Estimating PBL height using Doppler lidar using machine learning

June 22, 2021

Raghu Krishnamurthy, Rob Newsom and Larry Berg

Pacific Northwest National Laboratory

PNNL is operated by Battelle for the U.S. Department of Energy

Planetary boundary layer height (z_i)

- z_i definition: The PBL height is generally defined as the altitude of a transition layer where air temperature or humidity gradient are significant within the lowest 1-5 kilometers above the surface.

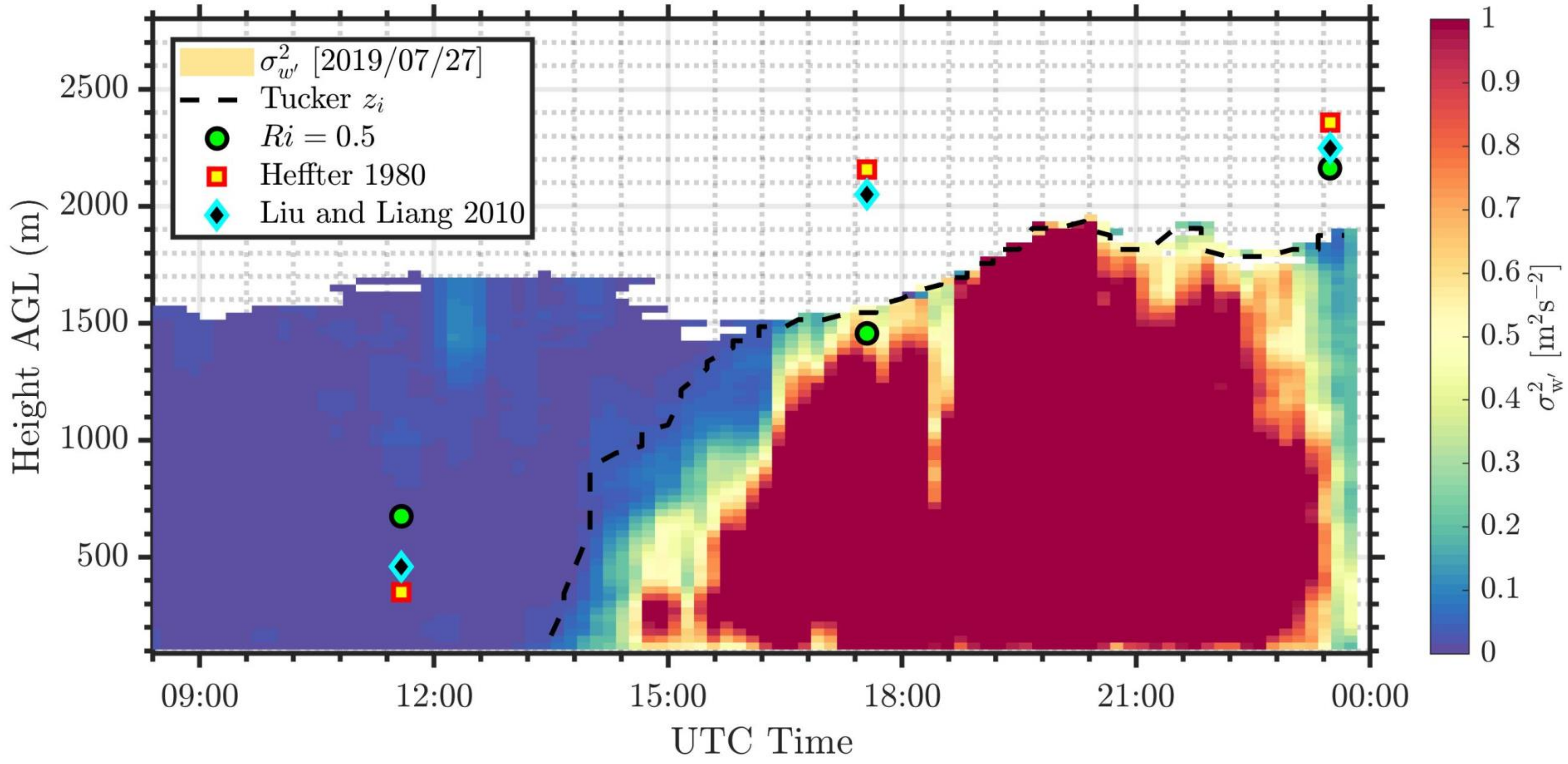
- During daytime, a mixed layer of vigorous turbulence grows in depth, capped by a statically stable entrainment zone of intermittent turbulence.
  - Radiosondes are standard data for calculating z_i

- Radiosonde releases are intermittent (4 times daily at SGP C1), so assessing z_i using remote sensing data is important.

- Doppler lidars can measure the turbulence growth directly, rather than secondary assumptions from aerosol backscatter.

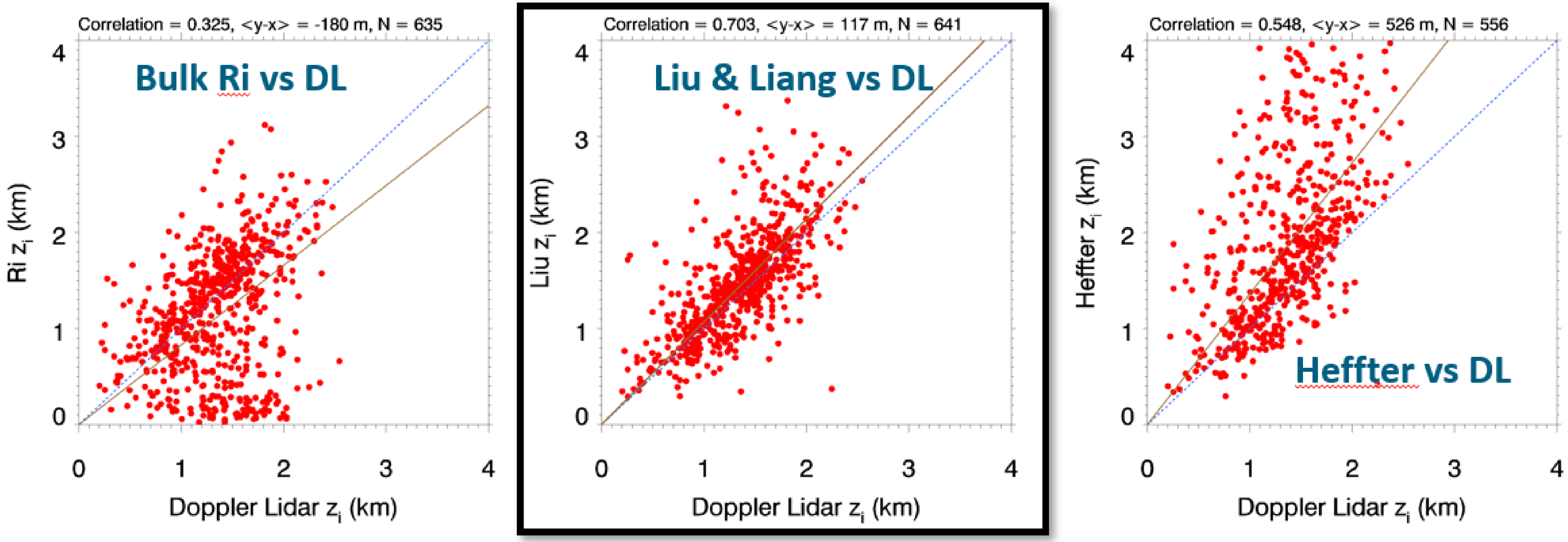
Motivation

- Doppler lidar estimates of PBLH (z_i) are based on a vertical velocity variance thresholding technique (Tucker et al., 2009)
  - $z_i = z\ (\sigma_w^2 < 0.04\ m^2 s^{-2})$ – sensitive to the threshold.
  - Tends to underestimate PBLH estimates compared to radiosondes during peak convective conditions
  - Only applicable for convective boundary layer
  - Poor SNR > 2 km at SGP C1



