as “No MABL.”

The expert-interpreted MABL height dataset used in this study was constructed primarily from profiles in which a single MABL top could be clearly defined. Therefore, the applicability of this reference dataset may be limited for profiles in



295 which a single MABL top cannot be robustly assigned because of strong transition layers, multiple inversion layers, upper
residual layers, or complex decoupling structures. The performance of the proposed MABL height estimation framework
was evaluated for these well-defined MABL cases, and its extension to complex multilayer cases requires further
examination using independent observations and additional interpretation criteria. In addition, the expert-interpreted MABL
heights used in this study should be regarded as physically based and consistent references assigned through a
300 comprehensive examination of vertical thermodynamic and kinematic structures, rather than as absolute truths. Therefore, all
subsequent comparisons should be interpreted as relative evaluations of the degree to which each method agrees with the
reference dataset.

2.5 K-means-based thermodynamic regime classification

305 In this study, K-means clustering was performed to classify the thermodynamic environments of the MABL. T850 and LTS
derived from the dropsonde vertical profiles were used as input variables for the clustering. T850 represents the air-mass
characteristics of the lower free troposphere, whereas LTS represents the thermodynamic stability difference between the
near-surface air and the lower free troposphere.

T850 was calculated as the temperature corresponding to 850 hPa in the vertical pressure–temperature profile of each
dropsonde. When no observations were available at this pressure level, T850 was interpolated in log-pressure coordinates
310 using adjacent observed layers bracketed at 850 hPa. The LTS was computed as the difference between the potential
temperature at 700 hPa and the near-surface potential temperature. The 700 hPa potential temperature, θ_{700} , was calculated
using the temperature and pressure corresponding to 700 hPa in the pressure–temperature vertical profile of each dropsonde.
When no observations were available at 700 hPa, the corresponding value was interpolated in log-pressure coordinates using
315 adjacent observed layers. The near-surface potential temperature, θ_{surface} , was calculated from the temperature and pressure of
the lowest valid observation within the lower 2–30 m layer of each dropsonde profile; i.e., $\text{LTS} = \theta_{700} - \theta_{\text{surface}}$.

K-means clustering analysis was applied to the 642 ML-complete profiles for which both expert-interpreted MABL heights
and all 11 ML input variables were available. Before clustering, T850 and LTS were standardized using a z-score
transformation. Specifically, the mean of each variable was subtracted and the result was divided by its standard deviation.
320 Standardization placed T850 and LTS on comparable scales, preventing differences in their numerical ranges and variances
from disproportionately affecting the K-means distance calculation. The number of clusters was set to three based on both
statistical metrics and physical interpretability. Although the silhouette scores for $k = 2$ and $k = 3$ were similar, $k = 3$ was
used in this study because it provided three physically interpretable environments: warm-stable, transitional, and cold/weakly
stable. MABL height was not used as a clustering input; instead, it was used after clustering to evaluate the MABL-height
325 characteristics of each thermodynamic regime.

In this study, thermodynamic regime classification was not intended to train separate prediction models for each regime.
Instead, it was used as a diagnostic framework to interpret the physical differences in the MABL height and regime-
dependent predictor importance in the global model.



330 2.6 Machine-learning input variables and model construction

The machine-learning input variables for estimating the MABL height were constructed from dropsonde observations, ERA5 SST, and COARE 3.0-based surface fluxes calculated from the lower-level dropsonde observations and ERA5 SST. The input variables were selected to represent the combined effects of sea surface and lower-atmospheric thermal conditions, moisture state, lower-atmospheric stability, mechanical mixing, and surface heat and moisture exchange.

335 A total of 11 meteorological and flux variables were considered in this study. The input variables consisted of SST, lowest-level air temperature (T_a), lowest-level relative humidity (RH), lowest-level wind speed (U), T850, LTS, 925 hPa temperature (T925), 925 hPa relative humidity (RH925), wind-speed difference (ΔWS), SHF, and LHF.

SST was extracted from ERA5 single-level data. T_a , RH, and U were defined as the lowest valid observations within the lower 2–30 m layer of each dropsonde profile. SHF and LHF were calculated by applying the iterative bulk-flux formulation
340 described in Sect. 2.3 to these lower-level dropsonde observations and ERA5 SST. T925 and RH925 were derived from the vertical pressure–temperature–humidity profile of each dropsonde at 925 hPa. When no observations were available at this pressure level, the corresponding values were interpolated in log-pressure coordinates using adjacent observed layers bracketed at 925 hPa. T850 and LTS were calculated as described in Sect. 2.5.

The wind-speed difference was defined as the wind speed at 850 hPa minus the wind speed at the lowest valid observation
345 within the 2–30 m layer. The wind speed at 850 hPa was obtained from the dropsonde wind-speed profile using interpolation in log-pressure coordinates.

SST and T_a represent the thermal contrast between the sea surface and lower atmosphere, whereas RH and RH925 represent the moisture state near the sea surface and the lower free troposphere or upper part of the boundary layer, respectively. T850, T925, and LTS reflect the temperature structure and stability of the lower atmosphere. The lowest-level wind speed is
350 directly involved in the calculation of surface bulk fluxes and represents the strength of near-surface mechanical forcing, whereas the wind-speed difference represents the vertical contrast in wind-speed magnitude between the near-surface layer and 850 hPa. SHF and LHF were used as variables representing the thermal and moisture forcing from the sea surface to the lower atmosphere.

Each dropsonde profile constituted one machine-learning sample represented by an 11-variable predictor vector and one
355 scalar target: its expert-interpreted MABL height in metres. The full observed θ_v , q , RH, and wind profiles informed the expert interpretation but were not supplied directly to the machine-learning algorithms as profile arrays; instead, only the 11 summarized meteorological and flux variables listed above were used as predictors. ERA5 PBLH was not used as an input variable and was retained solely as a benchmark. For each algorithm, a pooled global model was trained using the combined profiles from all thermodynamic regimes within each cross-validation training fold; separate regime-specific models were
360 not trained.

Five machine-learning algorithms were applied in this study: support vector regression (SVR), random forest (RF), XGBoost, gradient boosting, and a deep neural network (DNN). MABL height estimation performance was evaluated by comparing



algorithms with different learning structures using the same dataset and validation conditions, rather than relying on a single specific algorithm.

365 The model settings were as follows: SVR used a radial basis function kernel with $C = 100$ and $\gamma = 0.01$. The RF used 100 decision trees with a maximum depth of 10. XGBoost used 100 estimators, a maximum depth of 4, and a learning rate of 0.05. Gradient boosting used 100 estimators with a maximum depth of 4. The DNN consisted of two hidden layers with 64 and 32 nodes, respectively. The maximum number of iterations was set to 1000, and early stopping was applied. The random-state parameter was set to 42 for K-means, Random Forest, XGBoost, Gradient Boosting, and DNN.

370 All model settings were fixed before the final observation-date-based cross-validation, and hyperparameters were not re-tuned using validation data from individual folds. The objective of this study was not to maximize the performance of individually optimized models but to compare representative machine-learning algorithms using the same input dataset and validation conditions.

375 **2.7 Machine-learning model validation, benchmark comparison, and variable-importance analysis**

The five machine-learning models described in Sect. 2.6 were used to predict the expert-interpreted MABL height. Thermodynamic regime information was used as a diagnostic criterion for interpreting MABL height characteristics and the regime-dependent predictor importance of the global model, rather than to train separate models for each regime.

Model performance was evaluated using five-fold GroupKFold cross-validation, in which the observation date was used as the grouping variable. The 642 ML-complete profiles were grouped into 179 observation-date groups based on KST dates. All dropsonde profiles collected on the same observation date were treated as one group and assigned to the same fold. Therefore, profiles observed on the same date were not simultaneously included in the training and validation datasets. This procedure was applied to reduce the possible overestimation of the model performance caused by similarities among profiles observed on the same flight date or under identical synoptic and oceanic conditions.

385 In each fold, approximately four-fifths of the observation-date groups were used for training and the remaining approximately one-fifth were used for validation. The number of profiles in each validation fold ranged from 128 to 129. In each cross-validation fold, the 11 input variables were standardized using only the training data. Specifically, the mean and standard deviation of each input variable were calculated from the training data for each fold, and these values were then used to transform the training data. The same transformation was applied to the validation dataset. Thus, scaling-sensitive models such as SVR and DNN used inputs standardized solely from the training data; statistics from the validation data were not used in the standardization process. The validation predictions from all folds were combined to calculate the final out-of-fold performance. In other words, each profile was predicted once as unseen validation data and the resulting out-of-fold predictions were compared with the expert-interpreted MABL height. Model performance was evaluated using the correlation coefficient (R), RMSE, MAE, and mean bias (MB). MB was defined as the mean difference between the model prediction or ERA5 PBLH and the expert-interpreted MABL height. ERA5 PBLH was used as the benchmark. The machine-learning models and ERA5 PBLH were evaluated against expert-interpreted MABL heights using the same 642 ML-



complete profiles. In addition, the MAE reduction rate relative to the ERA5 PBLH was calculated for each machine-learning model to quantify the improvement in MABL height estimation over the reanalysis-based PBLH.

