International Journal of Applied Earth Observation and Geoinformation 130 (2024) 103917

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A physically informed multi-scale deep neural network for estimating foliar nitrogen concentration in vegetation

Mohammad Hossain Dehghan-Shoar [a,c,*], Gabor Kereszturi [a], Reddy R. Pullanagari [c,d], Alvaro A. Orsi [b], Ian J. Yule [c], James Hanly [a]

[a] School of Agriculture and Environment, Massey University, Tennent Drive, Palmerston North 4414, New Zealand
[b] Institute of Environmental Science and Research Limited (ESR), Mt Albert Science Centre, 120 Mt Albert Road, Auckland 1025, New Zealand
[c] Stoneleigh Consulting Limited, 70a Francis Road, Rd 2, Whakamarama 3172, New Zealand
[d] Australian Plant Phenomics Facility, The Plant Accelerator, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA 5064, Australia

ARTICLE INFO

Keywords:
Deep learning
Spectroscopy
Nitrogen concentration estimation
Radiative transfer models
Physically informed machine learning

ABSTRACT

This study introduces a Physically Informed Deep Neural Network (PINN) that leverages spectral data and Radiative Transfer Model insights to improve nitrogen concentration estimation in vegetation, addressing the complexities of physical processes. Utilizing a comprehensive spectroscopy dataset from various species across dry/ground (n = 2010), leaf (n = 1512), and canopy (n = 6007) scales, the study identifies 13 spectral bands key for chlorophyll and protein quantification. Key bands at 2276 nm, 755 nm, 1526 nm, 2243 nm, and 734 nm emerged vital for accurate N% prediction. The PINN outperforms partial least squares regression and standard deep neural networks, achieving an R^2 of 0.71 and an RMSE of 0.42 (%N) on an independent validation set. Results indicate dry/ground data performed best (R^2 = 0.9, RMSE = 0.24 %N), with leaf and canopy data showing lower efficacy (R^2 = 0.67, RMSE = 0.45 %N; R^2 = 0.65, RMSE = 0.46 %N, respectively). This multi-scale approach provides insights into spectral and N% relationships, enabling precise estimation across vegetation types and facilitating the development of transferable models. The PINN offers a new avenue for analyzing remote sensing data, demonstrating the significant potential for accurate, scale-spanning N% estimation in vegetation. Further enriching our analysis, the inclusion of seasonal data significantly enhanced our model's performance in field spectroscopy, with notable improvements observed across summer, spring, autumn, and winter. This adjustment underlines the model's increased accuracy and predictive capability at the field spectroscopy scale, emphasizing the vital role of integrating environmental factors, including climatic and physiological aspects, in future research.

1. Introduction

Grasslands are indispensable for the global economy and ecological sustainability, serving as a crucial source of essential resources like food, income, and biofuels. They play a pivotal role in mitigating climate change and providing critical ecosystem services such as soil conservation, water regulation, and biodiversity conservation, all of which are vital for human well-being (FAO, 2020; Taubert et al., 2012; Dehghan-Shoar, 2018; Petermann and Buzhdygan, 2021). Accurately estimating grassland nitrogen levels is paramount for effective management (Goetz and Whitehead, 1971). Notably, grasslands are a significant source of food for livestock, highlighting the importance of adopting sustainable management practices, including precise nitrogen level monitoring, to ensure a stable and sustainable global food supply (Ramankutty et al., 2008).

Assessing nitrogen concentration (N%) in grasslands faces various challenges, including labor-intensive processes, limited spatial coverage, intrusive sampling methods, accuracy concerns, and high costs. This has prompted interest in alternative monitoring methods (Pullanagari et al., 2021; Dehghan-Shoar et al., 2023a,b). Spectroscopy, applied at different scales like dry/ground vegetation, leaf, and canopy, has emerged as a valuable technique for estimating N% in vegetation (Mutanga, 2004). While this method can enhance performance, it often leads to empirical and data-driven models, which may lack generalizability. Despite achieving reasonable accuracy at individual scales,

* Corresponding author at: School of Agriculture and Environment, Massey University, Tennent Drive, Palmerston North 4414, New Zealand.
E-mail address: Mohammad.Dehghan.Shoar@massey.ac.nz (M.H. Dehghan-Shoar).

https://doi.org/10.1016/j.jag.2024.103917
Received 25 February 2024; Received in revised form 27 April 2024; Accepted 15 May 2024
Available online 28 May 2024

accurately estimating N% across multiple scales remains challenging due to diverse factors affecting spectral reflectance at each scale. For instance, measurements at the canopy scale, sensitive to energy-matter interactions, are crucial for determining vegetation biochemical and biophysical variables (Kokaly, 2001). However, accuracy at this scale can be influenced by variables like signal-to-noise ratio (SNR), sensor specifications, and environmental conditions such as soil exposure and water on leaves (Queally et al., 2022; Pullanagari et al., 2021).

To reliability monitoring N% across different scales, it is essential to consider factors like leaf area index (LAI), canopy-scale leaf angle, and internal leaf structure and water content at leaf and ground levels (Féret et al., 2021; Verhoef, 1984). This multiscale approach represents a significant advancement over traditional methods reliant on soil or plant tissue sampling and subsequent laboratory analysis. It offers improved performance and generalizability, capturing nuanced variations in nitrogen levels across diverse grasslands. Additionally, it enables more precise assessments and can be scaled up, providing a comprehensive understanding of grassland nitrogen distribution. Compared to traditional sampling and laboratory testing, this method is more efficient and cost-effective, facilitating timely decision-making critical for responsive environmental management and agricultural practices (Dehghan-Shoar et al., 2023a).

Models estimating N% at multiple scales, including leaf, canopy, and ground, are crucial for effective grassland nitrogen monitoring. This approach provides a nuanced understanding of nitrogen distribution and addresses the complexities of diverse grassland ecosystems. By integrating these scales, this strategy fills a critical research gap, offering a comprehensive solution for nitrogen monitoring. Each scale presents unique challenges. Canopy scale data, offering extensive area coverage, are influenced by factors like atmospheric conditions and canopy structure (Verhoef, 1984; Queally et al., 2022). Although insightful for understanding plant responses to varying N%, leaf scale studies are labor-intensive and less suited for whole ecosystems (Féret et al., 2021). Dry/ground scale data, known for accurate N% measurements, face limitations due to potential sample preparation errors, the need for specialized equipment, and cost (Féret et al., 2021). Balancing cost and accuracy is the key challenge in remote sensing agriculture and environmental monitoring applications. Addressing these complexities necessitates the development of robust models that integrate physical principles, aiming to improve vegetation mapping for agricultural and environmental monitoring.

To enhance the accuracy of N% estimation in vegetation, we have devised a method that melds Physics-Informed Neural Networks (PINN) with Radiative Transfer Modeling (RTM) (Raissi et al., 2019). This innovative approach not only adheres to physical laws but also integrates vital biochemical and biophysical factors influencing plant growth. Our methodology underscores the significance of data directly linked to vegetation development, elucidating its pivotal role in precision agriculture. This is particularly evident in discerning key spectral features across various vegetation scales—such as at the dry/ground level, leaf, and canopy (Chen et al., 2022; Pullanagari et al., 2021). Further research bolsters our approach, highlighting the substantial benefits of merging spectral, structural, and textural data with traditional spectral reflectance methods. This amalgamation proves particularly effective in surmounting challenges commonly encountered with RTM, thereby enhancing the estimation of vegetation parameters (Ishaq et al., 2023; Dhakal et al., 2023; Dehghan-Shoar et al., 2023a,c; Pullanagari et al., 2021).

The adoption of a multi-dimensional strategy enhances the analysis of plant nutrients. For instance, studies involving unmanned aerial vehicles (UAVs) demonstrate that incorporating multispectral imaging with machine learning algorithms – factoring in canopy features – results in more accurate biomass estimation (Dhakal et al., 2023). This underscores the advantages of employing diverse data types for precise plant analysis (Camps-Valls et al., 2023; Ishaq et al., 2023). Furthermore, the fusion of spectral data with other physical information proves advantageous in monitoring vegetation and estimating forage biomass, thus accentuating the value of integrating varied data types in plant studies (Pullanagari et al., 2018; Chen et al., 2021). Emphasizing the importance of incorporating a wide array of information further bolsters our comprehensive approach to understanding plant behavior (Pullanagari et al., 2018; Ishaq et al., 2023). Through the application of PINN alongside RTM, our technique augments N% estimation accuracy in vegetation while mitigating the ill-posedness typically associated with RTM. This method addresses the complexities associated with N% estimation through a holistic and integrated strategy, meeting the demands of contemporary agricultural and environmental research (Camps-Valls et al., 2023). It underscores the essential role of incorporating a broad spectrum of physically-based data, thereby enhancing the precision and depth of vegetation nutrient monitoring and affirming the significance of PINN in advancing precision agriculture.

2. Material and methods

2.1. Canopy scale measurements

Between 2009 and 2016, we surveyed 17 dairy and hill country farms across New Zealand's South and North Islands. These farms were selected for their exposure to New Zealand's temperate climate, which allows for year-round grass growth moderated by annual precipitation ranging from 600 mm to 1600 mm. For detailed field campaign information and dataset specifics, refer to Pullanagari et al. (2021) and Dehghan-Shoar et al. (2023a). Spectral data were collected from canopies of seven principal grass species during spring (n = 2940), autumn (n = 2005), summer (n = 854), and winter (n = 208). We used ASD FieldSpec 4 Hi-Res and ASD FieldSpec PRO spectroradiometers with 512-element silicon photodiodes, offering high-resolution Visible Near Infrared (VNIR) (350–1000 nm) and Shortwave Infrared (SWIR) (1000–2500 nm) to collect spectral measurements. The Canopy Pasture Probe (CAPP) facilitated grassland spectral data capture, paired with a 50-W tungsten-quartz-halogen bulb for consistent illumination across varying weather conditions, calibrated against a white reference for optimal accuracy (Sanches et al., 2009). Stratified random sampling accounted for environmental factors like slope and elevation, with 20 spectral readings per 0.5 m^2 area to account for grassland diversity and reduce noise. Grass samples were collected, preserved, and geolocated using an RTK GPS for N% analysis by the Kjeldahl method and processed at the Analytical Research laboratory, Napier, New Zealand, involving sample drying, grinding, and homogenization for analysis with a QuikChem® 8500 analyzer.