Radiosonde z_i Model Evaluation

- Multiple models exist (Heffter 1986, Bulk Richardson number thresholds (0.25/0.5), & Liu & Liang 2010)
- Liu-Liang method chosen as baseline, since its based-on inversion height and wind shear and correlates well with standard lidar PBL height estimates



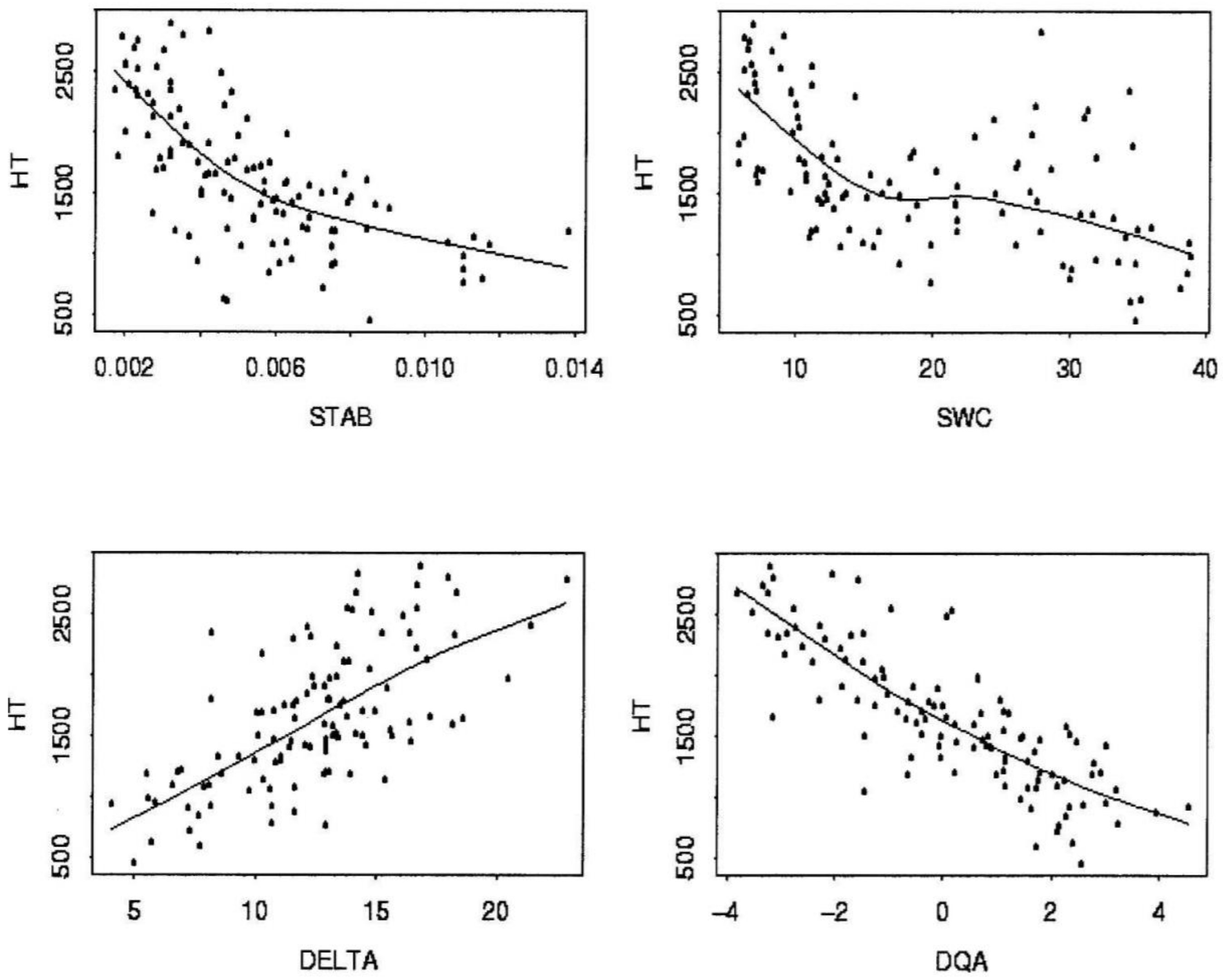
Surface link to convective z_i

- Surface properties, such as potential temperature, soil moisture, relative humidity, atmospheric stability etc., have shown positive correlations with convective z_i (Santanello et al., 2005)

- Other surface parameters, such as friction velocity (u_*), obukhov length (L) is also known to be important (Zilitinkevich 1972 or Brost and Wyngaard 1978)