A permutation importance analysis was performed for the RF model to quantify predictor importance. For the RF model trained in each fold, each input variable in the validation data was randomly permuted one at a time, and the resulting increase in the MAE was calculated. A larger increase in MAE after permutation indicates greater importance of the corresponding variable for MABL height prediction. The fold-wise permutation-importance values were averaged across the five validation folds and interpreted as the overall variable importance of the global RF model.

In addition, within each fold, the validation data were grouped by thermodynamic regime, and permutation importance was calculated separately for each regime using the corresponding global RF model. This analysis did not involve the training of separate RF models for each regime. Instead, it was used to examine how predictor importance in the global RF model varied across thermodynamic environments. Therefore, the regime-specific permutation-importance results were not intended to represent the performance or structure of regime-specific models; rather, they were used as a supplementary diagnostic analysis to interpret how predictor importance in the global model varied with the thermodynamic environment.

To enhance the reproducibility of this study, Supplementary Table S1 provides the complete dataset used for model evaluation, including the 11 input variables, expert-interpreted MABL height, T850–LTS-based thermodynamic regime, ERA5 PBLH, cross-validation grouping information (such as `obs_date_kst` and `fold_id`), and the out-of-fold predictions from each machine-learning model for all 642 ML-complete profiles.

3 Results

This section presents, in sequence, the spatial distribution of the multiyear dropsonde observations around the Korean Peninsula, the representation of ERA5 lower-atmospheric variables and surface fluxes, the error characteristics of traditional diagnostic methods, ERA5 PBLH relative to the expert-interpreted MABL height, and the performance of machine-learning-based MABL height estimation. Because different quality-controlled data subsets were used for different analyses, the sample size for each result is specified in the corresponding section and in the relevant figure or table captions.

3.1 Dropsonde observation distribution and evaluation of ERA5 lower-atmospheric variables and surface fluxes

The spatial distribution of the 1193 dropsonde release locations and the KMA buoy stations used for ERA5 SST validation is shown in Fig. 2. The dropsonde observations covered the Yellow Sea, the northern East China Sea, and the East Sea during 2020–2024.

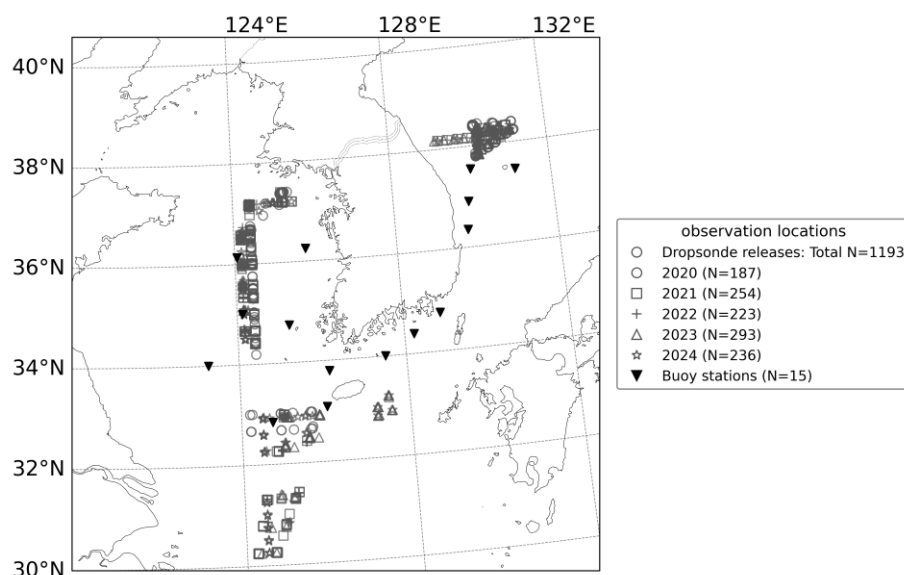


Figure 2. Spatial distribution of the 1193 dropsonde release locations around the Korean Peninsula during 2020–2024 and Korea Meteorological Administration (KMA) marine meteorological buoy stations used for ERA5 sea surface temperature (SST) validation. Open symbols denote the dropsonde observation years, as indicated in the legend, and filled inverted triangles denote the buoy stations used only for the SST comparison in Fig. 4.

Figure 3 compares dropsonde-input surface heat-flux estimates calculated using the COARE 3.0-based bulk formulation with native ERA5 surface fluxes for the 1138 lower-level-complete profiles that contained valid observations within the lower 2–30 m layer. For SHF, the correlation coefficient was 0.949, and the RMSE and MAE were 22.4 and 11.9 W m⁻², respectively. For LHF, the correlation coefficient was 0.918, and the RMSE and MAE were 54.7 and 34.5 W m⁻², respectively. Both fluxes showed high correlations, but the errors were larger for LHF than for SHF. This indicates that uncertainties in moisture-related variables may have a stronger influence on LHF estimates.

Figure 4 compares lower-atmospheric variables derived from the dropsonde observations and ERA5 across observation times and seasons. Comparisons of wind speed, air temperature, and relative humidity were based on 1138 lower-level-complete dropsonde profiles, whereas the SST comparison was based on 5412 quality-controlled buoy–ERA5 SST pairs. Overall, ERA5 reproduced seasonal and diurnal variability. However, lower-level wind speed was underestimated relative to the observations by 0.7–1.2 m s⁻¹ across all seasons and time periods, with the largest underestimation occurring during autumn. Air temperature biases were small (within ±0.7 °C) across all categories, with weak warm biases dominant in summer and winter afternoons. Relative humidity exhibited wet biases of approximately 4–7 % in spring and autumn, whereas the biases were small (within 2 %) in summer and winter afternoons. ERA5 SST agreed well with buoy observations across seasons and time periods, with biases within ±0.4 °C. The observed wet bias in relative humidity suggests that uncertainty may arise when only reanalysis-based lower-level humidity information is used for MABL height

estimation, and that dropsonde-based lower-level humidity information can provide complementary predictive information. In contrast, the strong agreement with ERA5 SST supports the use of ERA5 SST as an input for the bulk-flux calculation in this study.

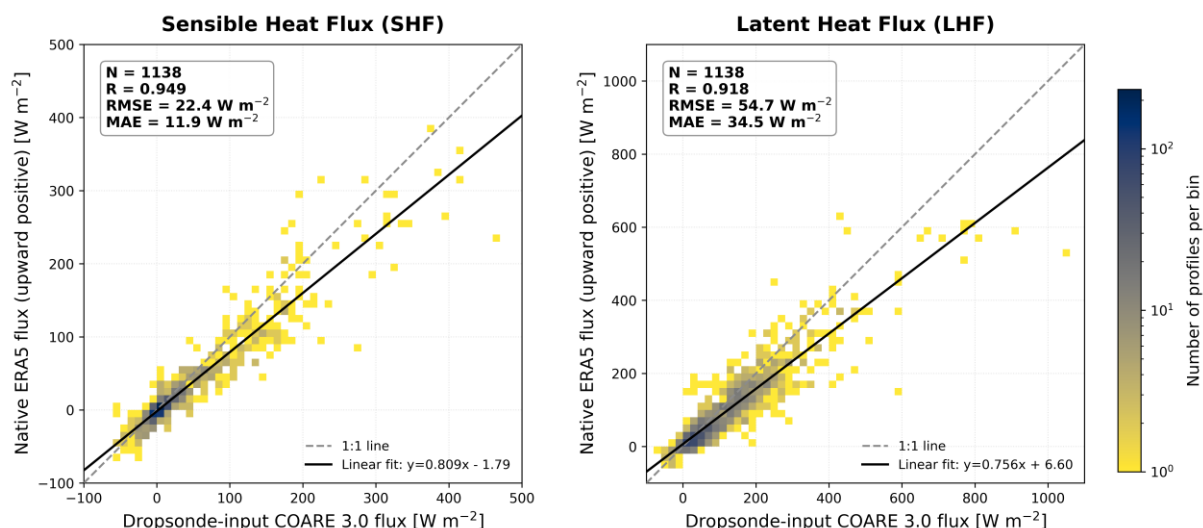


Figure 3. Two-dimensional frequency distributions comparing native ERA5 surface heat fluxes with dropsonde-input flux estimates calculated using the COARE 3.0-based bulk formulation for (a) sensible heat flux (SHF) and (b) latent heat flux (LHF). Signed flux values are presented using a common upward-positive convention. Colors indicate the number of profiles per bin on a logarithmic scale. The dashed and solid lines denote the 1:1 line and linear fit, respectively. Statistics in each panel denote N, R, RMSE, and MAE.