2.2. Dry/ground scaled measurements

In 2014, a study on six New Zealand hill country and dairy farms assessed the spectral reflectance of grass samples, collecting 2010 samples for in-situ and laboratory analysis. The samples were dried, ground, and homogenized to ensure consistency. Spectral reflectance measurements were conducted using an ASD FieldSpec PRO spectroradiometer with a contact probe, taking 20 readings per sample to reduce noise and enhance data reliability. The instrument was calibrated with Spectralon™ standards to ensure accuracy. This approach aimed to estimate N% in the grass, following methodologies from previous research (Dehghan-Shoar et al., 2023c,a), capturing variability within each sample through comprehensive spectral analysis.

2.3. Leaf scale measurements

In our study, we leveraged two paramount leaf scale datasets: the LOPEX93 dataset (n = 304 spectral reflectance samples) from the

Leaf Optical Properties Experiment Database (Hodáňová, 1985) and leaf spectra of nitrogen-stressed maize (n = 1210) from the Ecological Spectral Information System (EcoSIS) (Grzybowski et al., 2021; Ge et al., 2019), which span a wide array of plant species and conditions. The LOPEX93 dataset, developed at the Joint Research Centre in Ispra, Italy, encompasses a comprehensive collection of biophysical, biochemical, and spectral data for over 50 plant species. This collection includes data on fresh and dried leaves and conifer needles. It stems, from spectral data acquisition conducted via a Perkin Elmer Lambda 19 spectrophotometer, ensuring high spectral resolution across the 400–2500 nm wavelengths. The resolution ranged from 1–2 nm in the VNIR (Visible and Near Infrared) spectrum to 4–5 nm in the SWIR (Short Wave Infrared) spectrum. This dataset notably includes measurements of reflectance and transmittance, along with actual nitrogen percentages determined through the Kjeldahl method. Conversely, the maize dataset (Grzybowski et al., 2021; Ge et al., 2019) focuses on spectra from maize leaves subjected to nitrogen stress, recorded using a FieldSpec 4 spectroradiometer across the 350–2500 nm wavelengths. In this dataset, nitrogen levels were quantified employing the Dumas method, facilitated by a LECO FP428 analyzer. Many samples in the leaf dataset are accompanied by leaf nitrogen content per leaf dry mass (% DW) values. To ensure measurements were consistent and comparable across different environments or conditions, we selectively harmonized the data from the dry/ground, canopy, and leaf scales, aligning them based on the samples' total mass or volume as necessary. This harmonization was particularly important in scenarios where the nitrogen measurement units varied, ensuring our analysis maintained a uniform approach. In cases where the data was already consistent or the measurement discrepancies did not affect the analysis, we opted not to harmonize, preserving the original data integrity. This selective process of alignment allowed us to integrate biomass and dry weight information effectively, ensuring leaf scale datasets corresponded accurately with those from dry/ground and canopy levels. Our harmonization also involved processing reflectance data from each dataset, where we conducted preprocessing to exclude outliers in spectral signatures or samples with missing nitrogen values. This careful preprocessing safeguarded the data's reliability, essential for incorporating a wide array of spectral and biochemical data in a thorough examination of our leaf scale dataset. After this meticulous preparation, we had a total of 1512 samples ready for our study.

2.4. Inversion of biophysical and biochemical variables

We employed the SAIL and PROSPECT-PRO RTM models, integrating dry/ground, leaf, and canopy scale data, to accurately estimate key biophysical and biochemical variables such as LAI, Protein Content (Cp), and Chlorophyll Content (Cab). These variables are pivotal for assessing vegetation health and growth dynamics. PROSAIL, a merger of the SAIL and PROSPECT-PRO models, facilitates the precise inversion of these variables by modeling canopy reflectance and vegetation biochemical properties (Féret et al., 2021). SAIL, introduced by Verhoef (1984), simulates canopy reflectance, accounting for light-plant interactions, while PROSPECT-PRO, initially conceptualized by Jacquemoud and Baret (1990) and later refined by Féret et al. (2021), is renowned for its effectiveness in biochemical estimation at leaf scale. The inversion process involved extensive forward and inverse RTM simulations in minimizing discrepancies between observed and modeled spectra. This optimization utilized the Affine Invariant Markov Chain Monte Carlo (MCMC) method, supported by 1000 walkers and 50,000 simulations, leveraging the Python EMCEE ensemble toolkit for its robustness and precision in error minimization (Foreman-Mackey et al., 2019). We refined our inversion methodology for enhanced accuracy by fine-tuning key variables within the PROSAIL model, such as viewing and illumination angles, as detailed in Table 2. We calculated the mean and standard deviation for variables of each sample across three scales: dry/ground (n = 2010), leaf (n = 1512), and canopy (n = 6007), thus compiling a comprehensive dataset comprising n = 9529 samples (Table 1).

Table 1
Sample size and summary statistics for nitrogen concentration (N%) across different scales and combined.

Dataset	Max	Mean	Min	Samples	Std Dev
Dry/ground scale	5.11	2.85	0.97	2010	0.75
Leaf scale	6.00	2.61	0.68	1512	0.81
Canopy scale	5.32	2.54	0.73	6007	0.79
Combined	6.00	2.62	0.68	9529	0.80

Table 2
The search space for PROSAIL variables was optimized using MCMC for inversion, setting the Equivalent Water Thickness (Cw) to zero for dry/ground samples to enhance performance while leaving Cw unconstrained for leaf and canopy data.

Variable name	Min value	Max value	Abbreviation	Units
Structure variable	1	3	N	
Chlorophyll content	1	100	Cab	cm2.microg-1
Carotenoid content	1	30	Car	cm2.microg-1
Brown pigment content	0	0.2	Cbrown	
Equivalent water thickness	0	0.02	Cw	cm-1
Dry matter content	0	0.02	Cm	cm2.g-1
Protein content	0	0.02	Cp	cm2.g-1
Carbon based constituents	0	0.02	CBC	cm2.g-1
Anthocyanins content	1	10	Anth	cm2.nmol-1
Leaf area index	1	3	LAI	
Leaf angle distribution a	10	90	LIDFa	
Leaf angle distribution b	10	90	LIDFb	
Background reflectance	–	–	RSOIL	
Dry/Wet soil factor	–	–	PSOIL	
Hotspot parameter	0	1	HOTSPOT	
Solar zenith angle	0	0	SZA	Degrees
Observer zenith angle	20	20	VZA	Degrees
Relative azimuth angle	160	160	RAA	Degrees

For dry/ground and leaf scale data, we utilized the PROSPECT-PRO model, renowned for its accurate simulation of optical properties including the structure variable, chlorophyll content, carotenoid content, brown pigment content, equivalent water thickness, dry matter content, protein content, carbon-based constituents, and anthocyanins content. For canopy scale data, we employed the PROSAIL model, which integrates both leaf and canopy structural effects alongside leaf scale properties such as LAI, leaf angle distribution parameters (a and b), and the hotspot parameter (Fig. 1). To guarantee data independence and minimize sample overlap, we meticulously divided the dataset into independent calibration (n = 7623) and validation (n = 1906) subsets, ensuring that samples for canopy and dry/ground scales were collected from distinct spatial and temporal locations. Specifically, the calibration subset included dry/ground (1608 samples), LOPEX (244 samples), Maize stress (966 samples), and canopy data (4805 samples). The validation subset consisted of dry/ground (402 samples), LOPEX (60 samples), Maize stress (242 samples), and canopy data (1202 samples). This strategic division, underlining spatial and temporal separation, facilitates a detailed and unbiased evaluation of our models across various vegetative states and scales.

2.5. Physically based equation for estimation of N%

In our study, we formulated a physically-based equation for accurately estimating N% by focusing on spectral regions sensitive to Cab within the 410–780 nm range and Cp beyond 1400 nm, identifying these regions as critical for precise N% predictions (Féret et al., 2021). To optimize our model, we implemented Genetic Feature Selection (GA), a process that successfully pinpointed 13 vital spectral bands intimately linked with the presence of Cab and Cp (Jarvis and Goodacre, 2005). This selection was crucial for refining our approach, as GA methodically sifted through a vast dataset to identify the most informative spectral bands, significantly enhancing the predictive

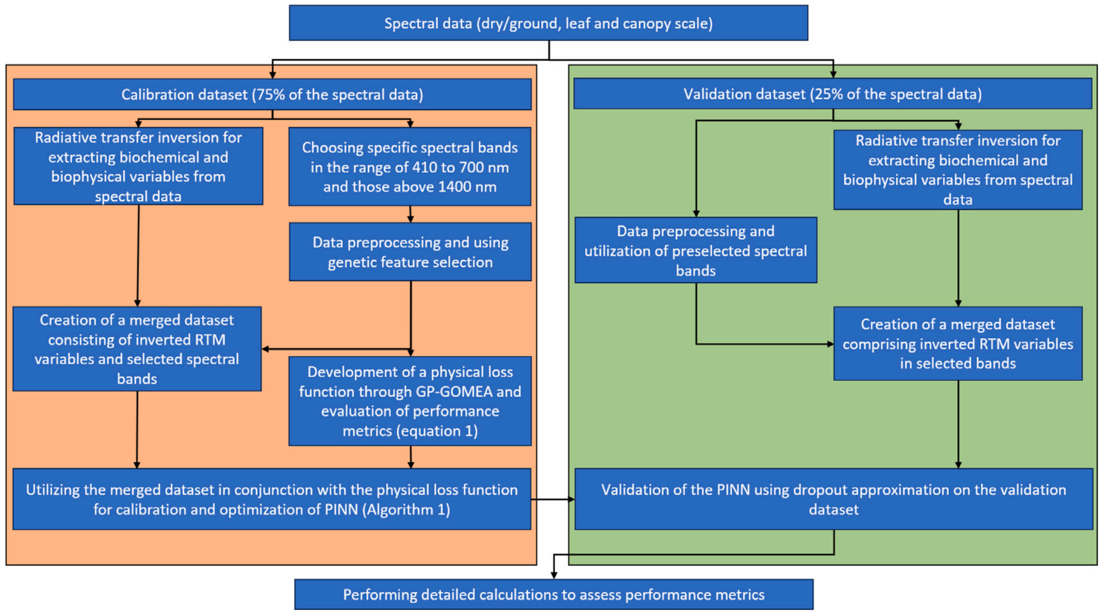


Fig. 1. Flowchart depicting the methodology used in this study to predict nitrogen content (N%) employing Physics-Informed Neural Networks (PINN). Spectral bands chosen through Genetic Feature Selection (GA) and inverted variables from leaf levels at dry/ground, leaf, and canopy scales served as input for the PINN model.

model's accuracy, especially in applications like remote sensing or spectroscopy (Dehghan-Shoar et al., 2023a). GA refined spectral band selection for N% estimation in vegetation by simulating natural evolutionary processes. Initially, GA generates a diverse array of potential solutions, each representing a different combination of spectral bands. Solutions are assessed for their predictive accuracy of N%, with the most effective ones selected for reproduction. During the crossover phase, these selected solutions exchange portions of their genetic code, creating new combinations that potentially enhance predictive performance. Mutation introduces random variations, ensuring diversity and aiding in the discovery of novel, effective band combinations. This cycle of selection, crossover, and mutation repeats, gradually honing in on a highly efficient set of spectral bands. The iterative, evolutionary nature of GA allows it to dynamically explore and optimize the selection space, effectively narrowing down the vast array of possible combinations to those most critical for accurate N% estimation. This process showcases GA's ability to focus on the most relevant spectral information, markedly improving the model's efficiency and accuracy in complex predictive modeling tasks.