$$z_i = 0.4 \left(\frac{u_* L}{|f|}\right)^{1/2}$$

f – Coriolis parameter

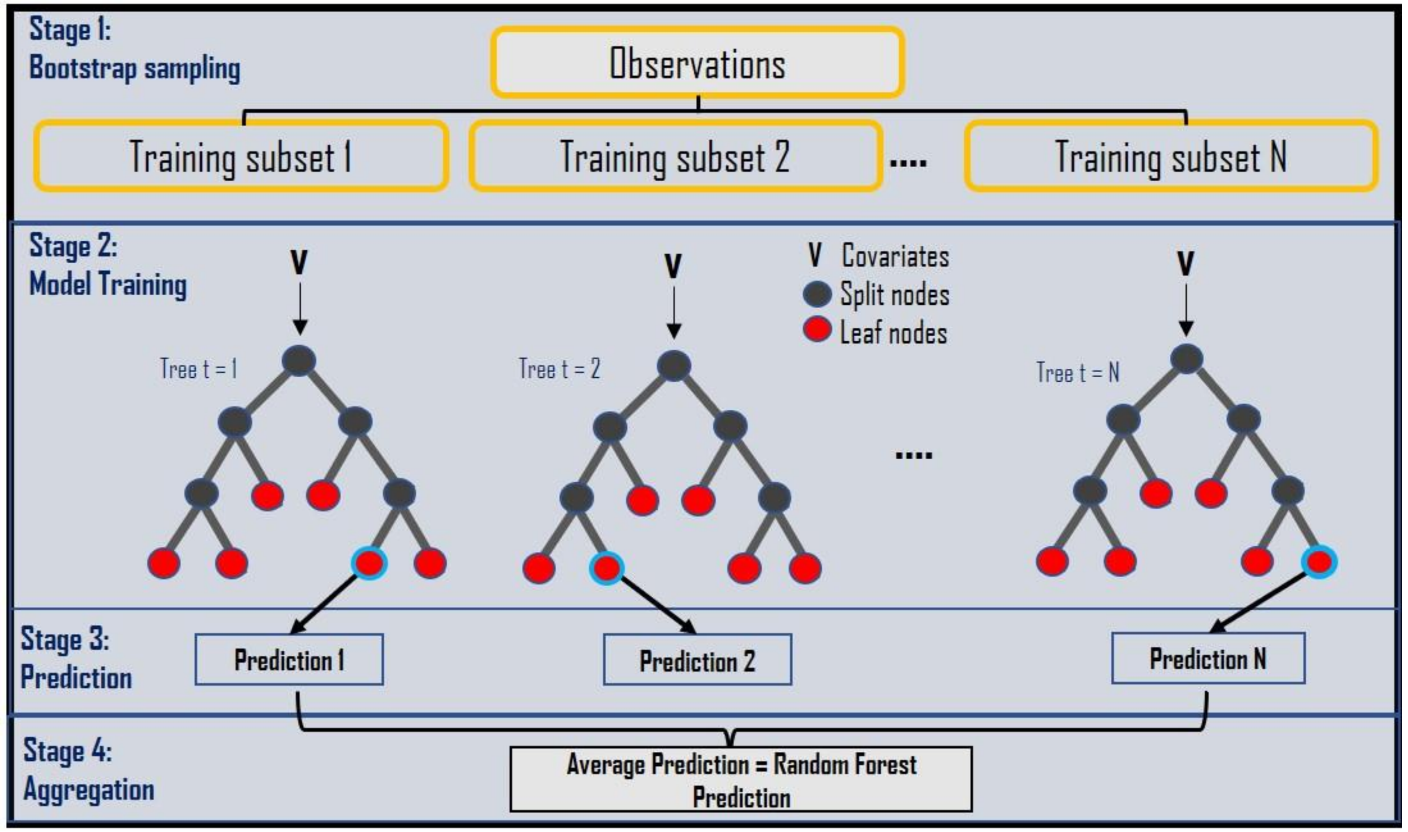


Observations of stability (STAB; K/m), soil water content (SWC; m^{3}m^3 percent volumetric), change in 2-m potential temperature (DELTA; K), and change in 2-m specific humidity (DQA; g kg1) plotted against the height of the PBL (HT; m) for all 132 days of the study. The lines represent local regression models fit to the data.

Random Forest Model

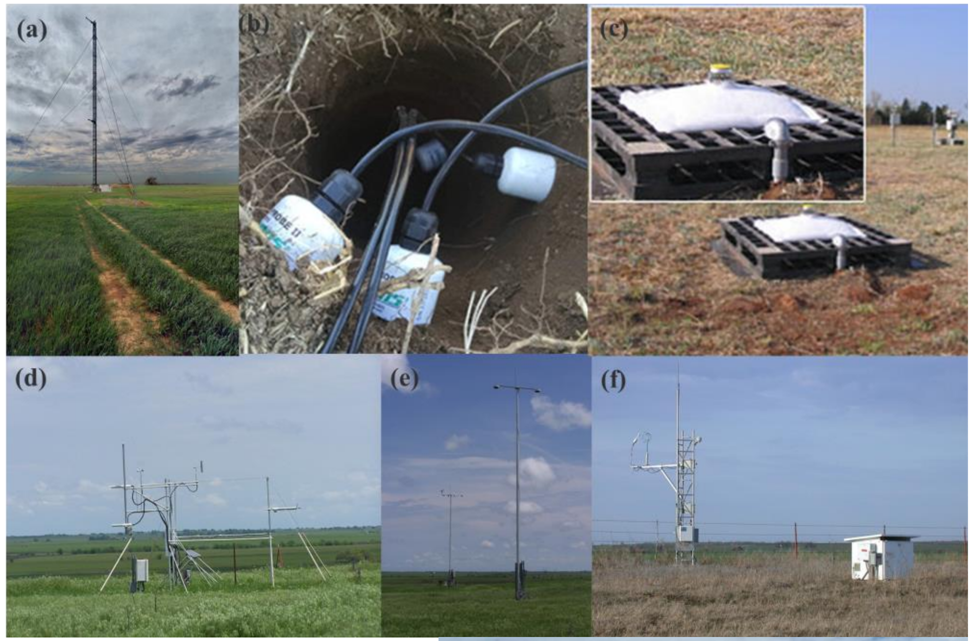
- **Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction**
  - Can be used for both classification and regression problems
- Predictive modeling tool not a descriptive tool

- Advantages
  - Versatility and
  - Relative importance to input features
- Disadvantages
  - Slow prediction with many trees
  - Possible overfitting of the data