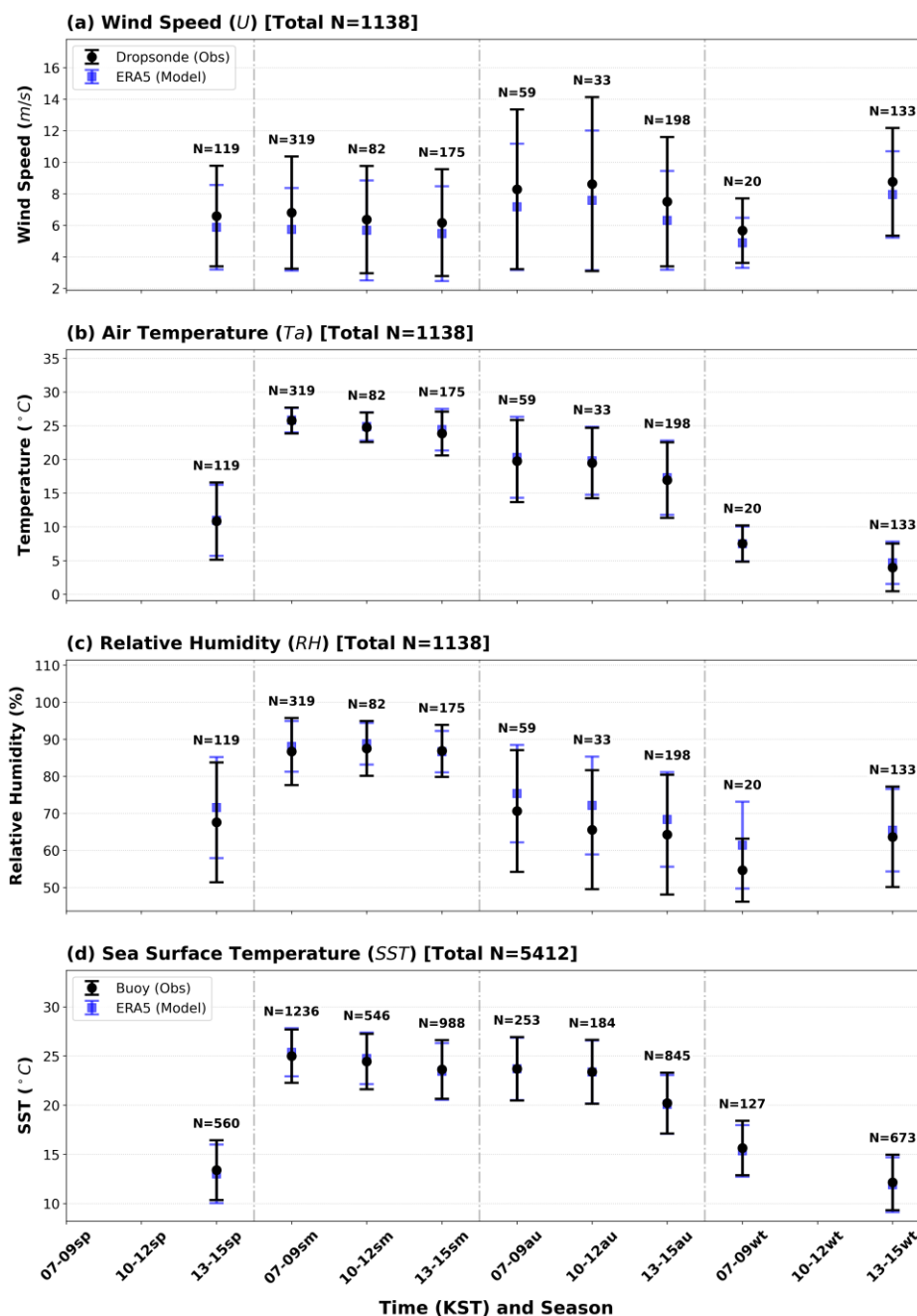


Figure 4. Seasonal and diurnal comparisons of lower-atmospheric variables from dropsonde observations and ERA5, including near-surface wind speed, air temperature, relative humidity, and SST. The wind-speed, air-temperature, and relative-humidity comparisons used 1138 lower-level-complete dropsonde profiles, whereas the SST comparison used 5412 quality-controlled buoy–ERA5 SST pairs. All comparisons were restricted to the daytime observation window of 07:00–15:00 Korea Standard Time (KST); spring observations were available only for 13:00–15:00 KST.



Figure 5 compares surface heat-flux estimates calculated using the same COARE 3.0-based iterative bulk-flux formulation with dropsonde and ERA5 lower-atmospheric inputs. For the dropsonde-input estimates, the lowest valid air temperature, relative humidity, wind speed, and pressure within the 2–30 m layer were applied at their observed height. For the ERA5-input estimates, 10 m wind speed and 2 m air temperature and relative humidity were applied at their respective reference heights without vertical interpolation or extrapolation. The ERA5-input SHF showed $R = 0.947$, $RMSE = 23.1 \text{ W m}^{-2}$, $MAE = 11.4 \text{ W m}^{-2}$, and a mean bias of -6.6 W m^{-2} . The ERA5-input LHF showed $R = 0.917$, $RMSE = 58.0 \text{ W m}^{-2}$, $MAE = 33.5 \text{ W m}^{-2}$, and a mean bias of -22.6 W m^{-2} . Although both fluxes showed high correlations, the larger error and negative bias for LHF indicate greater sensitivity to differences in the lower-atmospheric moisture inputs and air–sea humidity contrast. Because the same bulk-flux formulation and SST input were used for both estimates, this comparison primarily represents sensitivity to differences between the dropsonde and ERA5 lower-atmospheric inputs.

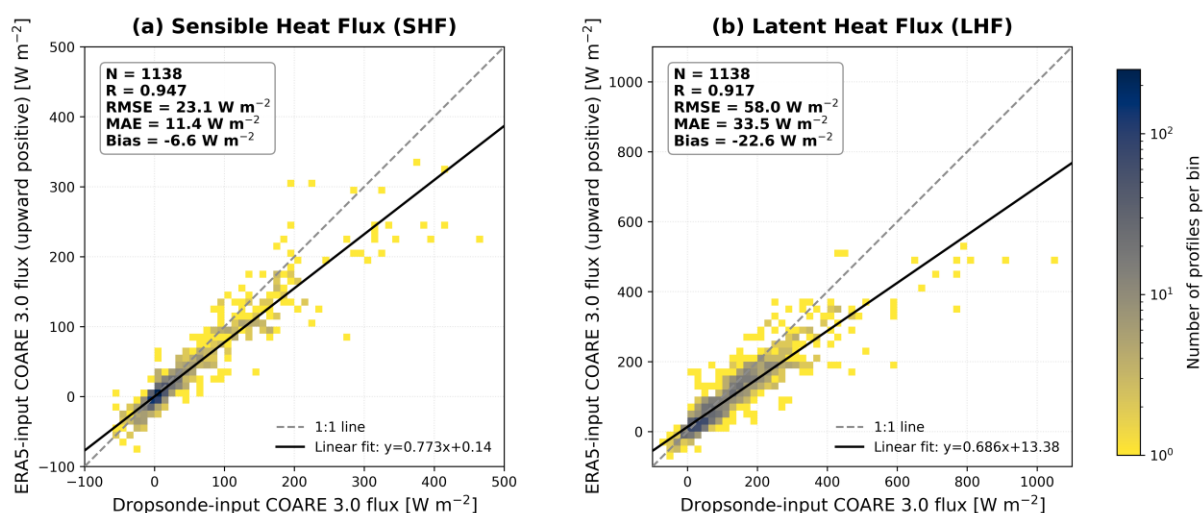


Figure 5. Two-dimensional frequency distributions comparing dropsonde-input and ERA5-input surface heat-flux estimates calculated using the same COARE 3.0-based iterative bulk-flux formulation for (a) sensible heat flux (SHF) and (b) latent heat flux (LHF). Signed flux values are presented using a common upward-positive convention. ERA5 10 m wind speed and 2 m air temperature and relative humidity were applied at their respective reference heights without vertical interpolation or extrapolation. Colors indicate the number of profiles per bin on a logarithmic scale. The dashed and solid lines denote the 1:1 line and linear fit, respectively. Statistics in each panel denote N, R, RMSE, MAE, and mean bias.

Overall, ERA5 reproduced the general variability of lower-atmospheric variables and surface heat fluxes over the seas surrounding the Korean Peninsula. However, relatively large differences were found in moisture-related variables and LHF, consistent with the greater sensitivity of LHF to differences in lower-atmospheric humidity inputs. These results support the inclusion of dropsonde-derived lower-level humidity structure and surface heat-flux estimates in the subsequent MABL height estimation, while ERA5 PBLH and traditional diagnostic methods were used as benchmarks.

3.2 Evaluation of traditional diagnostic methods and ERA5 PBLH against the expert-interpreted MABL height

Among the profiles spatiotemporally collocated with ERA5, 666 cases were assigned valid expert-interpreted MABL heights. Using this dataset as the reference, the error characteristics of the Heffter method, bulk Richardson number method, θ_v threshold (0.5 K) method, and ERA5 PBLH were evaluated. The expert-interpreted MABL heights were determined by comprehensively examining the vertical profiles of θ_v , q , RH, wind speed, and wind direction, and the resulting dataset was used as the common reference for evaluating the diagnostic methods.

Table 2 compares the quantitative performance of the traditional MABL diagnostic methods and the ERA5 PBLH against the expert-interpreted MABL height. The evaluation metrics used were R , RMSE, MAE, and MB. MB (defined in Sect. 2.7) is the diagnostic estimate minus the expert-interpreted MABL height; a positive MB indicates that the corresponding diagnostic method overestimated the MABL height relative to the expert-interpreted height. The Heffter method yielded valid MABL height estimates for 620 profiles, in which a height satisfying the potential-temperature inversion criterion was detected. The remaining 46 profiles, for which no level satisfied this criterion, were excluded from the quantitative evaluation.

Table 2. Performance of traditional MABL diagnostic methods and ERA5 PBLH against expert-interpreted MABL heights.