In our work, we leveraged the Gene-pool Optimal Mixing Evolutionary Algorithm for Genetic Programming (GP-GOMEA) to enhance N% estimation across varied vegetative conditions such as dry/ground, leaf, and canopy levels (Virgolin et al., 2021). GP-GOMEA excels by efficiently merging segments of pre-existing solutions, significantly elevating the fitness of successive generations. This is achieved through a comprehensive analysis of interactions between solution components, directing the algorithm towards the most viable solutions. Importantly, GP-GOMEA is adept at estimating interdependencies among model components, employing this knowledge to execute crossover operations of interdependent components en bloc (Virgolin et al., 2020). This approach ensures the preservation of their collective efficacy, a critical feature for solving complex problems that demand integrated solutions. GP-GOMEA is notably skilled at deriving models in the form of compact symbolic expressions, which are often interpretable, providing an added layer of understanding. By applying GP-GOMEA, we developed a precise model that selectively utilized 10 of the initially identified 13 key spectral bands associated with crucial plant elements – protein and chlorophyll – essential for accurate N% estimation. The algorithm's decision to select 10 features reflects a strategic focus on maximizing model efficiency and information utility, by sidelining features with limited additional value or those contributing to multicollinearity, thereby improving the model's predictive accuracy and generalizability. This process not only illustrates GP-GOMEA's capacity to pinpoint and prioritize the most impactful variables for complex estimations but also highlights its significant methodological contribution to precision agriculture, emphasizing its ability to find models that balance simplicity with interpretability and effectiveness.

Table 3
Genetic algorithm and symbolic regression hyperparameters used.

Variable	GA variables	GP-GOMEA variables
Fitness function	MSE	MSE
Population size	100	64
Crossover probability	0.1	0.5
Mutation probability	0.05	0.5
Number of generations	10 000	No limit
Independent crossover probability	0.5	0.5
Independent mutation probability	50	0.5
Tournament size	0.5	–
Max tree height	–	6
Linear scaling	–	True
Function set	–	[+ , −, _p, / _sqrt, _plog]

Our methodology intricately weaves the selected spectral bands into a PINN, leveraging the absorption characteristics of Cab and Cp to enhance N% predictions. This integration serves as a crucial regularization factor within the network. Further elevating our model's clarity and analytical depth, we incorporate SHapley Additive exPlanations (SHAP) for a comprehensive feature importance analysis of the 13 GA selected bands (Lanitis et al., 2017). This inclusion of SHAP values illuminates the specific contributions of various spectral inputs towards the model's predictions, significantly boosting both the precision and the interpretability of our results. By meticulously setting hyperparameters for both GA and GP-GOMEA, as detailed in Table 3, we ensure a refined and detailed process for N% estimation, reflecting our dedication to achieving a model that is both accurate and transparent in its predictive capabilities.

Table 4
Summary of key variables utilized in the multi-scale PINN model for N% in vegetation.

Parameter	Selected value	Search space
Model architecture	Sequential	
Hidden layers	Relu	[Linear, ReLU, and Sigmoid]
Dropout layer	0.2	[0.1–0.5]
Output layer	Linear	
Optimizer	Adam	
Number of layers	3	[2–5]
Number of neurons	100	[50–300]
Loss functions	MSE and physical loss	
Loss weights	[0.2, 0.8]	
Batch size	10 000	
Epochs	20 000	

2.6. Creating the PINN

To estimate N%, our study utilized a PINN incorporating a physically based equation derived through the GP-GOMEA and optimized via TensorFlow across 10,000 iterations using Bayesian optimization (Schell, 2022; grantbuster et al., 2021). We have integrated the physical-loss function into the PINN by implementing Eq. (1) within a custom loss function built using TensorFlow. Additionally, we have introduced weighting factors for both the standard MSE and the physical loss, as proposed by grantbuster et al. (2021). The PINN model inputs include inverted RTM variables (N, Cab, Car, Cbrown, Cw, Cm, Cp, CBC, and Anth) and 13 spectral bands identified by GA. Bayesian optimization refined the PINN's architecture, improving performance through a Gaussian process model on hyperparameter efficiency (Table 4 and Fig. 8) (Bergstra et al., 2015). Additionally, dropout-based Bayesian approximation quantified the uncertainty in PINN predictions, creating a predictive distribution that assesses model uncertainty, thus offering insights into the model's prediction confidence levels (Junhwan et al., 2022) (Algorithm 1). The PINN's efficacy was verified using an independent dataset, evaluating its accuracy with metrics like Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and coefficient of determination (R^2), detailed in our methodology section. We further benchmarked the PINN by comparing its performance with several established regression methods, each with unique characteristics and applications. Partial Least Squares Regression (PLSR) is a technique that reduces the predictors to a smaller set of uncorrelated components and performs linear regression on these components, making it effective in handling situations where predictors are highly collinear. Support Vector Regression (SVR) works by mapping input features into high-dimensional space and finding the hyperplane that best fits the data with a margin of tolerance, ideal for complex, non-linear relationships (Kecman, 2005). Random Forest Regression (RF) utilizes an ensemble of decision trees to improve prediction accuracy and control over-fitting, making it versatile for various data types without the need for extensive parameter tuning (Liang et al., 2018). Gaussian Process Regression (GPR), on the other hand, is a probabilistic model that provides not only predictions but also the uncertainty of those predictions, making it suitable for cases where estimation of prediction uncertainty is crucial (Pullanagari et al., 2021). In addition, to evaluate the robustness of each model against bias and noise, we introduced heteroscedastic noise into calibration datasets spanning different scales. Random values (−0.45 to 0.45) were added to the original values, creating variable noise variance.

3. Results

3.1. Variables inversion

We applied MCMC numerical optimization to invert the RTM variables, enabling the assessment of uncertainties in dry/ground, leaf, and canopy data (Fig. 2). Dry/ground data inversion showed superior performance and reduced uncertainty compared to leaf and canopy data. Canopy scale data exhibited lower uncertainty in the visible spectrum than NIR and SWIR, which is attributed to the stronger influence of biophysical factors. However, canopy inversion demonstrated reduced performance compared to leaf data, highlighting the critical role of interleaf interactions.

Algorithm 1 Algorithm depicting PINN with Bayesian approximation using dropout.

1: **Initialize:** the model variables (weights and biases) and set the dropout rate
2: **Load:** the calibration data, physical values (N%), and RTM variables and spectral bands
3: **Define:** the physical loss function, which calculates the difference between the network's outputs and physical values
4: **for** each calibration sample **do**
5: **Input:** the current data (spectral bands and RTM variables) into the PINN
6: **Apply Dropout :** randomly drop out units during forward Propagation with the specified rate
7: **Forward Propagation :** propagate the inputs through the network to calculate the outputs
8: **Calculate Loss :** calculate the physical loss and mean squared error
9: **Backpropagation:** use the physical loss and mean squared error to update the model variables
10: **end for**
11: **Make Predictions:** use the trained model to make predictions on new data by turning off the Dropout for prediction
12: **Bayesian Approximation :** use Monte Carlo sampling to obtain posterior distributions of the model variables, given the calibration data and physical loss function

3.2. Physically based equation

We utilized GA for feature selection, a crucial step that helped us identify the fundamental bands correlating with Cab and Cp. We then applied GP-GOMEA to derive Eq. (1), allowing us to integrate a physical loss function into the PINN. This function was designed to reduce the adverse effects of noise and bias (Fig. 3), and it is written as.

$$A = 1.27$$
$$B = 17.95$$
$$C = \log(569\,\mathrm{nm})$$
$$D = \sqrt{734\,\mathrm{nm}}$$
$$E = \sqrt{755\,\mathrm{nm}}$$
$$F = \log(2077\,\mathrm{nm})$$
$$G = \log\left(\frac{2276\,\mathrm{nm}}{2177\,\mathrm{nm}}\right)$$
$$H = \log(1559\,\mathrm{nm})$$
$$I = \frac{2276\,\mathrm{nm}}{2077\,\mathrm{nm}}$$
$$J = 467\,\mathrm{nm}$$

$$N\% = A - \frac{B \cdot (C \cdot \log(\frac{D}{E}))}{\left(\frac{E}{D}\right)^{\frac{1}{4}} \cdot F \cdot G \cdot \sqrt{\frac{E \cdot \log(D)}{J}} \cdot \log\left(\frac{\sqrt{H}}{\log(I)}\right)} \quad (1)$$

Our research identified key spectral bands at 2276 nm, 755 nm, 1526 nm, 2243 nm, 734 nm, 2177 nm, 2375 nm, 1559 nm, 569 nm,

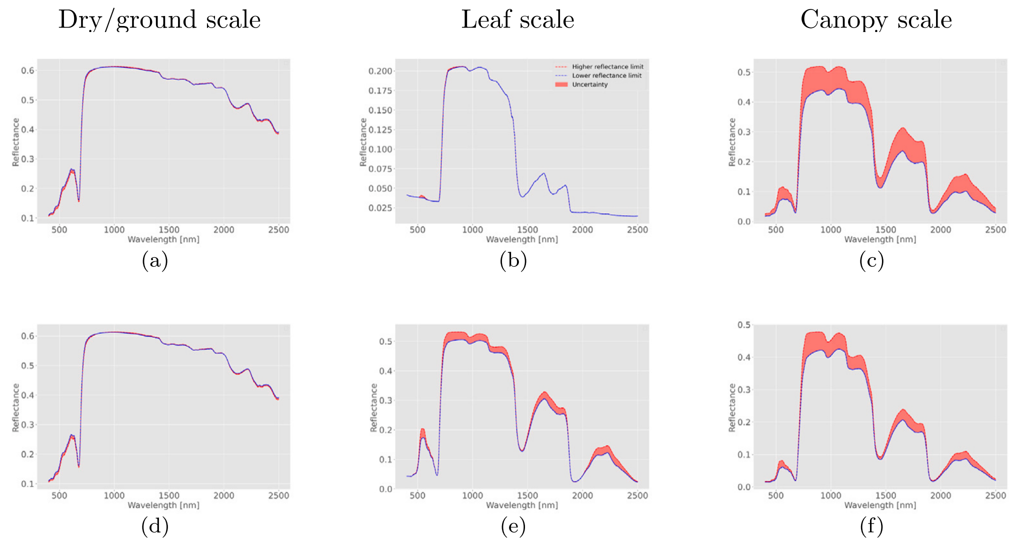


Fig. 2. Spectral band results from the inversion of PROSPECT-PRO and PROSAIL variables across dry/ground (a and d), leaf (b and e), and canopy (c and f) scales exhibit a trend of increasing uncertainty. This escalation is particularly significant from dry/ground to leaf, and even more so at the canopy scale, signaling issues related to the complexity of canopy structure and the inherent ill-posed nature of Radiative Transfer Models (RTM). These uncertainties underscore the challenges in accurately capturing the intricate interactions within canopy systems through spectral data.