Measurements

- List of variables used in Random Forest model

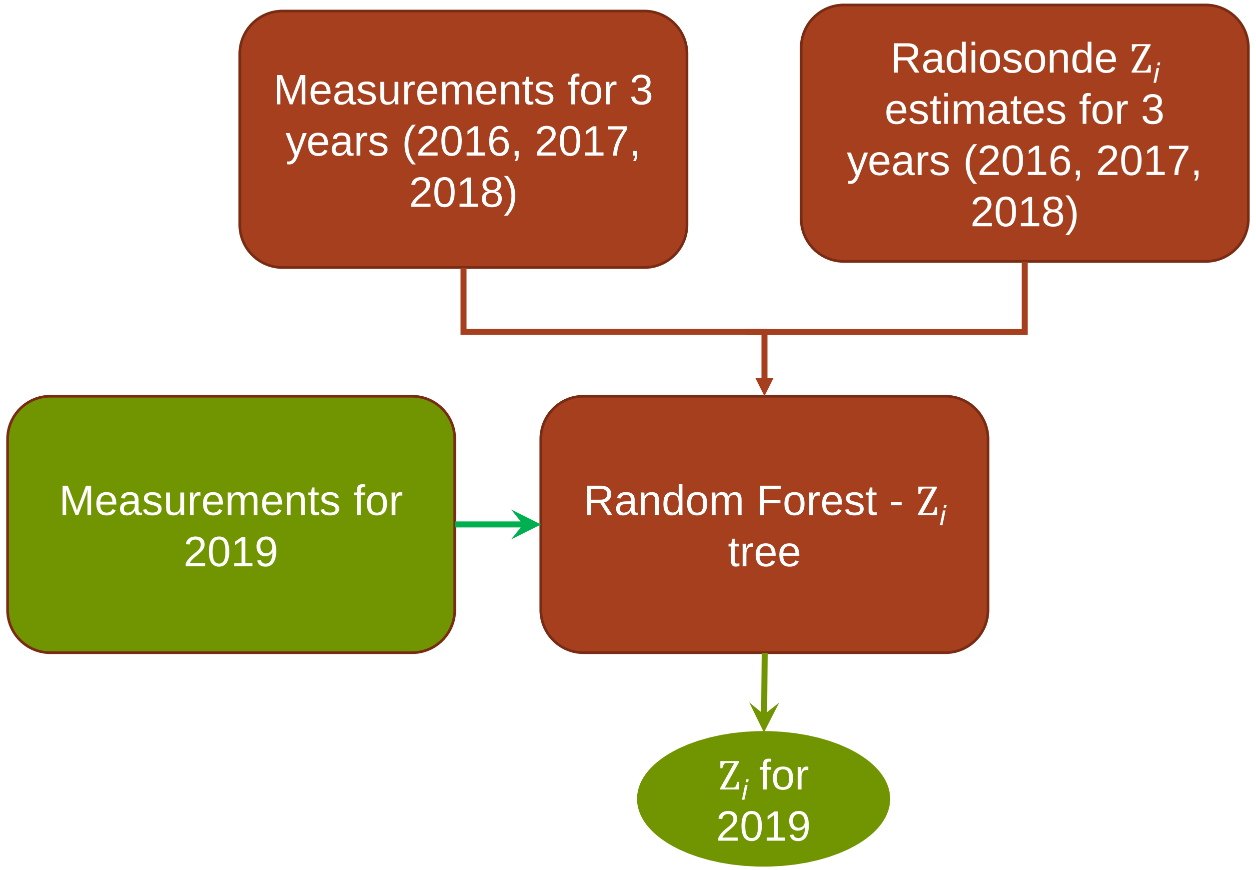


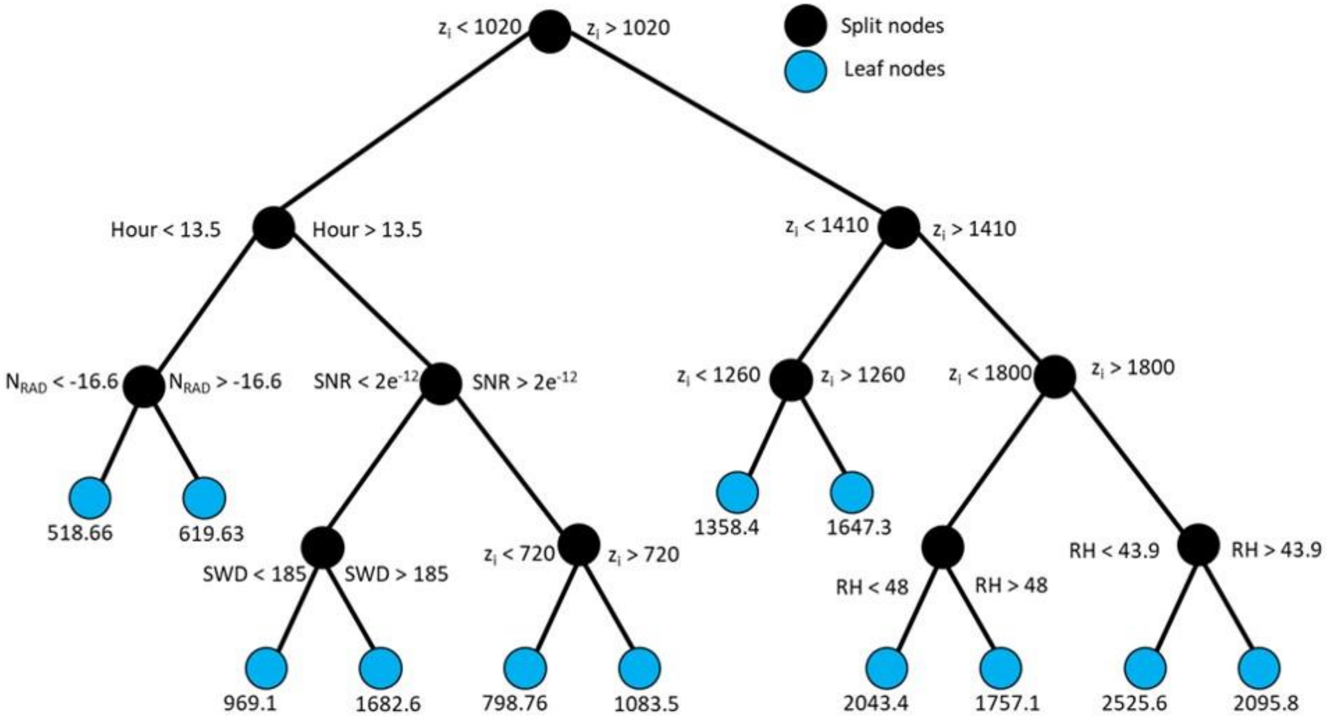
Instrumentation at SGP C1 a) 60-m tower, b) soil fluxes, c) disidrometer, (d) energy balance Bowen ratio (EBBR) station, (e) surface meteorological observation system (SMOS), (f) eddy-correlation (ECOR) flux station, and (g) Doppler lidar.

Instrument	ARM data stream	Measurements or features	Measurement height/ range	References
Radiosonde	sgppblhtsonde1mcfarlC1.c1	PBL height estimates (m)	100 to 5000 m a.g.l.	Sivaraman et al. (2013)
Surface eddy correlation station	sgp30co2flx25mC1.b1	Sensible heat flux (W m^{-2})	25 m a.g.l.	Cook (2018a) and Tang et al. (2019)
		Latent heat flux (W m^{-2})		
		Vertical velocity variance (m^2 s^{-2})		
		Friction velocity (m s^{-1})		
		Turbulence kinetic energy (m^2 s^{-2})		
		Monin–Obukhov length (m)		
		Wind speed (m s^{-1})		
		Wind direction (degrees from north)		
Surface meteorological station	sgpmetE13.b1	Air temperature (K)	4 m a.g.l.	Ritsche and Prell (2011)
		Relative humidity (%)		
Soil temperature and moisture probes	sgpstampE13.b1 or sgpswatsE13.b1	Soil moisture (m^3 m^{-3}) Soil temperature (°C)	−5 cm below surface	Cook (2018b)
Surface energy balance system/solar infrared radiation station	sgpqcrad1longE13.c1 and sgpqcrad1longE13.c2	Best estimate of longwave, shortwave, and normal radiation (W m^{-2})	2 m a.g.l.	Cook and Sullivan (2019)
Doppler lidar	sgpdlfptC1.b1	Range-corrected attenuated backscatter variance (m^{-1} sr^{-1}), SNR variance (dB), and average eddy dissipation rate (m^2 s^{-3})	90 to 800 m a.g.l.	Champagne et al. (1977), Newsom and Krishnamurthy (2020)
	sgpdlprofwstats4newsC1.c1	Cloud base height (m) CBL depth from Tucker method (m) Year, month, and hour of day	0 to 9000 m a.g.l.	Newsom et al. (2019b), Tucker et al. (2009)
	sgpdlprofwinds4newsC1.c1	Wind shear exponent $\left(\alpha = \log_{10}\left(\frac{U_1}{U_2}\right)/\log_{10}\left(\frac{Z_1}{Z_2}\right)\right)$, where U_i and Z_i are wind speed and height at altitude i	$z_1 = 90$ m to $z_2 = 300$ m a.g.l. (or lower, depending on data availability)	Newsom et al. (2019a), Wharton and Lundquist (2012)

Methodology

Use a supervised learning approach to estimate Doppler lidar boundary layer height estimates at SGP C1 from surface meteorological & Doppler lidar primary/secondary variables





Sample RF Z_i tree

Results

- Performance
  - Overall bias & RMSE is less for the RF model
  - Daytime – clear sky and cloudy conditions are performing better with RF
  - Small underestimation in RF z_i is observed
- Nighttime z_i estimates are reasonably correlated but since are typically constant at SGP, correlations are poor.