Method	Diagnostic basis	N	R	RMSE (m)	MAE (m)	MB (m)
Heffter (1980)	Potential temperature inversion	620	0.071	1650.5	1062.3	1055.3
Bulk Richardson (0.25)	Bulk Richardson threshold	666	0.651	328.7	196.4	−7.8
θ_v threshold (0.5 K)	Virtual potential temperature increase	666	0.850	287.6	205.0	176.1
ERA5 PBLH	Model-diagnosed PBL height	666	0.803	340.6	262.3	240.9

Note: Mean bias was calculated as the diagnostic estimate minus the expert-interpreted MABL height; positive values indicate overestimation. Heffter diagnostics were obtained for 620 profiles. In the remaining 46 profiles, no level satisfied the potential-temperature inversion criterion.

Among the traditional diagnostic methods, the Heffter method exhibited the largest errors in this dataset. Its correlation coefficient was very low at 0.071, and the RMSE and MAE were 1650.5 and 1062.3 m, respectively. The MB was 1055.3 m, indicating a strong tendency to overestimate the expert-interpreted MABL height. This result suggests that, even for cases in



which a single MABL top can be interpreted, the Heffter method may detect an upper inversion layer or a residual layer located above the lower MABL top.

The bulk Richardson number method showed a correlation coefficient of 0.651, RMSE of 328.7 m, and MAE of 196.4 m against the expert-interpreted reference. Its mean bias was -7.8 m, indicating little net bias across the 666 profiles, although individual differences remained substantial. This diagnostic combines thermal stratification with the observed vector wind difference between each level and the reference level, and its agreement can vary with the lower-atmospheric stability and wind structure.

The θ_v threshold (0.5 K) method produced the highest correlation among the three selected profile-based diagnostics ($R = 0.850$) and the smallest RMSE (287.6 m), whereas its MAE was 205.0 m and its mean bias was $+176.1$ m. Thus, the bulk Richardson number method gave a slightly lower MAE (196.4 m), while the θ_v threshold method gave a higher correlation and lower RMSE. The positive bias of the θ_v threshold method indicates that the first $+0.5$ K transition often occurred above the expert-interpreted MABL top. These statistics quantify agreement with an expert-interpreted thermodynamic reference rather than independent physical accuracy.

ERA5 PBLH showed a correlation coefficient of 0.803 when compared with the expert-interpreted MABL height; its RMSE and MAE were 340.6 and 262.3 m, respectively, and its MB was $+240.9$ m. These values reflect the corrected UTC-based hourly ERA5 matching. ERA5 PBLH partly captured the variability of expert-interpreted MABL height, but generally overestimated it in the present dataset. The positive ERA5 PBLH bias contrasts with the continental radiosonde-based results of Guo et al. (2024), who reported a mean radiosonde-minus-ERA5 PBLH bias of 95.7 m. This contrast suggests that the sign and magnitude of an ERA5–observational difference depend on environmental conditions and the observational reference definition.

These comparisons show that traditional diagnostic methods and ERA5 PBLH can partly represent the overall variability of the MABL height, but they have limitations in complex marine boundary layer structures. In particular, when multiple inversion layers, residual layers, stable layers, and humidity layers coexist, single-threshold-based methods may not reliably identify the lower MABL top. Therefore, in the subsequent machine-learning analysis, the expert-interpreted MABL height was used as the reference value for evaluating machine-learning models that simultaneously considered multiple lower-atmospheric thermodynamic and flux variables.

3.3 Blind re-reading consistency of expert-interpreted MABL heights

To evaluate the internal consistency of the expert-interpreted MABL heights, blind re-reading was performed on 90 profiles following the stratified sampling and blinding procedures described in Sect. 2.4.

Table 3 compares the initial expert-interpreted values with the blind re-reading results. The re-read MABL heights exhibited high consistency with the initial interpretations, with $R = 0.909$, RMSE = 177.0 m, and MAE = 99.2 m. The mean bias was -57.1 m, and the median absolute difference was 37.2 m. In addition, the differences between the two interpretations were within 50 m for 60.0 % of the profiles, within 100 m for 73.3 %, and within 150 m for 80.0 % of the profiles.



Table 3. Intra-analyst repeatability of expert-interpreted MABL heights based on blind re-reading.

Group	N	R	Bias (m)	RMSE (m)	MAE (m)	Median absolute difference (m)	Within 50 m (%)	Within 100 m (%)	Within 150 m (%)
All re-read profiles	90	0.909	-57.1	177.0	99.2	37.2	60.0	73.3	80.0

550 Note: Blind re-reading was performed approximately three months after the initial expert interpretation. Ninety profiles were selected from expert-interpreted MABL profiles for which T850 and LTS were available, with 30 profiles randomly sampled from each T850–LTS-based thermodynamic regime using a random seed of 42. During rereading, the analyst had no access to the initial MABL heights, automated diagnostic results, ERA5 PBLH, thermodynamic regime labels, T850, LTS, or other reference MABL information. Mean bias was calculated as the difference between re-read MABL height and the initial expert-interpreted MABL height.

555 These results indicated that the expert-interpreted MABL heights were generally internally consistent. However, some profiles exhibited differences exceeding 100 m between the initial interpretation and the re-reading. This implies that a certain level of interpretation uncertainty may remain, even for cases in which a single MABL top can be relatively clearly defined, particularly when the transition near the MABL top is gradual or when multiple thermodynamic indicators are not fully consistent. Therefore, the errors of the diagnostic methods, ERA5 PBLH, and machine-learning models presented in the following sections should be interpreted not as errors against an absolute truth, but as relative agreement metrics with respect to the expert-interpreted reference dataset. The blind re-reading assessment evaluated intra-analyst repeatability and did not assess inter-analyst uncertainty.

565 3.4 T850–LTS-based thermodynamic regimes and MABL height characteristics

The K-means clustering results in the T850–LTS space are presented in Fig. 6. T850 and LTS were derived from the dropsonde vertical profiles as described in Sect. 2.5. The K-means analysis classified the dataset into three thermodynamic regimes: warm-stable, transitional, and cold/weakly stable. As shown in Fig. 6, the warm-stable regime was concentrated in the relatively high-T850 region, whereas the cold/weakly stable regime was mainly distributed in the low-T850 and low-LTS regions. The transitional regime was identified as an intermediate thermodynamic environment characterized by a low T850 and a relatively high LTS.

575 Table 4 summarizes the sample size and the mean and standard deviation of the T850, LTS, and expert-interpreted MABL heights for each thermodynamic regime. The sample sizes were 370, 130, and 142 for the warm-stable, transitional, and cold/weakly stable regimes, respectively. The warm-stable regime had relatively high T850 and moderate-to-high LTS, and exhibited the lowest mean MABL height of 363.2 ± 190.0 m. The cold/weakly stable regime had a low T850 and relatively

low LTS, exhibiting the highest mean MABL height of 859.1 ± 437.7 m. The transitional regime, with both a low T850 and high LTS, had an intermediate mean MABL height of 596.3 ± 364.8 m.

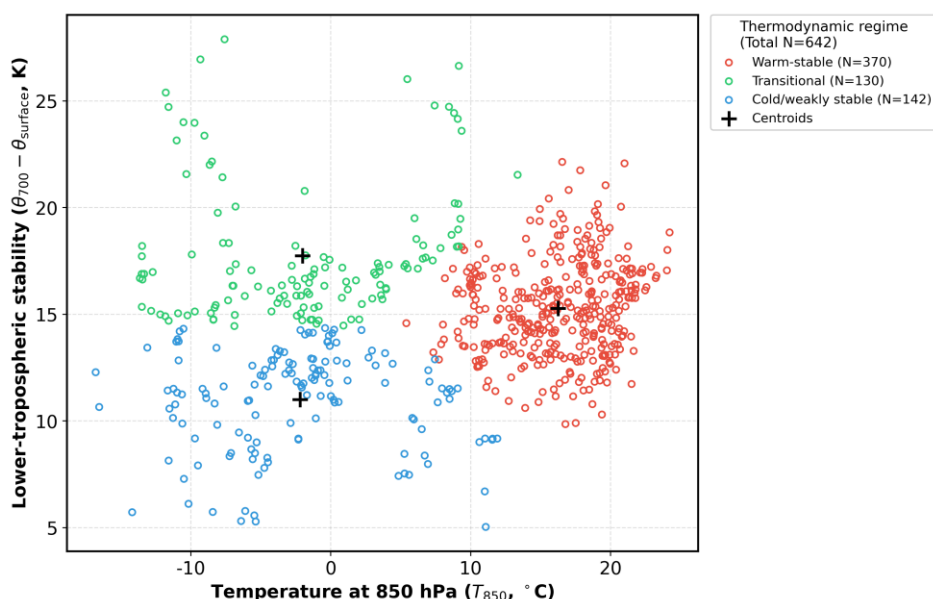


Figure 6. K-means classification of thermodynamic regimes in the T850–LTS space. Open circles indicate the 642 ML-complete dropsonde profiles, and black crosses denote the cluster centroids transformed back to the original T850–LTS units. The three regimes are labeled as cold/weakly stable (blue circle), transitional (green circle), and warm-stable (red circle).

Table 4. Characteristics of the thermodynamic regimes identified by K-means clustering using dropsonde-derived T850 and LTS.