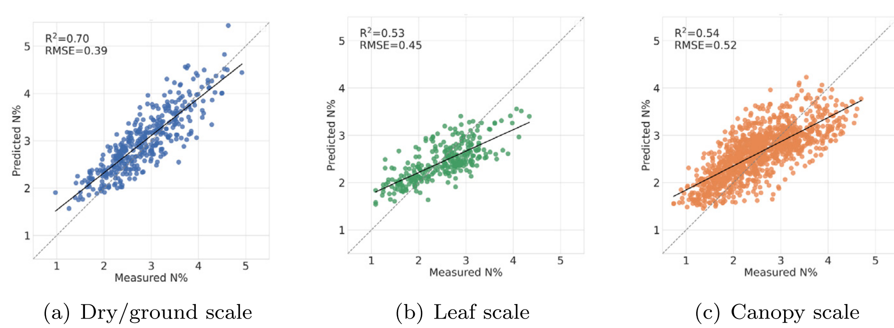


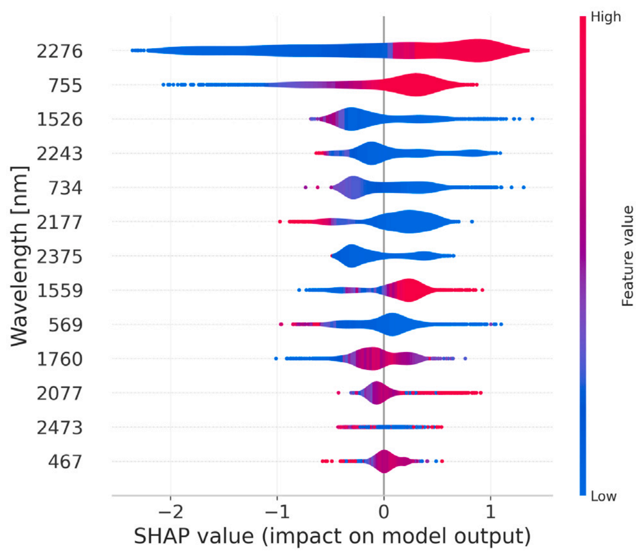
Fig. 3. Eq. (1), derived from GP-GOMEA with a physical loss function, effectively estimates N% across scales: dry/ground (panel a), leaf (panel b), and canopy (panel c). Its performance, influenced by key spectral bands from GA, demonstrates precision and adaptability in various grassland environments.

1760 nm, 2077 nm, 2473 nm, and 467 nm as crucial for estimating N% where most key spectral bands were identified in the SWIR (Fig. 4b). Using Eq. (1) and independent validation data (Fig. 3), we achieved 0.70 R^2 (RMSE = 0.39%N) for dry/ground, 0.53 R^2 (RMSE = 0.45%N) for leaf, and 0.54 R^2 (RMSE = 0.52%N) for canopy scale data. While these bands consistently influenced N% predictions across scales, their overall impact varied (Fig. 4a). Notably, bands at 2276 nm, 755 nm, 1526 nm, 2243 nm, and 734 nm were most influential for N% estimation across all scales (Fig. 4a). We observed an inverse correlation between reflectance at 1526 nm, 2243 nm, and 734 nm and increased N%, as detailed in Fig. 5. Significant non-linear relationships between N% and these spectral bands were evident (Fig. 5e, f, g, and j).

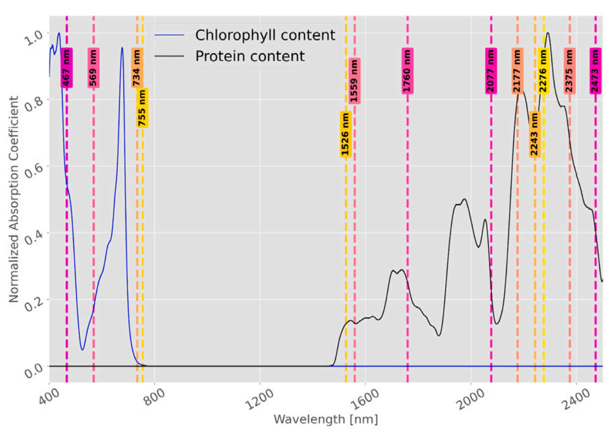
We created 13 dependence plots to scrutinize the interplay between pivotal spectral bands, as pinpointed by GA, and their correlation with N% content. This suite of plots vividly demonstrates how each chosen spectral band interacts with nitrogen levels within plant tissues, offering a deep dive into the predictive power of specific wavelengths (Fig. 5). Subplots (a) to (m) systematically explore these relationships, with particular emphasis on bands at 2276 nm and 755 nm (subplots c and d), marked as primary predictors of N%. The 755 nm wavelength, linked to Sun Induced Fluorescence (SIF), emerges as a significant factor in N% estimation, spotlighting the utility of SIF in agricultural monitoring (Mohammed et al., 2019).

Moreover, the 2276 nm band is distinguished by its high SHAP value, reinforcing its indispensable role in accurate N% prediction. Contrastingly, the 2375 nm and 2473 nm bands within the SWIR range, alongside the 467 nm band in the visible spectrum (subplots l, m, and a, respectively), exhibit lower SHAP values. This suggests a relatively lesser impact on N% estimation, though their inclusion provides a fuller spectral analysis narrative. Notably, some plots reveal U-shaped or parabolic trends, underscoring the non-linear dynamics between certain spectral bands and N%. This complexity is critical for understanding the multifaceted nature of spectral data analysis in agricultural contexts. Subplot (a) sets the stage by comparing 467 nm with 1760 nm, revealing distinct clusters that imply a strong dependency of spectral reflectance on nitrogen levels. Moving through the series, subplot (b) analyzes the 569 nm and 1760 nm wavelengths, suggesting an inverse relationship reflective of how vegetation absorption properties vary with nitrogen content across these frequencies.

Subplot (c) focuses on 734 nm and 755 nm, showcasing a dispersed data pattern that highlights the intricate relationship these wavelengths share with vegetation N%, pivotal for refining spectral



(a) Feature importance



(b) Spectral absorption

Fig. 4. Plot (a) shows a SHAP summary highlighting key spectral bands for N% estimation. Plot (b) reveals the sensitivity of these bands to protein and chlorophyll, crucial for the physical loss function and GA selection, with chlorophyll and protein absorption marked in blue and black, respectively. The absorption coefficients are generated using PROSPECT-PRO. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

analysis techniques. As we progress to subplots (d) through (m), each plot unfolds further layers of insight. For instance, subplots (d) and (e), which contrast 755 nm with 467 nm and 1526 nm with 2243 nm respectively, delve into the diverse ways N% influences plant reflectance, showcasing a range of correlation degrees and data dispersion
patterns.

3.3. Performance of PINN

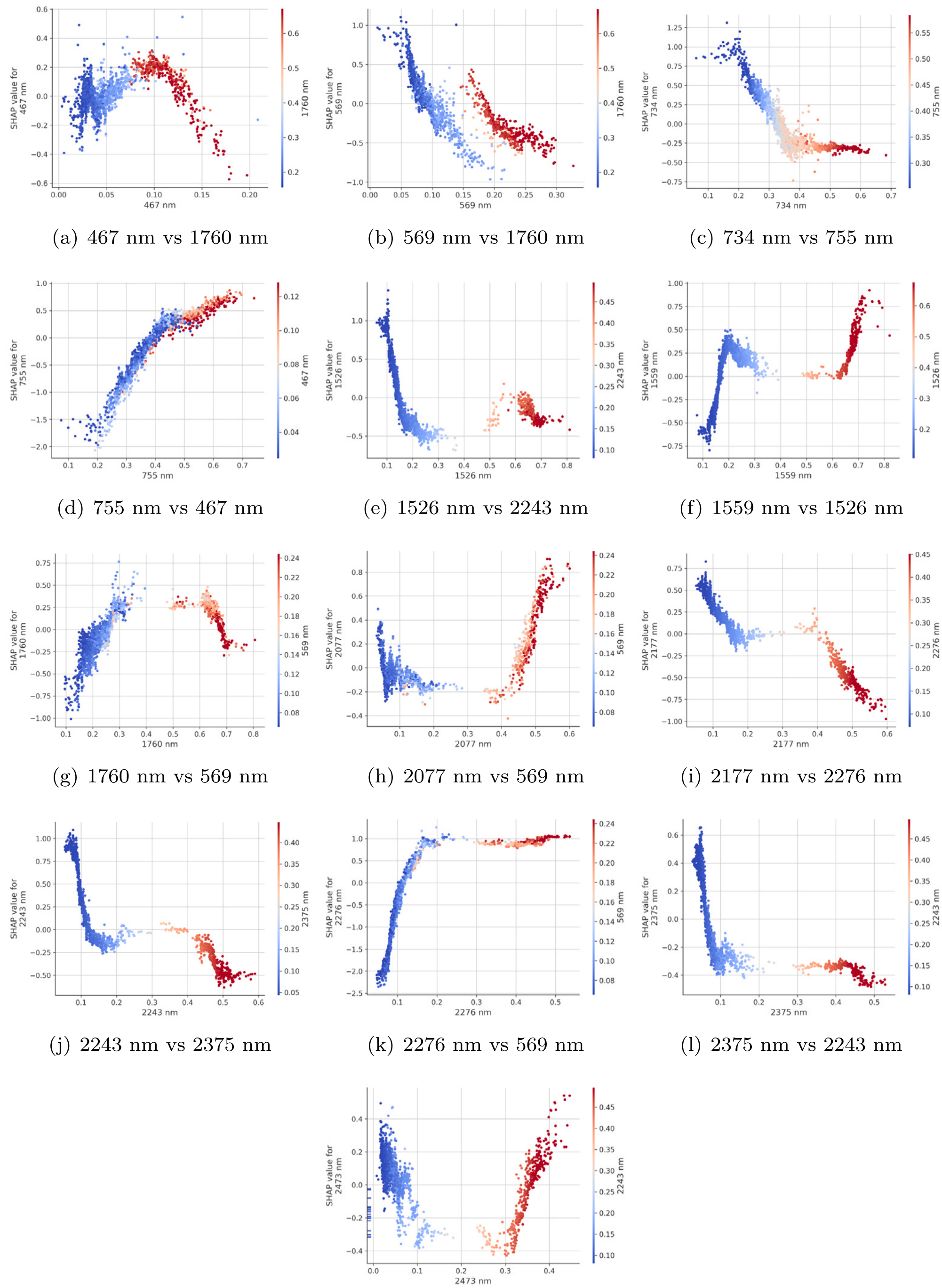
The PINN's performance was assessed with RTM-derived variables exclusively and in combination with important spectral bands. Critical RTM-derived variables for N% estimation were Cab, Cw, Cbc, and Cp. Integrating spectral bands with RTM features improved model validation (R^2 = 0.71, RMSE = 0.42 N%) compared to using RTM variables alone (R^2 = 0.57, RMSE = 0.51 N%) (Fig. 6). Including spectral bands reduced error and outliers, evidenced by a reduced Interquartile Range (IQR) of 1.5. A z-test ($p<0.05$) confirmed significantly lower uncertainty in models with spectral bands (Fig. 10).