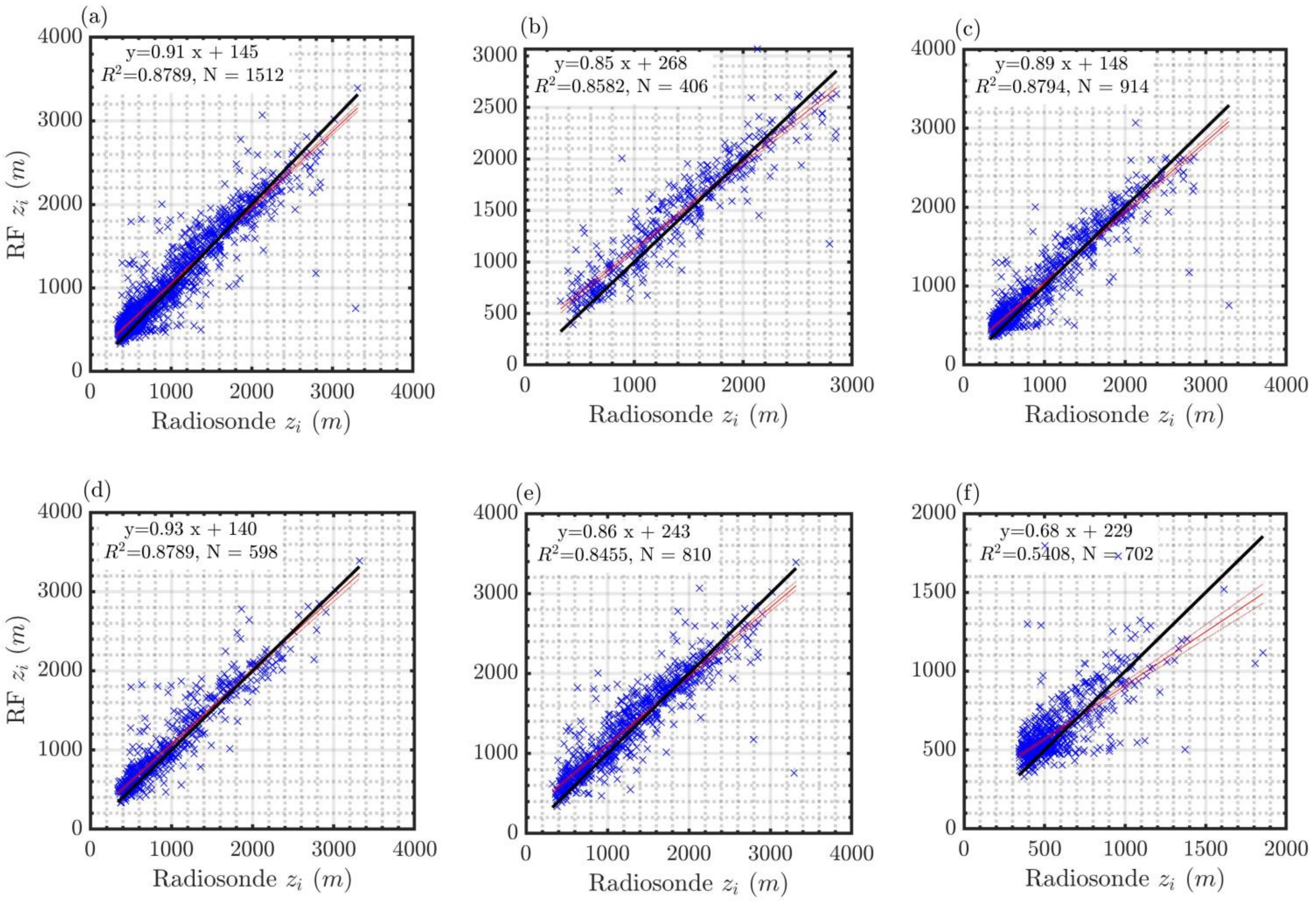


Figure: Correlations between RF PBL height and radiosonde PBL height for (a) all data in 2019, (b) daytime clear sky, (c) clear-sky daytime and nighttime, (d) cloudy daytime and nighttime, (e) daytime only, and (f) nighttime only.

Observed atmospheric conditions	MAE (m) RF	MAE (m) Tucker method	MAE (m) % improvement	RMSE (m) RF	RMSE (m) Tucker method	RMSE (m) % improvement	R^2 RF	R^2 Tucker method
Daytime only	167	311	46 %	249	441	43 %	0.845	0.545
Daytime clear sky	165	336	51 %	235	479	51 %	0.857	0.520
Daytime cloudy	141	255	45 %	208	363	43 %	0.878	0.725

Table: Systematic mean absolute errors, root-mean-square error, and correlation coefficient (R2) between RF, Tucker method, and radiosonde z_i estimates in 2019.