Regime	N	T850 (°C)	LTS (K)	MABL height (m)
Warm-stable	370	16.3 ± 3.8	15.3 ± 2.2	363.2 ± 190.0
Transitional	130	-2.0 ± 6.9	17.7 ± 3.1	596.3 ± 364.8
Cold/weakly stable	142	-2.2 ± 6.5	11.0 ± 2.3	859.1 ± 437.7

Note: Values are presented as mean $\pm$ standard deviation. T850 and LTS were derived from the dropsonde profiles, as described in Sect. 2.5.

The regime-wise statistics in Table 4 show that MABL height varies with both the thermal state of the lower free troposphere and lower-tropospheric stability. The warm-stable regime, characterized by relatively high T850 and LTS, exhibited the shallowest MABL, suggesting that strong lower-atmospheric stability limited its vertical development. In contrast, the cold/weakly stable regime exhibited the deepest MABL, indicating that weak lower-atmospheric stability favored greater

MABL development. The transitional regime, characterized by low T850 and high LTS, showed an intermediate mean MABL height.

3.5 Performance of the global machine-learning models for MABL height estimation

The out-of-fold predictions from the global ML models are compared with the expert-interpreted MABL heights in Fig. 7. All machine-learning models and the ERA5 PBLH were evaluated using the same 642 ML-complete profiles.

All machine-learning models exhibited prediction distributions closer to the expert-interpreted MABL height than the ERA5 PBLH. SVR achieved $R = 0.766$, $RMSE = 241.7$ m, and $MAE = 162.7$ m, while Random Forest achieved $R = 0.792$, $RMSE = 220.6$ m, and $MAE = 155.5$ m. XGBoost achieved $R = 0.791$, $RMSE = 221.8$ m, and $MAE = 155.9$ m, while Gradient Boosting achieved $R = 0.788$, $RMSE = 223.0$ m, and $MAE = 158.5$ m. DNN showed the highest correlation ($R = 0.815$) and the lowest RMSE (209.5 m), while its MAE (155.6 m) was comparable to those of Random Forest and XGBoost (155.5–155.9 m). Among these, Random Forest achieved a comparable MAE with the additional advantage of interpretable variable-importance analysis. Table 5 summarizes the quantitative performance of the global machine-learning models and the ERA5 PBLH, complementing the results shown in Fig. 7.

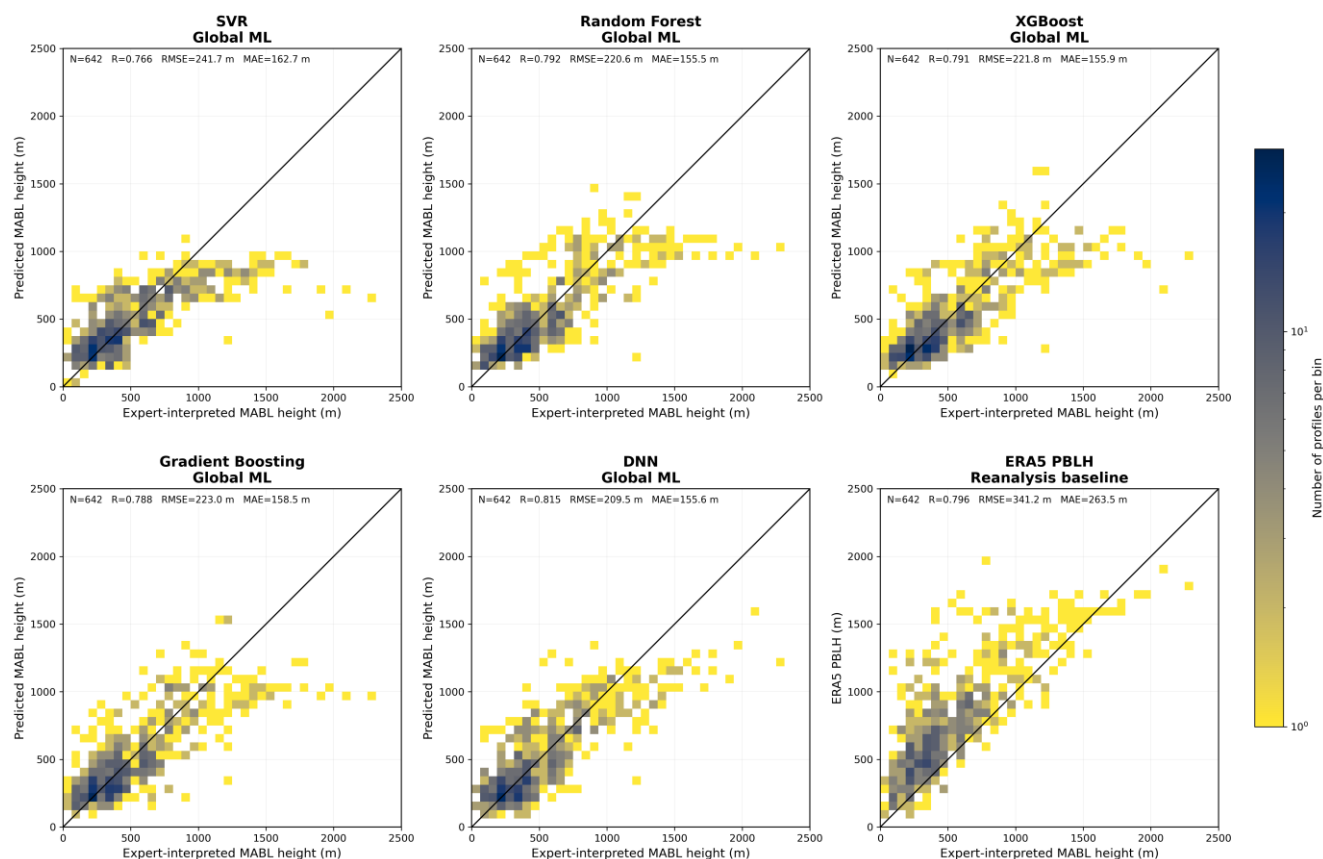


Figure 7. Two-dimensional frequency distributions comparing expert-interpreted MABL heights with date-grouped out-of-fold estimates from the global machine-learning models and ERA5 PBLH. Colors indicate the number of profiles per bin using a logarithmic scale. The solid line indicates the 1:1 line. Statistics in each panel denote N, R, RMSE, and MAE.

For the same 642 profiles, ERA5 PBLH achieved $R = 0.796$, $RMSE = 341.2$ m, and $MAE = 263.5$ m. All machine-learning models showed relatively small negative mean biases, ranging from -34.4 to -6.8 m, indicating slight underestimation on average, whereas ERA5 PBLH showed a large positive mean bias of 243.3 m. Although the correlation coefficient of ERA5 PBLH was comparable to those of some machine-learning models, its RMSE and MAE were larger than those of all machine-learning models. This indicates that ERA5 PBLH partly reproduced the linear variability of the MABL height but had relatively large absolute height errors. In contrast, the machine-learning models produced predictions closer to the expert-interpreted MABL height using the combined information from lower-atmospheric thermodynamic variables, surface fluxes, and the wind-speed difference. Therefore, although the best-performing metrics differed slightly among the models, Random Forest, XGBoost, and DNN all showed clearly lower errors than the ERA5 PBLH relative to the expert-interpreted MABL height reference.

Table 5. Performance of the global machine-learning models and ERA5 PBLH against the expert-interpreted MABL height for the 642 ML-complete profiles.

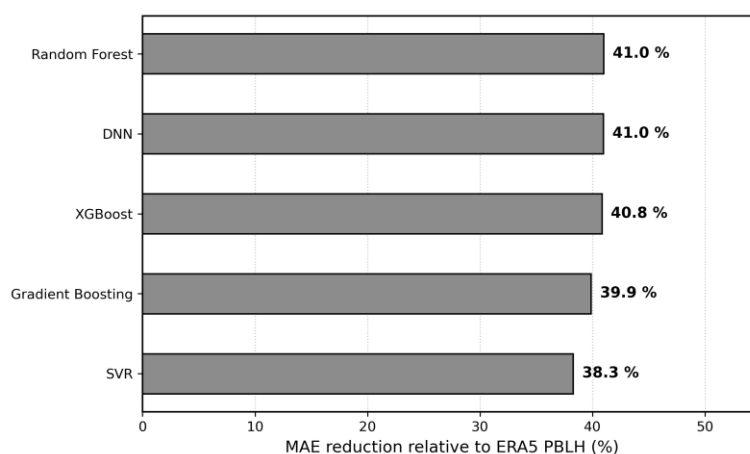
Method	N	R	RMSE (m)	MAE (m)	MB (m)
SVR	642	0.766	241.7	162.7	-34.4
Random Forest	642	0.792	220.6	155.5	-12.3
XGBoost	642	0.791	221.8	155.9	-18.6
Gradient Boosting	642	0.788	223.0	158.5	-18.0
DNN	642	0.815	209.5	155.6	-6.8
ERA5 PBLH	642	0.796	341.2	263.5	243.3

Note: Mean bias was calculated as the prediction minus the expert-interpreted MABL height. Machine-learning performance metrics were calculated based on out-of-fold predictions obtained using date-grouped five-fold GroupKFold cross-validation with observation date as the grouping variable. ERA5 PBLH was evaluated using the same 642 ML-complete profiles.