In our study, we conducted a comprehensive analysis of PINN, PLSR, GPR, RF, and SVR performances, utilizing RTM variables and spectral bands across various scales (Fig. 6 Table 5. At the dry/ground scale, the PINN method utilizing solely RTM variables demonstrated a moderate correlation, with an R^2 of 0.61 and RMSE of 0.47 N%. This performance was significantly enhanced upon the inclusion of spectral bands, achieving an R^2 of 0.90 and RMSE of 0.24 N%. Notably, GPR closely matched PINN's performance with spectral bands, achieving an R^2 of 0.89 and RMSE of 0.23 N%, indicating its strong predictive capability in this context. Other methods like RF and SVR also showed improvements with spectral integration, with RF achieving an R^2 of 0.79 and RMSE of 0.34 N%, and SVR reaching an R^2 of 0.83 and RMSE of 0.30 N%.

At the leaf scale, PINN models using RTM variables alone indicated moderate efficacy (R^2 = 0.44, RMSE = 0.59 N%), which improved with the addition of spectral bands (R^2= 0.67, RMSE = 0.45 N%). The performance of PLSR, GPR, RF, and SVR was generally lower in this scale compared to PINN, with PLSR showing the weakest performance with RTM variables alone (R^2 = 0.40, RMSE = 0.74 N%). The canopy scale analysis revealed moderate correlations for PINN models both with RTM alone (R^2 = 0.57, RMSE = 0.51 N%) and when combined with spectral bands (R^2 = 0.65, RMSE = 0.46 N%). Similar to the leaf scale, GPR, RF, and SVR showed moderate performances, with GPR reaching an R^2 of 0.63 and RMSE of 0.47 N% when combining RTM variables with spectral bands, highlighting the competitive nature of these methods at larger scales. Across all scales combined, the PINN model integrating RTM variables with spectral bands outperformed its single-variable counterpart, achieving an R^2 of 0.71 and RMSE of 0.42 N%. This comparison underscores the enhanced predictive capability of PINN with spectral integration. This pattern was consistent but varied in magnitude among the other methods (PLSR, GPR, RF, and SVR), reflecting the distinct advantages and limitations of each method in leveraging RTM variables and spectral bands for regression analysis.

3.4. Performance of the PINN in the presence of noise

This assessed PINN performance and the equation's behavior under different noise conditions, offering comprehensive applicability insights. Leveraging Eq. (1) as a physical loss function improved PINN accuracy (R^2 = 0.66, RMSE = 0.46 N%), compared to lower performance without it (R^2 = 0.40, RMSE = 0.61 N%) (Table 6 and Fig. 7). Utilization of the physical loss function in the PINN reduced uncertainty, minimized bias, and enhanced the depiction of input spectral data relationships with N%. Testing across scales – dry/ground, leaf, and canopy – revealed consistent improvements with the physical loss function. Canopy scale model performance notably increased (R^2 = 0.62, RMSE = 0.48 N% with physical loss; R^2 = 0.41, RMSE = 0.59 N% without). Similar enhancements were observed at leaf (R^2 = 0.50, RMSE = 0.56 N% with; R^2 = 0.30, RMSE = 0.65 N% without) and dry/ground scales (R^2 = 0.85, RMSE = 0.29 N% with; R^2 = 0.48, RMSE = 0.54 N% without). To enhance the validity of our results amidst the presence of noise and uncertainty in the data, we undertook a comparative analysis. This analysis juxtaposed the performance of PINN, characterized by a physical loss function, with that of traditional machine learning algorithms, namely PLSR, RF, SVR, and GPR. PINN, distinguished by its explicit incorporation of underlying physical principles, exhibited superior R^2 values, underscoring a more robust model fit in comparison to its conventional counterparts.PLSR and SVR both have an R^2 of 0.57 and an RMSE of 0.51, implying moderate performance. Lastly, RF and GPR have slightly lower R^2 values of 0.52 and 0.53, respectively, with GPR also having an RMSE of 0.53, indicating a compromise between fit quality and predictive error. This finding underscores the potential of integrating physical knowledge to markedly enhance predictive accuracy (Table 6 and Fig. 9).



(a) 467 nm vs 1760 nm
(b) 569 nm vs 1760 nm
(c) 734 nm vs 755 nm
(d) 755 nm vs 467 nm
(e) 1526 nm vs 2243 nm
(f) 1559 nm vs 1526 nm
(g) 1760 nm vs 569 nm
(h) 2077 nm vs 569 nm
(i) 2177 nm vs 2276 nm
(j) 2243 nm vs 2375 nm
(k) 2276 nm vs 569 nm
(l) 2375 nm vs 2243 nm
(m) 2473 nm vs 2253 nm

Fig. 5. Integrated reflectance interaction plots across vegetation scales for N% estimation. The plots depict mixed data from dry/ground, leaf, and canopy scales, plotting the reflectance of one spectral band on the x-axis against another band on the y-axis. The color intensity represents the SHAP values for the band on the y-axis, offering insights into its predictive relevance. To interpret the figure, observe how the data points' position and color variation signify the inter-band reflectance relationships and their combined predictive influence on N% across the mixed scales. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

4.1. Inversion of RTM variables

The reliability of MCMC methods for inverting RTM variables from spectral data can be compromised by insufficient preliminary observations, leading to gaps, and by uninformative priors, which are crucial for the Bayesian framework and can result in unstable posterior distributions. These issues are evident in our graphical uncertainty analysis across various scales – dry/ground, leaf, and canopy – presented in Fig. 2. At the dry/ground scale (subplots a and d), uncertainty decreases with increasing wavelength, hinting at more consistent model predictions in the infrared spectrum due to less absorption by soil and ground components. Conversely, at the leaf scale (subplots b and e),

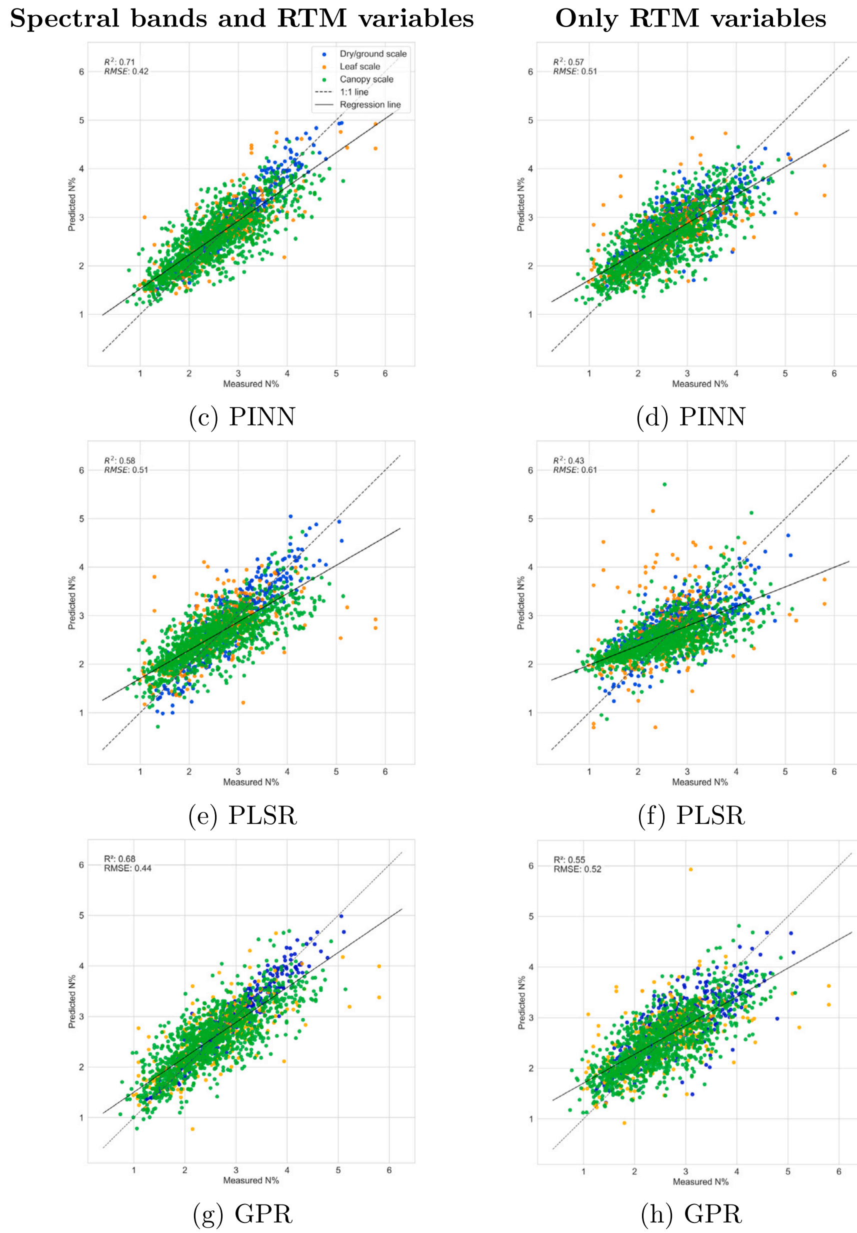


Fig. 6. The comparative analysis of PINN, PLSR, GPR, RF, and SVR in estimating N% based on spectral bands related to chlorophyll and protein content demonstrates the superior performance of PINN. These results illustrate that PINN outperforms traditional models in accurately estimating N%.

uncertainty peaks at key photosynthetic absorption wavelengths, reflecting complex light interactions with leaf biochemistry. The canopy scale (subplots c and f) shows even greater uncertainty due to multiple light scattering and absorption, influenced by factors such as leaf orientation, canopy structure, and background reflectance, highlighting the intricacies of canopy modeling. These complexities illustrate the need for tailored strategies in remote sensing to reduce uncertainty at each scale. Additionally, MCMC's inherent uncertainties – attributable to its stochastic nature and potential Monte Carlo errors, especially in complex or limited models – can affect its utility for nitrogen management decisions. Alternative approaches like variational data assimilation and particle filters offer some resolution but face their challenges, such as handling non-Gaussian errors and the 'curse of dimensionality', which requires more computational resources with increasing model dimensions, as discussed by Del Moral et al. (2006).