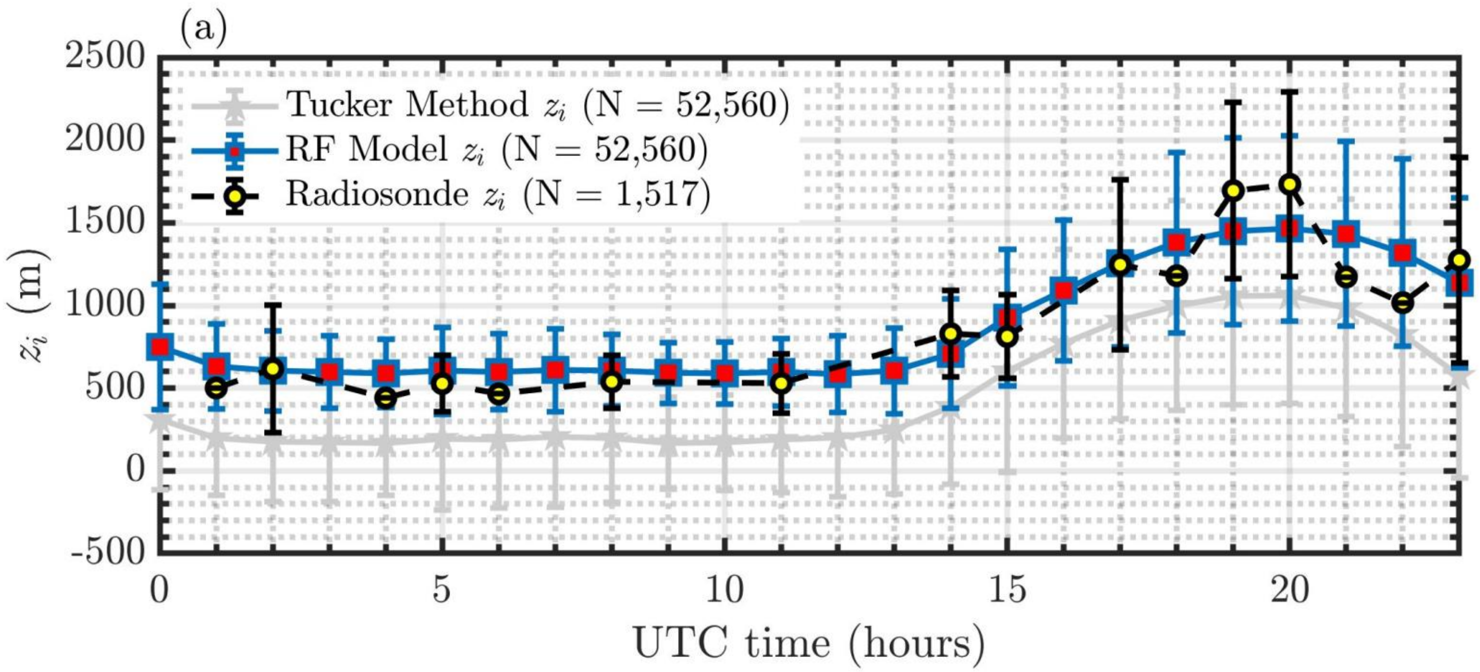
Results

- Lidar samples → 52,560
  Radiosondes → 1,517

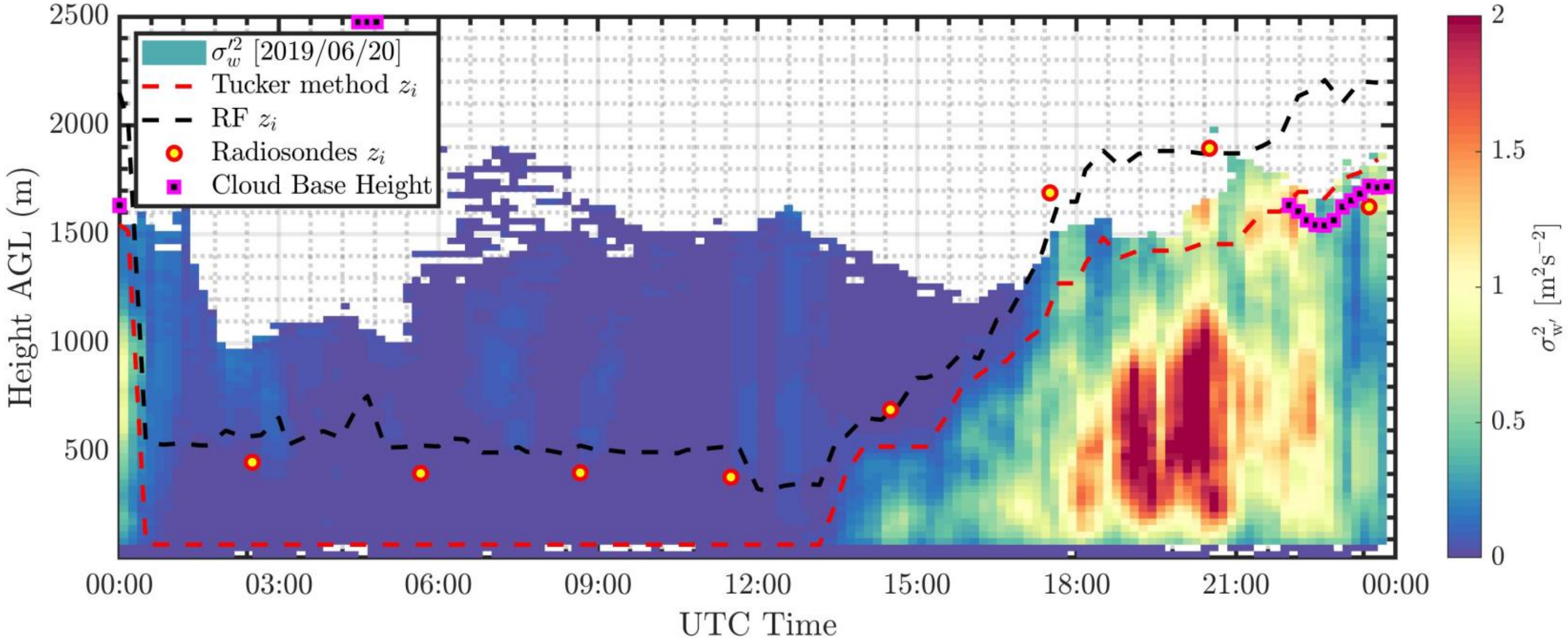
- The transition and daytime conditions are well captured by the RF model

- Nocturnal conditions are constant with little variability

- Time series RF z_i shows good agreement with radiosonde



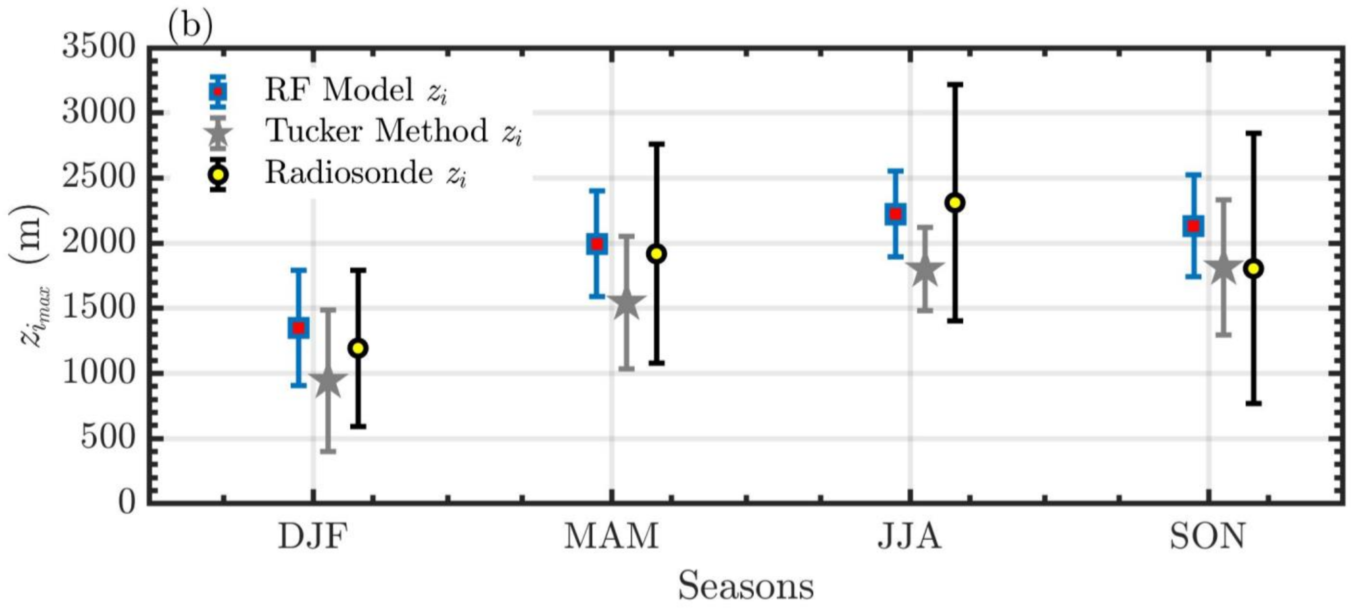
Hourly averaged z_i estimates at the SGP central facility for 2019 from RF, the Tucker method, and radiosondes. Total number of samples (N) for each dataset is also shown in the legend. The bars in both plots represent 1 standard deviation.