Figure 8 presents the relative MAE reduction of each machine-learning model using the ERA5 PBLH MAE as the reference. All machine-learning models reduced the MAE by 38.3–41.0 % relative to ERA5 PBLH. The 38.3–41.0 % reduction in MAE obtained in this study is broadly consistent with the improvement reported by Krishnamurthy et al. (2021), who found an approximately 50 % reduction in daytime MAE for an RF-based boundary-layer-height estimate relative to a Doppler-lidar threshold method. Although the datasets, atmospheric environments, reference definitions, and validation strategies



635 differed, both studies demonstrate that multivariable machine-learning approaches can reduce errors relative to conventional boundary-layer-height estimates. The reduction rates were 38.3 %, 39.9 %, 40.8 %, 41.0 %, and 41.0 % for the SVR, Gradient Boosting, XGBoost, DNN, and Random Forest, respectively. Therefore, although some differences were observed among the algorithms, all global machine-learning models had lower MAEs than ERA5 PBLH when evaluated against the same expert-interpreted MABL height reference.



640 **Figure 8. MAE reductions of the machine-learning models relative to ERA5 PBLH. All MAE values were calculated against the expert-interpreted MABL heights.**

Overall, in the observation-date-based cross-validation using all 642 ML-complete profiles, the global machine-learning models showed lower errors than the ERA5 PBLH with respect to the expert-interpreted MABL height reference. In particular, Random Forest, XGBoost, and DNN showed similar MAEs (155.5–155.9 m), corresponding to MAE reductions of 40.8–41.0 % relative to ERA5 PBLH. This comparison should be interpreted as an assessment of agreement with the expert-interpreted dropsonde reference dataset rather than as validation against an independent observational reference. Furthermore, the results indicate that a machine-learning approach combining dropsonde-based lower-level and vertical variables with COARE 3.0-based surface fluxes provides a useful observation-based alternative to ERA5 PBLH for estimating MABL height in well-defined cases.

3.6 Random Forest-based variable importance and thermodynamic-regime-stratified conditional diagnostics

The permutation importance results for the global Random Forest model are shown in Fig. 9. The permutation importance was calculated based on the increase in MAE in the validation data when each input variable was randomly permuted. Therefore, larger values indicate greater predictive importance.

Based on the permutation-importance values averaged across the five validation folds, the most important variable was lowest-level relative humidity (RH). When RH was randomly permuted, the MAE increased by an average of 145.4 m, indicating that the global Random Forest model used the lower-level moisture structure as a key predictor for MABL height



estimation. The importance of RH is consistent with the findings of Krishnamurthy et al. (2021), who also identified surface
660 relative humidity as one of the influential variables in RF-based boundary-layer-height estimation. However, the particularly
large importance of RH in the present study may also reflect the use of distinct changes in humidity and moisture structure as
criteria for constructing the expert-interpreted MABL-height reference. The next most important variable was the 925 hPa
relative humidity (RH925), for which the mean increase in MAE after permutation was 23.0 m. Because RH925 can reflect
the moisture structure of the lower free troposphere or the region near the upper part of the boundary layer, it can be
665 interpreted as contributing, together with lower-level RH, to identifying the MABL top. However, because distinct decreases
in RH and changes in moisture structure were used as major criteria for identifying the MABL top during the expert
interpretation process, the high permutation importance of RH and RH925 should not be interpreted as evidence of an
independent causal relationship. Rather, it likely reflects, at least in part, the model's ability to reproduce the expert-
interpreted reference dataset.

670

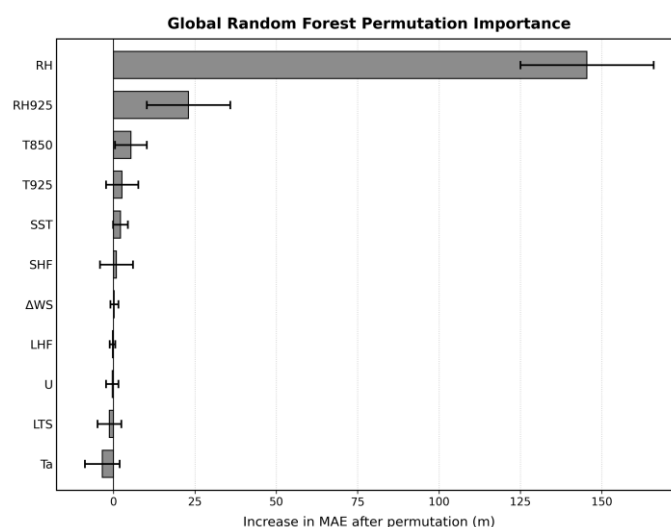


Figure 9. Fold-wise permutation importance of the global Random Forest model using date-grouped cross-validation. Bars indicate the mean increase in MAE after feature permutation, and error bars indicate ± 1 standard deviation across the five cross-validation folds. Δ WS denotes the wind-speed difference calculated as the wind speed at 850 hPa minus the lowest-level wind speed within the 2–30 m layer.
675

T850, T925, SST, and SHF exhibited relatively small positive permutation importance values. The mean increases in the
MAE for these variables were 5.3, 2.6, 2.1, and 0.9 m, respectively. In contrast, the permutation importance values of the
wind-speed difference, LHF, U, LTS, and Ta were close to zero or negative. This does not necessarily mean that these
680 variables are physically unimportant for MABL height estimation. Rather, this may indicate that predictive information was
distributed among the correlated input variables or that the error changes caused by permutations were not consistent across



folds because of the limited validation sample size. Therefore, the permutation importance in this study was interpreted as the relative additional information provided by each variable when the trained Random Forest model predicted the expert-interpreted MABL height rather than as evidence of the independent physical causal effect of each variable.

Figure 10 shows the permutation importance of the global Random Forest model calculated separately for each T850–LTS-based thermodynamic regime, as described in Sect. 2.7. This analysis was not intended to compare the performance of regime-specific models, but to interpret the regime-dependent predictor importance of the global model.

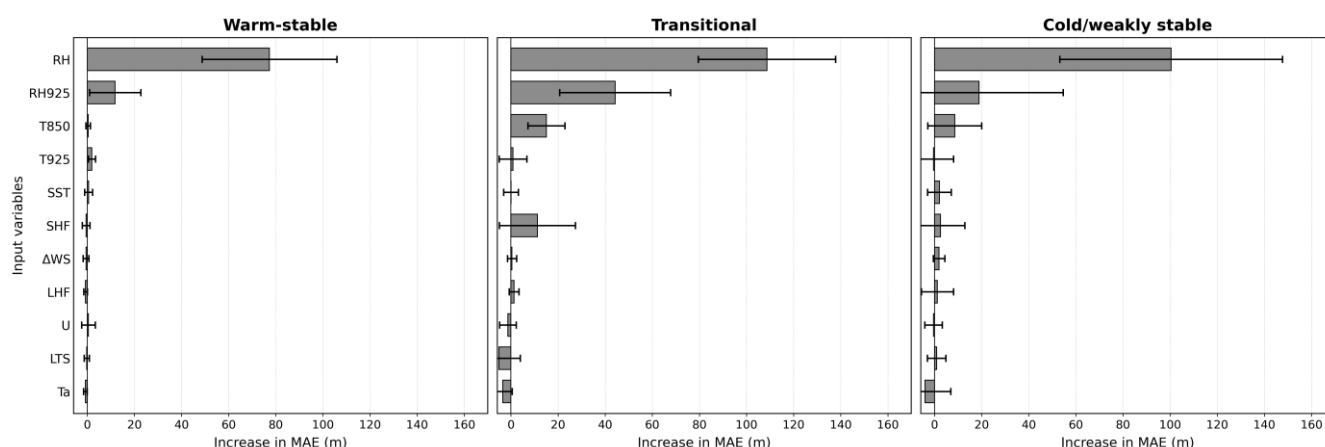


Figure 10. Regime-stratified fold-wise permutation importance of the global Random Forest model using date-grouped cross-validation. The same global Random Forest model fitted to all training profiles in each fold was evaluated separately for the warm-stable, transitional, and cold/weakly stable validation subsets. Bars indicate the mean increase in MAE after feature permutation, and error bars indicate ± 1 standard deviation across the validation folds used for each regime.

RH was the most important variable across all three thermodynamic regimes, and RH925 served as an important auxiliary variable. Together, these findings indicate that the lower-tropospheric moisture structure provided important predictive information for reproducing the expert-interpreted MABL height reference across all three regimes. However, the relative importance of the auxiliary variables differed among regimes. In the warm-stable regime, the influence of lower-level humidity variables was particularly pronounced, suggesting that humidity gradients may have served as an important clue for identifying the MABL top under stable lower-atmospheric conditions. In the transitional and cold/weakly stable regimes, T850, T925, and surface-flux-related variables made relatively larger contributions than they did in the warm-stable regime, although RH remained the dominant predictor. This pattern suggests that thermodynamic conditions associated with cold-air advection or air–sea temperature differences may have provided additional information on variations in MABL depth.

In summary, the Random Forest-based variable importance analysis showed that the lowest-level RH and RH925 provided the largest predictive contributions for reproducing the expert-interpreted MABL height reference. The thermodynamic-regime-stratified conditional diagnostics further indicated that the relative importance of the auxiliary predictors varied with



environmental conditions. These results suggest that the proposed machine-learning model used statistical information on the lower-level moisture structure and thermodynamic environment related to MABL height variability. However, the importance of these variables should be interpreted as the relative predictive contribution within the trained model for reproducing the expert-interpreted reference dataset, rather than as an independent physical causal effect of each variable.