Our analysis further explores challenges in spectral data inversion, such as noise and complex biophysical structures, which increase uncertainty. We leverage prior knowledge to minimize these and focus on spectral bands that yield accurate RTM estimates (Dehghan-Shoar et al., 2023a). We find the highest accuracy in dry/ground scale reflectance data, attributing it to reduced complexity in vegetation's biophysical and interleaf characteristics. The dry/ground scale model outperforms canopy scale models, highlighting the significant impact of canopy structure on N% estimation and the effectiveness of controlled environments in mitigating canopy complexity's effect on N% estimation (Liu et al., 2023). Imposing constraints on biophysical and interleaf variables enhances inversion accuracy and reduces uncertainty (Jacquemoud and Baret, 1990). Including additional spectral bands has decreased uncertainty and improved performance, attributed to the unique absorption features across the spectral curve. However, an

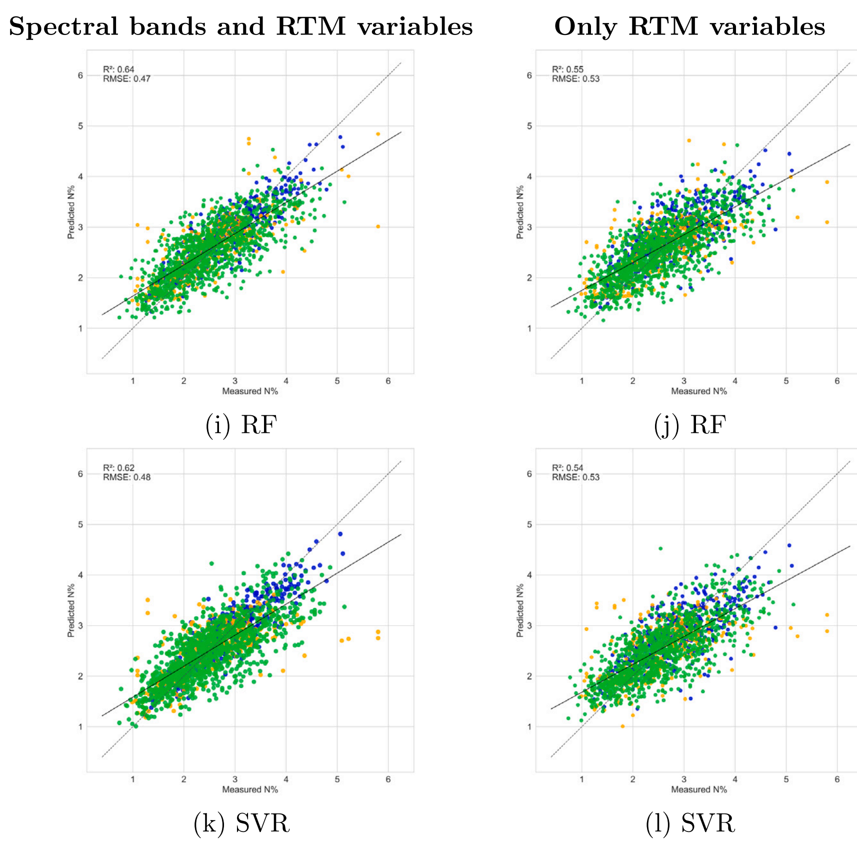


Fig. 6. (*continued*).

Table 5
Comparison of R^2 and Root Mean Square Error (RMSE) values across various methods: Physics Informed Neural Network (PINN), Partial Least Squares Regression (PLSR), Gaussian Process Regression (GPR), Random Forest Regression (RF), and Support Vector Regression (SVR).

Category	Method	R^2	RMSE (N%)
	PINN (RTM variables)	0.61	0.47
Dry/ground scale	PINN (RTM variables and spectral bands)	0.90	0.24
	PLSR (RTM variables)	0.55	0.50
	PLSR (RTM variables and spectral bands)	0.78	0.35
	GPR (RTM variables)	0.61	0.46
	GPR (RTM variables and spectral bands)	0.89	0.23
	RF (RTM variables)	0.58	0.48
	RF (RTM variables and spectral bands)	0.79	0.34
	SVR (RTM variables)	0.60	0.47
	SVR (RTM variables and spectral bands)	0.83	0.30
	PINN (RTM variables)	0.44	0.59
Leaf scale	PINN (RTM variables and spectral bands)	0.67	0.45
	PLSR (RTM variables)	0.40	0.74
	PLSR (RTM variables and spectral bands)	0.50	0.60
	GPR (RTM variables)	0.4	0.60
	GPR (RTM variables and spectral bands)	0.51	0.55
	RF (RTM variables)	0.43	0.59
	RF (RTM variables and spectral bands)	0.57	0.50
	SVR (RTM variables)	0.37	0.64
	SVR (RTM variables and spectral bands)	0.44	0.57
	PINN (RTM variables)	0.57	0.51
Canopy scale	PINN (RTM variables and spectral bands)	0.65	0.46
	PLSR (RTM variables)	0.50	0.58
	PLSR (RTM variables and spectral bands)	0.55	0.52
	GPR (RTM variables)	0.57	0.50
	GPR (RTM variables and spectral bands)	0.63	0.47
	RF (RTM variables)	0.54	0.52
	RF (RTM variables and spectral bands)	0.59	0.49
	SVR (RTM variables)	0.55	0.51
	SVR (RTM variables and spectral bands)	0.58	0.50

(*continued on next page*)

Table 5 (*continued*).

	PINN (RTM variables)	0.57	0.51
Combined	PINN (RTM variables and spectral bands)	0.71	0.42
	PLSR (RTM variables)	0.43	0.61
	PLSR (RTM variables and spectral bands)	0.58	0.51
	GPR (RTM variables)	0.55	0.52
	GPR (RTM variables and spectral bands)	0.68	0.44
	RF (RTM variables)	0.55	0.53
	RF (RTM variables and spectral bands)	0.64	0.47
	SVR (RTM variables)	0.54	0.53
	SVR (RTM variables and spectral bands)	0.62	0.48

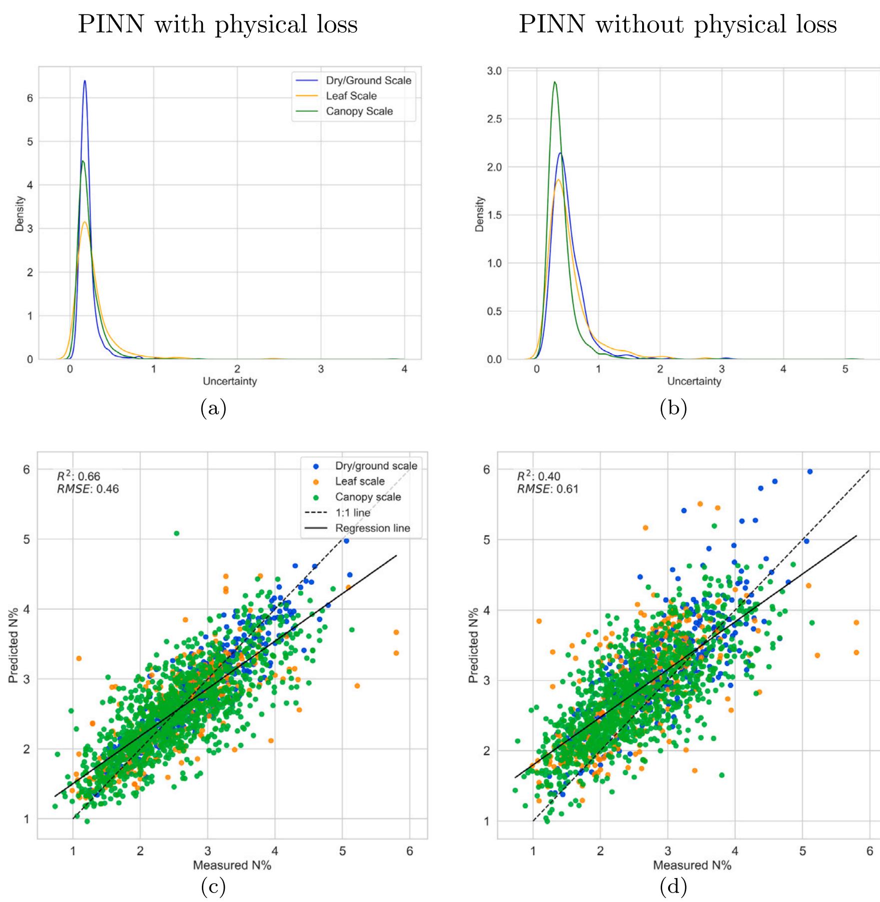


Fig. 7. The efficacy of a physical loss function in improving the performance of the PINN model in the presence of noisy data is demonstrated in this study. Plots c and d show a comparative analysis of the model's accuracy with and without the inclusion of heteroscedastic noise in the dataset, indicating that incorporating physical knowledge significantly improves accuracy and reduces bias. Plots a and b illustrate the uncertainty of the estimations.

excessive number of spectral bands may lead to overfitting, suggesting future research should examine the impact of spectral band selection on model performance.