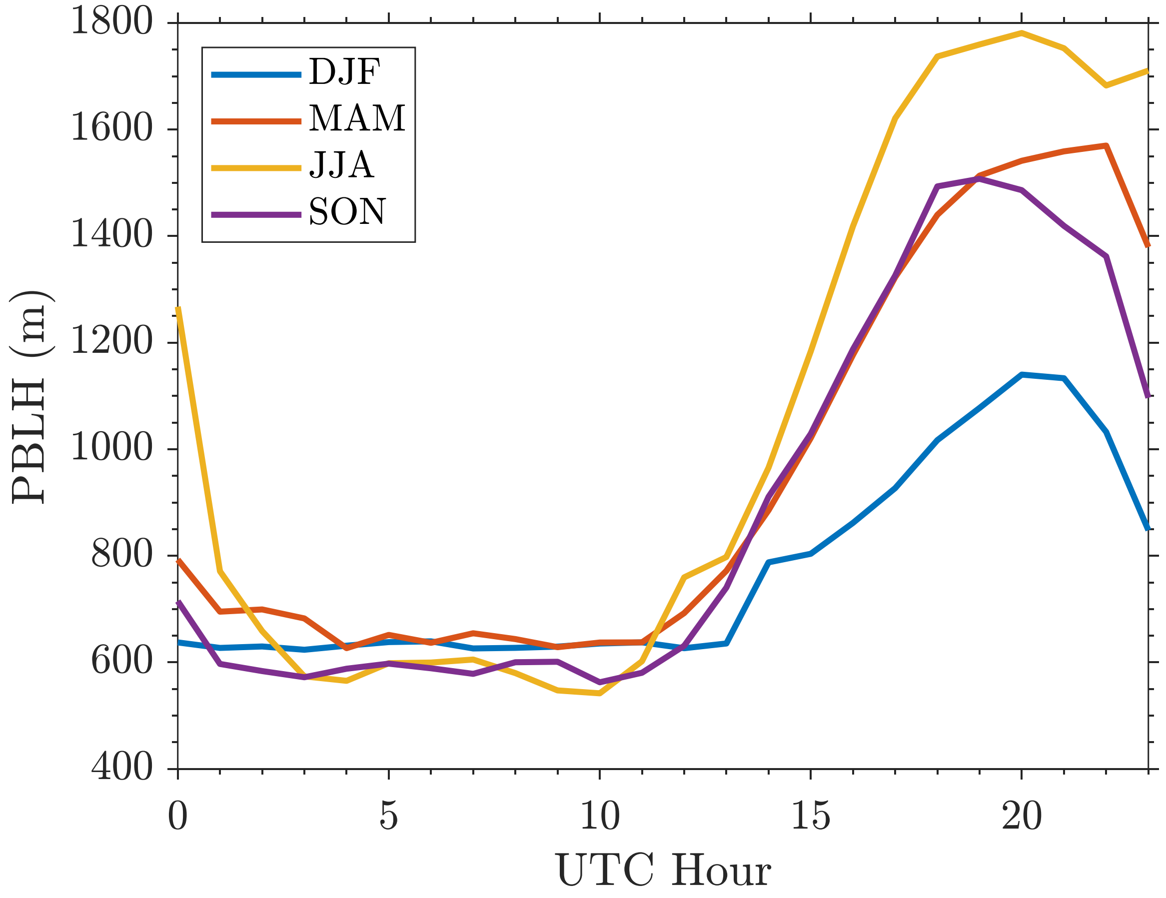
Boundary layer height estimates at the SGP central facility on 20 June 2019 from the Tucker method (Tucker et al., 2009), RF model z_i, radiosondes z_i (Sivaraman et al., 2013), cloud base height estimates from lidar (Newsom et al., 2019b), and the background colours represent vertical velocity variance measurements from Doppler lidar.

Seasonal variability

- Summer observes largest median daily maximum PBLHs
  - Transition periods are steeper with quiescent nocturnal conditions
- Higher nocturnal PBLHs are observed during Spring (storms & nocturnal convection)



Daily daytime maximum z_i estimates for four seasons (DJF, MAM, JJA, SON) from RF, radiosonde, and the Tucker method.



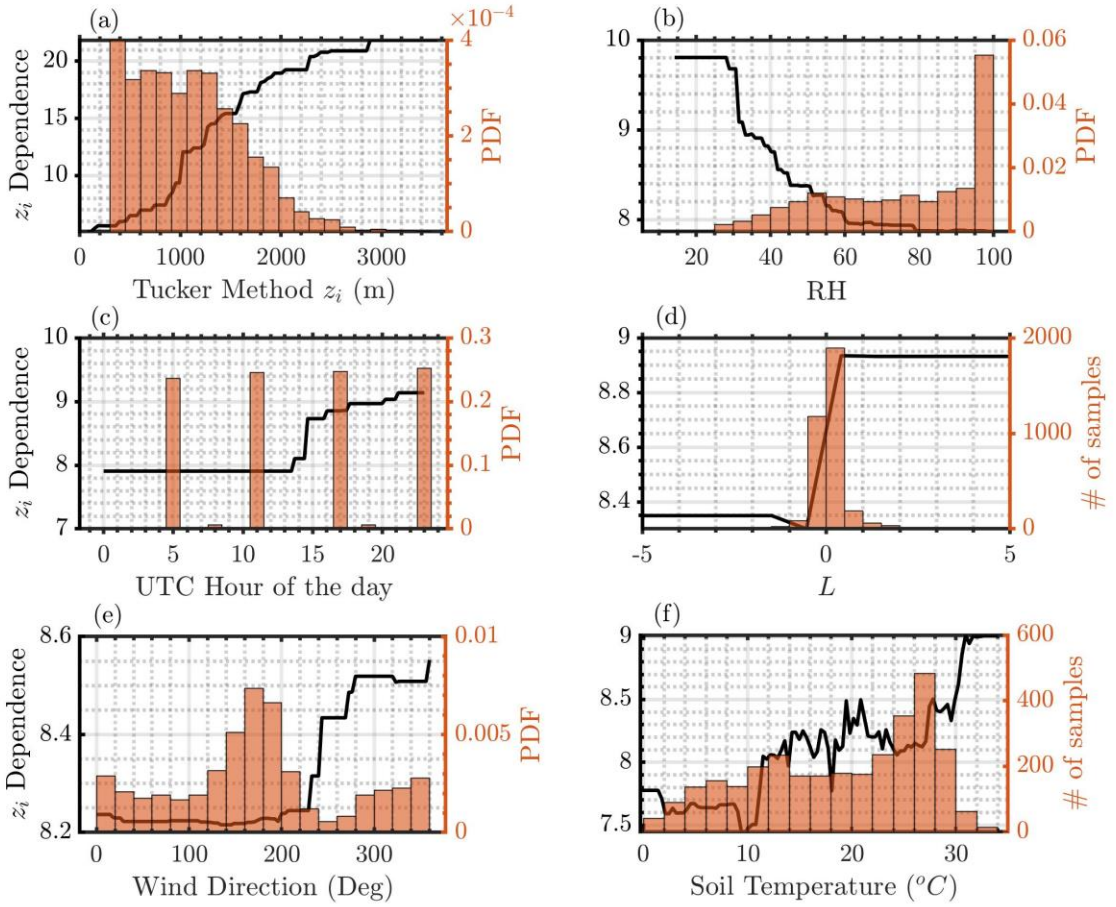
Seasonal z_i estimates for four seasons (DJF, MAM, JJA, SON) from RF method.