Overall, the results of this study show that traditional threshold-based diagnostic methods and the ERA5 PBLH can capture the general variability of well-defined MABL cases but may differ substantially from the expert-interpreted reference dataset under complex marine boundary layer conditions. In contrast, the proposed machine-learning models consistently showed a lower MAE than the ERA5 PBLH in the observation-date-based GroupKFold validation, suggesting that dropsonde-based lower-atmospheric thermodynamic information and COARE 3.0-based surface fluxes provide useful predictive information for reproducing the MABL height reference. However, these results should be interpreted as an assessment of agreement with the expert-interpreted reference dataset rather than as validation against an independent observational reference. The RH and moisture flux information used during the expert interpretation may also be reflected in the permutation importance results. Therefore, the machine-learning framework developed in this study should be interpreted as an observation-based tool for estimating and evaluating MABL height in well-defined cases over the seas around the Korean Peninsula. Additional external validation and comparisons across different observing platforms are required before applying the framework to independent regions or real-time operational applications.

4 Conclusions

In this study, a framework for MABL height estimation and evaluation was developed using aircraft dropsonde observations collected over the seas around the Korean Peninsula during 2020–2024. A total of 1193 dropsonde profiles were collected, of which 1179 were spatiotemporally collocated with ERA5. The 1138 profiles with valid observations in the lower 2–30 m layer were used to evaluate the ERA5 lower-atmospheric variables and surface fluxes, and the 666 profiles with valid expert-interpreted MABL heights were used to evaluate the traditional diagnostic methods and ERA5 PBLH. Finally, the performance of the machine-learning-based MABL height estimation was evaluated using 642 ML-complete profiles for which both the expert-interpreted MABL height and all 11 machine-learning input variables were available.

When selected profile-based diagnostics and ERA5 PBLH were compared with the expert-interpreted MABL heights, the Heffter criterion strongly overestimated the reference height. The corrected bulk Richardson number method yielded the lowest MAE (196.4 m) among the three profile-based diagnostics, whereas the simplified θ_v threshold (0.5 K) method showed the highest correlation ($R = 0.850$) and the lowest RMSE (287.6 m). ERA5 PBLH generally exceeded the expert-interpreted height (MB = +240.9 m). These results reflect agreement with the reference constructed from clearly identifiable thermodynamic transitions rather than an independent verification of absolute turbulent boundary layer height. They also demonstrate that method ranking depends on the error metric used.

To evaluate the internal consistency of the expert-interpreted MABL heights, a blind re-reading was performed on 90 profiles. The initial and re-read interpretations showed consistent results, with $R = 0.909$, RMSE = 177.0 m, and MAE = 99.2



m. The mean bias was -57.1 m, and the median absolute difference was 37.2 m. The differences between the two interpretations were within 50 m for 60.0 % of the profiles, within 100 m for 73.3 %, and within 150 m for 80.0 % of the profiles. These results indicate that the expert-interpreted MABL heights are generally repeatable, but a certain level of interpretation uncertainty could remain for complex or ambiguous vertical structures. Therefore, the machine-learning performance reported in this study should be interpreted within the uncertainty range associated with the dropsonde-profile-based reference dataset, rather than treating the expert-interpreted MABL height as an absolute truth.

K-means clustering using T850 and LTS classified MABL environments around the Korean Peninsula into three thermodynamic regimes: warm-stable, transitional, and cold/weakly stable. The warm-stable regime was characterized by relatively shallow MABL heights, whereas the cold/weakly stable regime showed the highest mean MABL height. This indicates that the MABL height can vary depending on both the temperature state of the lower free troposphere and the stability difference between the near-surface layer and the lower free troposphere. However, the thermodynamic regimes used in this study were not used to train separate regime-specific prediction models. Instead, they were used as diagnostic criteria to interpret the physical differences in the MABL height and regime-dependent characteristics of the global model.

Based on observation-date-based five-fold GroupKFold cross-validation, all machine-learning models showed lower RMSE and MAE values than the ERA5 PBLH for the same 642 ML-complete profiles. Random Forest, XGBoost, and DNN showed nearly identical MAEs of 155.5 – 155.9 m. DNN showed the lowest RMSE of 209.5 m and the highest correlation coefficient ($R = 0.815$). Compared with the ERA5 PBLH MAE of 263.5 m, the machine-learning models achieved a 38.3 – 41.0 % reduction in MAE. These results demonstrate that machine-learning models integrating dropsonde-based lower-atmospheric and vertical-profile variables, ERA5 SST, and COARE 3.0 surface fluxes produced MABL-height estimates that agreed more closely with the expert-interpreted reference than ERA5 PBLH.

In the permutation importance analysis of the global Random Forest model, lowest-level RH and RH925 made the largest contributions to predicting the expert-interpreted MABL height. The thermodynamic-regime-stratified conditional diagnostics also showed that RH was the key predictor across all regimes, whereas the relative contributions of T850, T925, and surface-flux-related variables varied among the regimes. These results suggest that the lower-level moisture structure provided important predictive information for reproducing the expert-interpreted MABL height reference. However, because distinct decreases in RH and changes in the moisture structure were also used as major indicators for identifying the MABL top during the expert interpretation, the high importance of RH and RH925 should be interpreted as reflecting their contributions to reproducing the expert-interpreted reference dataset rather than as evidence of independent causal effects.

Overall, this study demonstrates the potential of machine-learning-based MABL height estimation over the seas around the Korean Peninsula by combining multiyear aircraft dropsonde observations with an expert-interpreted MABL height reference dataset. However, the dataset was limited to cases and regions sampled by aircraft observations, and the expert-interpreted reference dataset contained a certain level of uncertainty. Future studies should further evaluate the generalizability of this framework to independent regions with different observation platforms and under real-time application conditions. In addition, by combining physically based diagnostic methods with machine-learning approaches,



775 this framework could be developed into a more reliable tool for MABL height estimation and numerical-model boundary layer evaluation over marine and coastal regions.

Code availability

780 The analysis code used in this study, including dropsonde preprocessing, ERA5 collocation, COARE 3.0-based bulk flux calculation, traditional MABL-height diagnostics, K-means regime classification, and machine-learning training, validation, and permutation importance analysis, is available from the corresponding author upon reasonable request.

Data availability

785 ERA5 hourly single-level reanalysis data are publicly available from the Copernicus Climate Data Store (<https://doi.org/10.24381/cds.adbb2d47>; Hersbach et al., 2023). The moored-buoy sea surface temperature observations used in this study were provided by the Korea Meteorological Administration through the Open MET Data Portal (<https://data.kma.go.kr/resources/html/en/aowdp.html>). Aircraft dropsonde observations were collected by the Korea Meteorological Administration/National Institute of Meteorological Sciences (KMA/NIMS) for research purposes and are subject to the data policy of the operating agency; therefore, they were not publicly redistributed in this study. Access to dropsonde data can be requested from the Observation Research Department, National Institute of Meteorological Sciences, 790 33 Seohobuk-ro, Seogwipo 63568, Jeju-do, Republic of Korea, or from the corresponding author. The processed per-profile ML-complete table, including the 11 predictor variables, expert-interpreted MABL heights, T850–LTS-based thermodynamic regimes, ERA5 PBLH, cross-validation fold information, and out-of-fold model predictions for the 642 ML-complete profiles, is provided in Supplementary Table S1.

Author contributions

795 MSK conceived the idea. MSK designed and conducted the experiments. MSK conducted the investigation and developed algorithms and software. MSK performed formal analysis of the results. MSK wrote the original draft and reviewed and edited the final manuscript. MSK supervised this study. TYG acquired the necessary funding for the project.

Competing interests

The authors declare that they have no competing interests.



800 Acknowledgements

The authors used AI-based language tools (ChatGPT, OpenAI; Claude, Anthropic) to assist in the translation of the manuscript from Korean to English, improve its grammar and readability, and develop and debug the analysis code. All scientific content, analyses, and interpretations were developed solely by the authors, who carefully reviewed and verified the translated text and analysis code and took full responsibility for the content of this manuscript.

805 Financial support

This work was funded by the Korea Meteorological Administration Research and Development Program “*Development of Application Technology on Atmospheric Research Aircraft*” (Grant No. KMA2018-00222).