4.2. Selection of important spectral bands for N% estimation

Our study employed a GA for feature selection to identify key spectral bands crucial for N% estimation (Leardi et al., 1992). The bands predominantly identified fall within the 2000–2500 nm, 1400–2000 nm, and 410–780 nm ranges, consistent with previous research findings (Féret et al., 2021; Pullanagari et al., 2012; Dehghan-Shoar et al., 2023b,a; Curran, 1989). However, we have noted minor discrepancies between the selected bands and the absorption peaks for fundamental nitrogen-containing compounds as delineated in the PROSPECT-PRO model (Fig. 4b). Specifically, while the band at 2276 nm directly aligns with the nitrogen compound absorption peaks, reinforcing its selection (Féret et al., 2021), discrepancies in other bands, such as the 755 nm band's close association with the O2 A absorption band, prompt further reflection (Mohammed et al., 2019). These discrepancies might stem from several factors. For instance, certain selected bands, though not directly overlaying the nitrogen absorption peaks, may exhibit the highest correlation with N% due to their indirect influence on photosynthetic efficiency and plant health, as evidenced by their sensitivity to chlorophyll content and photosynthetic activity. This is particularly true for the 734 nm band within the red-edge region, which, despite its theoretical detachment from direct N% estimation, provides significant insights due to nitrogen's impact on photosynthesis. It is also conceivable that the optimal selection of bands results from a combination of factors, including their direct and indirect relationships with nitrogen content, the complex interplay of spectral signatures, and plant physiological responses to varying nitrogen levels. These elements, coupled with the GA's capability to discern patterns within multi-dimensional datasets, might explain the preference for bands that do not strictly align with known nitrogen absorption peaks. Given these

Table 6
The comparison of the coefficient of determination (R^2) and Root Mean Square Error (RMSE) values for the Physics Informed Neural Network (PINN) utilizing RTM variables alongside critical spectral bands, both with and without the inclusion of physical loss, is examined alongside the performance of Gaussian Process Regression (GPR), Random Forest (RF), and Partial Least Squares Regression (PLSR) under noisy conditions at various scales (dry/ground, leaf, canopy, and a combined).

Category	Method	R^2	RMSE (N%)
Dry/ground scale	PINN (with physical loss)	0.85	0.29
	PINN (without physical loss)	0.48	0.54
	PLSR	0.71	0.41
	RF	0.43	0.56
	SVR	0.61	0.47
	GPR	0.69	0.42
Leaf scale	PINN (with physical loss)	0.50	0.56
	PINN (without physical loss)	0.30	0.65
	PLSR	0.31	0.66
	RF	0.38	0.62
	SVR	0.40	0.61
	GPR	0.41	0.65
Canopy scale	PINN (with physical loss)	0.62	0.48
	PINN (without physical loss)	0.41	0.59
	PLSR	0.55	0.52
	RF	0.47	0.57
	SVR	0.57	0.51
	GPR	0.58	0.50
Combined	PINN (with physical loss)	0.66	0.46
	PINN (without physical loss)	0.40	0.61
	PLSR	0.57	0.51
	RF	0.52	0.54
	SVR	0.57	0.51
	GPR	0.53	0.53

considerations, our findings underscore the need for further research to elucidate the underlying mechanisms governing the correlation between these spectral bands and N% estimation. This exploration could reveal novel insights into the spectral characteristics of vegetation and their intricate relationship with nitrogen availability, potentially leading to more refined approaches in remote sensing applications for nutrient management.

Our investigation into spectral characteristics, particularly within the critical ranges of 2000–2500 nm, 1400–2000 nm, and 410–780 nm, has yielded valuable insights for estimating N%. However, this study also highlights the complex challenges associated with varying scales of spectral data collection. The complexity of harmonizing spectral data is underscored by the differences in measurement approaches, such as the bidirectional reflectance from dry/ground samples versus directional-hemispherical reflectance in fresh leaf samples, as seen in the LOPEX93 dataset. Bridging the gap between the hemispherical-directional reflectance factor (HDRF) and hemispherical reflectance (HRF) facilitates accurate data integration and analysis. We developed Eq. (1), focusing on selecting spectral bands after thoroughly examining HDRF effects across the spectrum. This process ensures these bands exhibit minimal HDRF influence, enhancing the equation's reliability amidst diverse measurement techniques.

4.3. Performance of the physical loss function

Our study implements the GP-GOMEA algorithm to develop a physical loss function (Eq. (1)) for calibrating the PINN, focusing on spectral bands at 569 nm, 734 nm, 755 nm, 2077 nm, 2177 nm, 1559 nm, and 467 nm due to their direct correlation with N% estimation (Fig. 4a). This approach necessitates meticulous spectral band selection to balance the risks of overfitting and underfitting, given the symbolic expression's complexity generated by the algorithm, which might pose interpretation challenges.

We highlight the complexity of non-linear interactions between these spectral bands in achieving accurate N% predictions. This perspective challenges the traditional reliance on linear models, such as PLSR,GPR RF and SVR, underscoring the need for models that capture complex dynamics, as linear models often oversimplify the nuanced realities of spectral data analysis (Fig. 5). Notably, bands at 1526 nm and 2243 nm show lower reflectance correlating with higher N%, emphasizing the importance of a diverse spectral range to reduce prediction noise and uncertainty.

In the context of our study, these high SHAP values underscore the particularly influential nature of these pairs in estimating N%. This observation illuminates the interplay between different spectral bands and their cumulative impact on the model's predictive capability, adding another layer of complexity and understanding to our predictive framework.

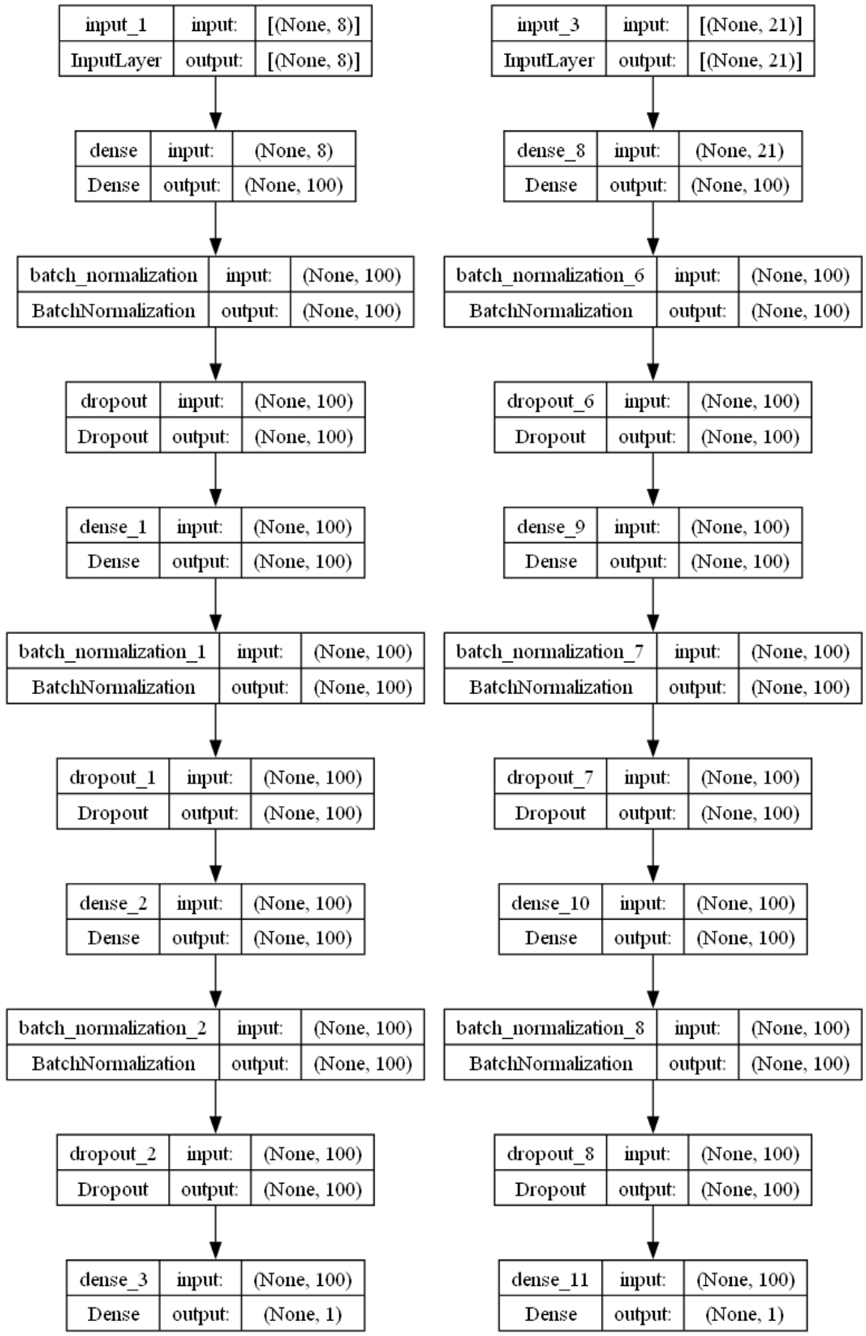
The complexity of interpreting the algorithm's symbolic expression indicates the need for advancements in transparent and effective feature selection and regression techniques. Scientific computing involves using advanced computing capabilities to solve complex scientific problems. It integrates mathematics, physics, and computer science to simulate and analyze natural and engineered systems. This field relies on numerical methods, algorithms, and computational tools to perform experiments in a virtual environment, allowing scientists to understand phenomena that are difficult, dangerous, or impossible to study physically. Scientific computing enables precise predictions, optimizations, and the exploration of scenarios across various disciplines, from climate modeling to bioinformatics, enhancing our understanding of the world.

Scientific Machine Learning (SciML) represents a cutting-edge synthesis of machine learning and traditional scientific disciplines, aimed at improving predictive models by incorporating physical laws into algorithms such as Physics-Informed Neural Networks (PINNs). This inclusion of physical loss functions, as noted by Camps-Valls et al. (2023), enhances the accuracy and interpretability of the models. It reflects a paradigm shift towards models that are not just data-driven but also rooted in the principles governing natural phenomena. The differential performance of PINNs with and without physical constraints emphasizes the significant benefits of domain-specific knowledge in boosting prediction accuracy for nitrogen estimation—a critical factor in resource management for sectors like agriculture and forestry. Furthermore, the principles underpinning SciML have broader implications for spectral data analysis, offering potential advancements in the detection and quantification of elements such as arsenic and antimony, which are correlated with nitrogen levels, as reported by Chakraborty et al. (2022), Rodriguez-Gomez et al. (2023).

4.4. Role of PINN in an improved N%

We introduce a novel PINN that enables the integration of domain knowledge in deep neural networks for monitoring N% in vegetation. This approach, grounded in scientific principles (e.g., protein absorption), can significantly improve model performance and mark a new era in vegetation spectroscopy research (Ballschmiter and Katz, 1969). Our study demonstrates that PINNs can outperform traditional models like PLSR, GPR, RF, and SVR by incorporating system physics knowledge, a first in this field (Fig. 6 and Table 5), echoing the success of the PINN with multiscale spectral data (Wan et al., 2022).