Predictor Importance

- Based on RF model inputs, the predictor importance estimates provide ranking to variables that are important in predicting PBLH
- Daytime and night-time have different variables deemed important (for example night-time, Obukhov length has shown to have the highest importance)

Parameters/features	% importance
Tucker method z_i	58.67 %
Hour of the day	10.05 %
Surface relative humidity	6.82 %
Attenuated backscatter	2.90 %
Surface air temperature	2.77 %
Monin–Obukhov length	2.77 %
Soil temperature	1.92 %
Surface wind direction	1.78 %
Turbulence kinetic energy	1.32 %
Others	< 11 %

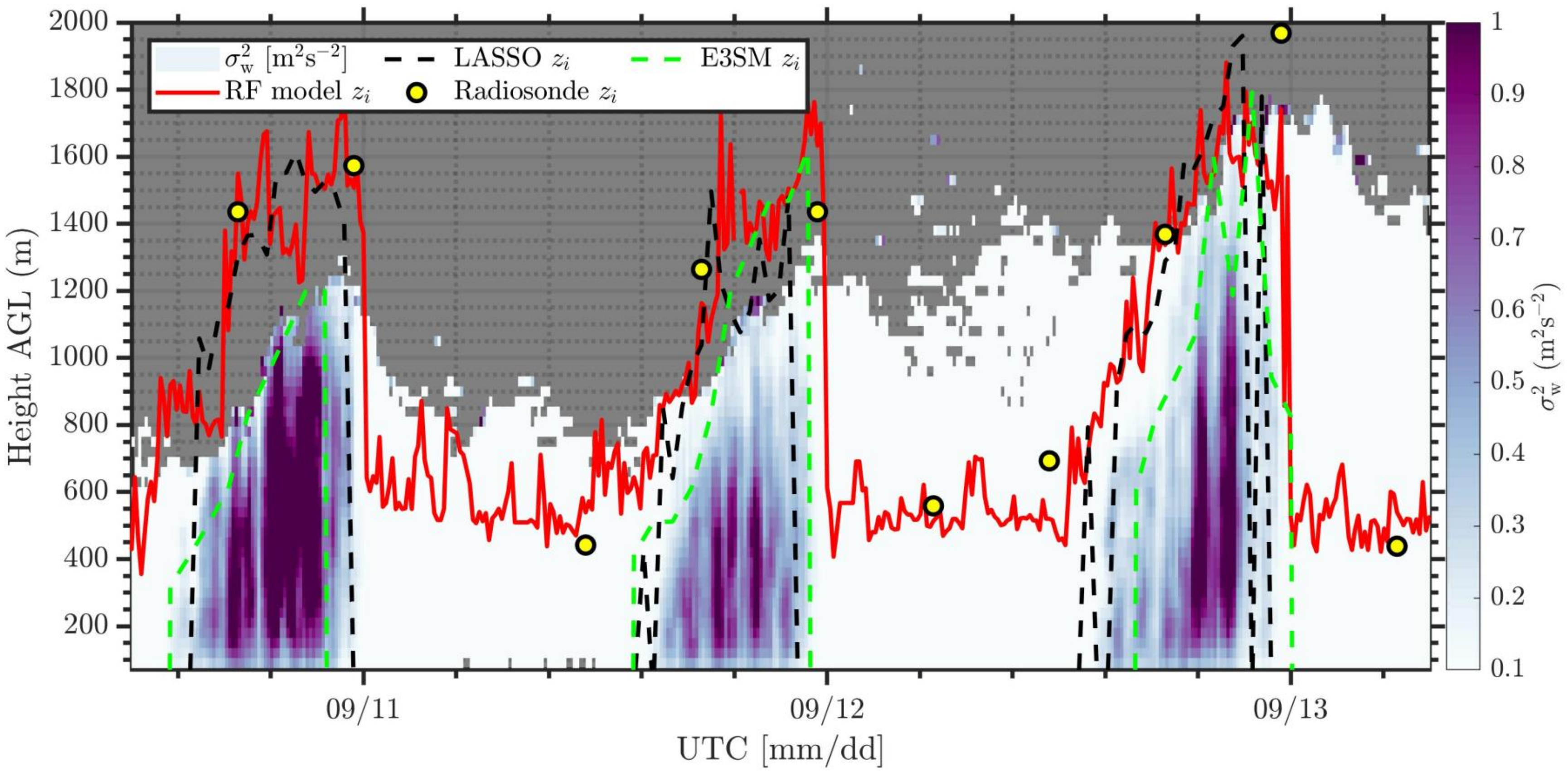
Key parameter/feature unbiased importance estimates during all conditions.



RF partial dependence during all conditions from (a) the Tucker method z_i , (b) relative humidity, (c) hour of the day, (d) Monin–Obukhov length, (e) surface wind direction, and (f) soil temperature to boundary layer height at the central facility. High dependence shows more sensitivity of the RF model to the bin of feature values.

Preliminary model comparison

- The RF model z_i compares better with LASSO type models, although the decay of turbulence is not well characterized (see Larry Berg's poster)
- E3SM type models performs well but under-estimate the z_i *(mostly due to the model resolution)*



Vertical velocity variance (σ^2_w) estimates from Doppler lidar for 3 days (10, 11, and 12 September 2016) with z_i estimates from **(a)** the RF model (red solid line), **(b)** radiosondes (yellow circles), **(c)** the LASSO model (black dashed line), and **(d)** the E3SM model (green dashed line).

Pacific Northwest NATIONAL LABORATORY

Summary

- Overall, the RF model estimates the convective boundary layer height with good accuracy (R^2 > 85%, MAE < 160 m)

- RF model show reduction in MAE by more than 50% compared to standard lidar outputs, in certain atmospheric conditions
  - RF model bias corrects night-time PBLH estimates, but the correlations are still poor. Part of future work.

- Predictor importance provides insight on the key atmospheric variables affecting boundary layer height at SGP C1

- Inter-comparison with models show that the LASSO models compare well with RF z_i estimates, but transition time periods are not well captured

Paper reference: Krishnamurthy et al., 2021, Atmos. Meas. Tech., 14, 4403–4424
https://doi.org/10.5194/amt-14-4403-2021

https://engineering.arm.gov/~raghuvaidhya/sgpdlC1/RF_PBLH/

Thank you