References

- Arya, S. P.: Introduction to micrometeorology, 2nd edn., Academic Press, San Diego, CA, ISBN 978-0-12-059354-5, 810 420 pp., 2001.
- Blount, D. V., Evans, C., Jirak, I. L., Dean, A. R., and Kravtsov, S.: An objective method for clustering observed vertical thermodynamic profiles by their boundary layer structure, *Weather Forecast.*, 38, 1143–1156, <https://doi.org/10.1175/WAF-D-22-0195.1>, 2023.
- 815 Carneiro, R. G., Ribeiro, M. M., Gatti, L. V., de Souza, C. M. A., Dias-Júnior, C. Q., Tejada, G., Domingues, L. G., Rykowska, Z., Dos Santos, C. A., and Fisch, G.: Assessing the effectiveness of convective boundary layer height estimation using flight data and ERA5 profiles in the Amazon biome, *Clim. Dyn.*, 63, 109, <https://doi.org/10.1007/s00382-025-07609-8>, 2025.
- 820 Chen, X., Yang, T., Wang, Z., Wang, F., and Wang, H.: An ensemble method for improving the estimation of planetary boundary layer height from radiosonde data, *Atmos. Meas. Tech.*, 16, 4289–4302, <https://doi.org/10.5194/amt-16-4289-2023>, 2023.
- 825 Fairall, C. W., Bradley, E. F., Hare, J. E., Grachev, A. A., and Edson, J. B.: Bulk parameterization of air–sea fluxes: updates and verification for the COARE algorithm, *J. Clim.*, 16, 571–591, [https://doi.org/10.1175/1520-0442\(2003\)016<0571:BPOASF>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2), 2003.



- Feng, J. and Wang, X.: Impact of assimilating upper-level dropsonde observations collected during the TCI field campaign on the prediction of intensity and structure of Hurricane Patricia (2015), *Mon. Weather Rev.*, 147, 3069–3089, 830 <https://doi.org/10.1175/MWR-D-18-0305.1>, 2019.
- Giangrande, S. E., Wang, D., and Mechem, D. B.: Cloud regimes over the Amazon Basin: perspectives from the GoAmazon2014/5 campaign, *Atmos. Chem. Phys.*, 20, 7489–7507, <https://doi.org/10.5194/acp-20-7489-2020>, 2020.
- 835 Guo, J., Zhang, J., Shao, J., Chen, T., Bai, K., Sun, Y., Li, N., Wu, J., Li, R., Li, J., Guo, Q., Cohen, J. B., Zhai, P., Xu, X., and Hu, F.: A merged continental planetary boundary layer height dataset based on high-resolution radiosonde measurements, ERA5 reanalysis, and GLDAS, *Earth Syst. Sci. Data*, 16, 1–14, <https://doi.org/10.5194/essd-16-1-2024>, 2024.
- Heffter, J. L.: Air resources laboratories atmospheric transport and dispersion model (ARL-ATAD), NOAA Technical 840 Memorandum ERL ARL-81, Air Resources Laboratories, Silver Spring, MD, 1980.
- Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S): Climate Data Store (CDS), data set, <https://doi.org/10.24381/cds.adbb2d47>, 845 2023.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., 850 Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Q. J. R. Meteorol. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Hock, T. F. and Franklin, J. L.: The NCAR GPS dropwindsonde, *Bull. Amer. Meteor. Soc.*, 80, 407–420, 855 [https://doi.org/10.1175/1520-0477\(1999\)080<0407:TNGD>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<0407:TNGD>2.0.CO;2), 1999.
- Intrieri, J. M., de Boer, G., Shupe, M. D., Spackman, J. R., Wang, J., Neiman, P. J., Wick, G. A., Hock, T. F., and Hood, R. E.: Global Hawk dropsonde observations of the Arctic atmosphere obtained during the Winter Storms and Pacific Atmospheric Rivers (WISPAR) field campaign, *Atmos. Meas. Tech.*, 7, 3917–3926, [https://doi.org/10.5194/amt-7-3917-](https://doi.org/10.5194/amt-7-3917-2014) 860 2014, 2014.



Jaimes de la Cruz, B., Shay, L. K., Wadler, J. B., and Rudzin, J. E.: On the hyperbolicity of the bulk air–sea heat flux functions: insights into the efficiency of air–sea moisture disequilibrium for tropical cyclone intensification, *Mon. Weather Rev.*, 149, 1517–1534, <https://doi.org/10.1175/MWR-D-20-0324.1>, 2021.

865

Kim, J.-H., Goo, T.-Y., Jung, S.-P., Kim, M.-S., Lee, K., Kang, M., Lee, C., Yang, J., Hong, S., Ko, H., and Yun, J. H.: Overview of the KMA / NIMS Atmospheric Research Aircraft (NARA) and its data archive: annual airborne observations over the Korean peninsula, *Geosci. Data J.*, 10, 447–460, <https://doi.org/10.1002/gdj3.182>, 2023a.

870 Kim, M.-S., Kwon, B. H., Goo, T.-Y., and Jung, S.-P.: Dropsonde-based heat fluxes and mixed layer height over the sea surface near the Korean Peninsula, *Remote Sens.*, 15, 25, <https://doi.org/10.3390/rs15010025>, 2023b.

Klein, S. A. and Hartmann, D. L.: The seasonal cycle of low stratiform clouds, *J. Clim.*, 6, 1587–1606, [https://doi.org/10.1175/1520-0442\(1993\)006<1587:TSCOLS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<1587:TSCOLS>2.0.CO;2), 1993.

875

Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D.: On the estimation of boundary layer heights: a machine learning approach, *Atmos. Meas. Tech.*, 14, 4403–4424, <https://doi.org/10.5194/amt-14-4403-2021>, 2021.

880 Lee, C., Chang, K.-H., Jung, J.-W., Cha, J.-W., Choi, Y.-J., and Kim, K.: Strategy for the meteorological and environmental airborne observations over the Korean Peninsula, *Asia-Pacific J. Atmos. Sci.*, 47, 91–96, <https://doi.org/10.1007/s13143-011-1005-8>, 2011.

885 Li, H., Vulova, S., Rocha, A. D., and Kleinschmit, B.: Spatio-temporal feature attribution of European summer wildfires with Explainable Artificial Intelligence (XAI), *Sci. Total Environ.*, 916, 170330, <https://doi.org/10.1016/j.scitotenv.2024.170330>, 2024.

Liu, S. and Liang, X.-Z.: Observed diurnal cycle climatology of planetary boundary layer height, *J. Clim.*, 23, 5790–5809, <https://doi.org/10.1175/2010JCLI3552.1>, 2010.

890 Ozupek, E., Teke, A., Celik, N., and Kavzoglu, T.: Explainable artificial intelligence to explore the intrinsic characteristics of climatic parameters governing meteorological drought forecasting: opening the black box, *Stoch. Environ. Res. Risk Assess.*, 39, 3201–3222, <https://doi.org/10.1007/s00477-025-03007-y>, 2025.



- Peng, K., Xin, J., Zhu, X., Wang, X., Cao, X., Ma, Y., Ren, X., Zhao, D., Cao, J., and Wang, Z.: Machine learning model to
895 accurately estimate the planetary boundary layer height of Beijing urban area with ERA5 data, *Atmos. Res.*, 293, 106925,
<https://doi.org/10.1016/j.atmosres.2023.106925>, 2023.
- Ralph, F. M., Simpson, M., Iacobellis, S., Cordeira, J. M., Cannon, F., Cobb, A., Michaelis, A. C., and Delle Monache, L.:
900 Dropsonde observations of the stable marine boundary layer beneath atmospheric rivers, *Mon. Weather Rev.*, 152, 2735–
2751, <https://doi.org/10.1175/MWR-D-23-0207.1>, 2024.
- Rieutord, T., Aubert, S., and Machado, T.: Deriving boundary layer height from aerosol lidar using machine learning: KABL
and ADABL algorithms, *Atmos. Meas. Tech.*, 14, 4335–4353, <https://doi.org/10.5194/amt-14-4335-2021>, 2021.
- 905 Seidel, D. J., Ao, C. O., and Li, K.: Estimating climatological planetary boundary layer heights from radiosonde observations:
Comparison of methods and uncertainty analysis, *J. Geophys. Res.-Atmos.*, 115, D16113,
<https://doi.org/10.1029/2009JD013680>, 2010.
- Silva, S. J. and Keller, C. A.: Limitations of XAI methods for process-level understanding in the atmospheric sciences, *Artif.*
Intell. Earth Syst., 3, e230045, <https://doi.org/10.1175/AIES-D-23-0045.1>, 2024.
- 910 Wood, R. and Bretherton, C. S.: On the relationship between stratiform low cloud cover and lower-tropospheric stability, *J.*
Clim., 19, 6425–6432, <https://doi.org/10.1175/JCLI3988.1>, 2006.
- 915 Xi, X., Yang, Q., Liu, C., Shupe, M. D., Han, B., Peng, S., Zhou, S., and Chen, D.: Evaluation of the planetary boundary
layer height from ERA5 reanalysis with MOSAiC observations over the Arctic Ocean, *J. Geophys. Res.-Atmos.*, 129,
e2024JD040779, <https://doi.org/10.1029/2024JD040779>, 2024.
- Zhang, J. A., Cione, J. J., Kalina, E. A., Uhlhorn, E. W., Hock, T., and Smith, J. A.: Observations of infrared sea surface
920 temperature and air–sea interaction in Hurricane Edouard (2014) using GPS dropsondes, *J. Atmos. Ocean. Technol.*, 34,
1333–1349, <https://doi.org/10.1175/JTECH-D-16-0211.1>, 2017.
- Zheng, M., Torn, R., Delle Monache, L., Doyle, J., Ralph, F. M., Tallapragada, V., Davis, C., Steinhoff, D., Wu, X., Wilson,
A., Papadopoulos, C., and Mulrooney, P.: An assessment of dropsonde sampling strategies for atmospheric river
925 reconnaissance, *Mon. Weather Rev.*, 152, 811–835, <https://doi.org/10.1175/MWR-D-23-0111.1>, 2024.