Integrating a physical loss function into our model leads to notable performance improvements, even with noisy data, by reducing prediction bias and uncertainty (Fig. 7 and Table 6). These results align with Buster et al. (2022)'s findings that combining physical principles with deep learning offers effective regularization, enhancing model generalizability. However, the applicability of PINNs is contingent upon the availability of a relevant physical loss function. Hence, their estimation relies on indirect measures through other biochemical and biophysical proxies, like the association between phosphorus content and SWIR spectra, mediated by protein absorption bands (Chakraborty et al., 2022; Rodriguez-Gomez et al., 2023). This



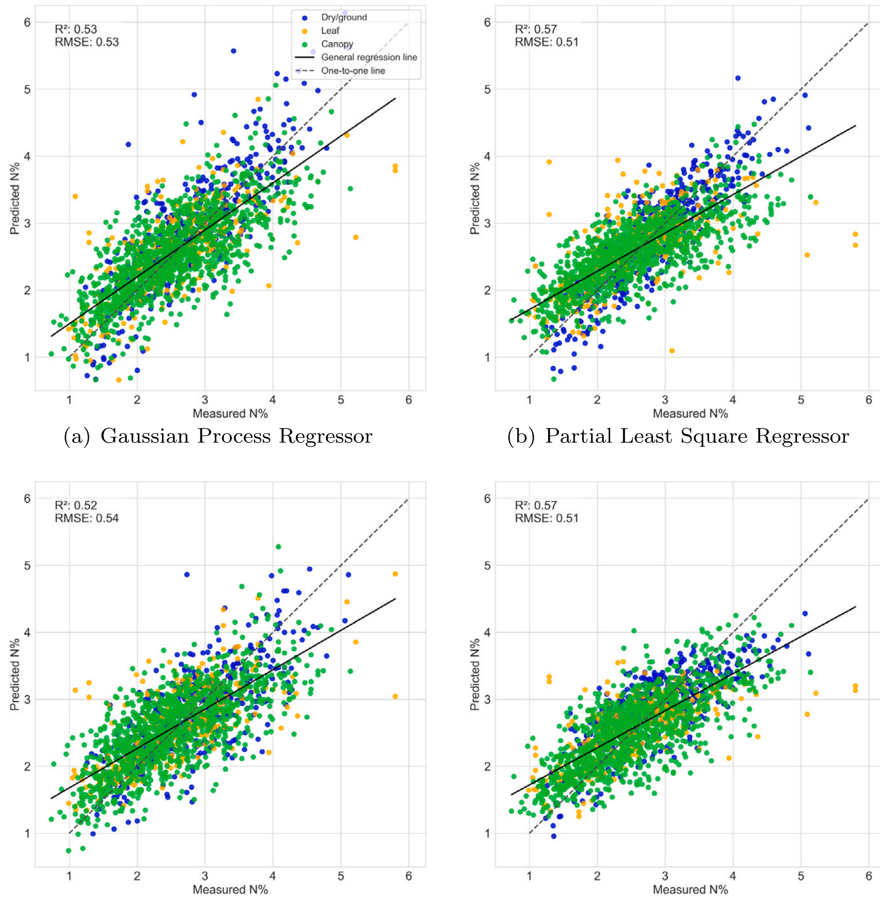
(a) PINN (RTM variables) (b) PINN (spectral bands and RTM variables)

Fig. 8. Diagram depicting the architecture of a dual-pathway neural network, with each path incorporating batch normalization and dropout layers that converge into dense layers. Image (a) represents the architecture tailored for a Physics-Informed Neural Network (PINN) utilizing solely Radiative Transfer Model (RTM) variables. Image (b) corresponds to the architecture adapted for a PINN that integrates both RTM variables and key spectral bands, indicating a more complex input feature set.

highlights the potential limitations of PINNs in contexts where direct physical relationships are not well-defined.

5. Conclusion

This study leveraged a comprehensive multiscale spectral dataset to develop a novel PINN model for predicting N%, integrating spectral and radiative transfer model (RTM) attributes with a custom physical loss function. This model identified critical spectral bands (2276 nm, 755 nm, 1526 nm, 2243 nm, and 734 nm) strongly correlated with N%, essential for accurate estimation. It exhibited outstanding performance across different scales, with a notable performance decrease from dry/ground to canopy scales due to canopy complexity and variable water content. Specifically, at the dry/ground scale, the model achieved an R^2 of 0.90 and an RMSE of 0.24%N, while at canopy



(a) Gaussian Process Regressor   (b) Partial Least Square Regressor

(c) Random Forest Regressor   (d) Support Vector Regressor

Fig. 9. Comparative Performance of Regression Models. Subfigures (a), (b), (c), and (d) illustrate the predictive accuracy of four different regression models. (a) Gaussian Process Regression, highlighting its capability to model complex, non-linear relationships with an inherent measure of uncertainty. (b) Partial Least Squares Regression, emphasizing its effectiveness in dealing with highly multivariate and collinear data. (c) Random Forest Regressor, showcases its ensemble learning approach for capturing non-linear interactions and providing robustness against overfitting. (d) Support Vector Regression, demonstrating its application of kernel functions to capture non-linear patterns in data with high dimensionality. Each subfigure includes a scatter plot of the measured vs. predicted values, along with a one-to-one reference line, regression fit, and corresponding R^2 and RMSE values to assess model performance.

and leaf scales, it recorded R^2 values of 0.65 and 0.67 and RMSEs of 0.46%N and 0.45%N, respectively, resulting in an overall R^2 of 0.71 and an RMSE of 0.42%N, outperforming models using fewer spectral bands.

Future research aims to enhance model accuracy by incorporating aerial and spaceborne remote sensing data, examining the impacts of topography and atmospheric conditions on N% predictions. This approach could improve robustness and real-world applicability. Our model's success over the PLSR model, by capturing non-linear relationships between spectral bands and RTM variables, offers insights for refining N% estimation. Accurate N% estimation has potential implications for advancing sustainable land-use and vegetation management practices.

CRediT authorship contribution statement

Mohammad Hossain Dehghan-Shoar: Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Gabor Kereszturi:** Writing – review & editing, Writing – original draft, Methodology. **Reddy R. Pullanagari:** Writing – review & editing, Writing – original draft, Supervision. **Alvaro A. Orsi:** Writing – review & editing. **Ian J. Yule:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **James Hanly:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

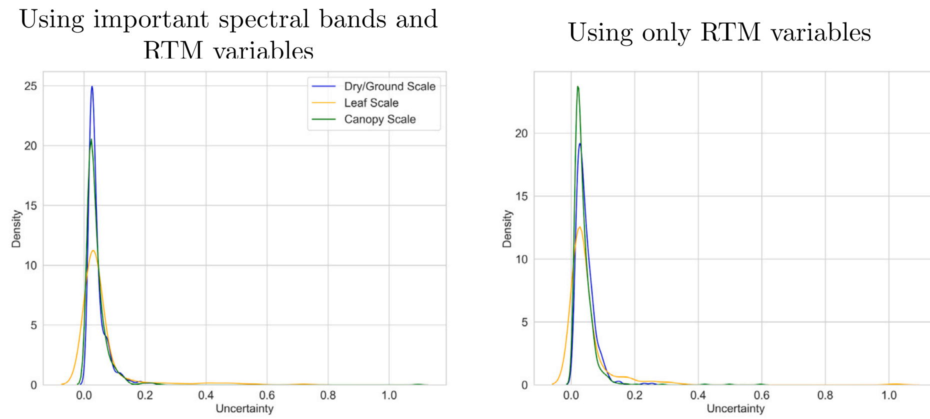
The data that has been used is confidential.

Acknowledgment

We sincerely thank the Pastoral 21 research program, The C. Alma Baker Trust, Ministry for Primary Industries, Ravensdown, New Zealand, and all the individuals who contributed to this study. We would also like to acknowledge the staff and students of Massey University and the staff of AgResearch New Zealand for their efforts in collecting the field samples.

Appendix

See Figs. 8–10.



(a) Uncertainty of PINN (b) Uncertainty of PINN

Fig. 10. Graphs (a) and (b) detail PINN model uncertainties highlighting the impact of including or excluding significant spectral bands in PINN.

References

Ballschmiter, K., Katz, J.J., 1969. Infrared study of chlorophyll-chlorophyll and chlorophyll-water interactions. J. Am. Chem. Soc. 5 (1), 57–67. http://dx.doi.org/10.1021/ja01038a044.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., Cox, D.D., 2015. Hyperopt: A Python library for model selection and hyperparameter optimization. Comput. Sci. Discov. 8 (1), 014008. http://dx.doi.org/10.1088/1749-4699/8/1/014008.

Buster, G., Bannister, M., Habte, A., Hettinger, D., Maclaurin, G., Rossol, M., Sengupta, M., Xie, Y., 2022. Physics-guided machine learning for improved accuracy of the National Solar Radiation Database. Sol. Energy 232, 483–492. http://dx.doi.org/10.1016/j.solener.2022.01.004.

Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., Runge, J., 2023. Discovering causal relations and equations from data. Phys. Rep. 1044, 1–68. http://dx.doi.org/10.1016/j.physrep.2023.10.005, URL: https://www.sciencedirect.com/science/article/pii/S0370157323003411, Discovering causal relations and equations from data.

Chakraborty, R., Kereszturi, G., Pullanagari, R., Durance, P., Ashraf, S., Anderson, C., 2022. Mineral prospecting from biogeochemical and geological information using hyperspectral remote sensing - Feasibility and challenges. J. Geochem. Explor. 232, 106900. http://dx.doi.org/10.1016/j.gexplo.2021.106900.

Chen, Q., Zheng, B., Chen, T., Chapman, S., 2021. Integration of APSIM and PROSAIL models to develop more precise radiometric estimation of crop traits using deep learning. bioRxiv.

Chen, Q., Zheng, B., Chen, T., Chapman, S.C., 2022. Integrating a crop growth model and radiative transfer model to improve estimation of crop traits based on deep learning. J. Exp. Bot. 73 (19), 6558–6574.

Curran, P.J., 1989. Remote sensing of foliar chemistry. Remote Sens. Environ. 30 (3), 271–278. http://dx.doi.org/10.1016/0034-4257(89)90069-2.

Dehghan-Shoar, M.H., 2018. Use of New Generation Geospatial Data and Technology for Low Cost Drought Monitoring and SDG Reporting Solution: A Thesis Presented in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Science at Massey University, Manawatū, New Zealand (Ph.D. thesis). Massey University.

Dehghan-Shoar, M.H., Orsi, A.A., Pullanagari, R.R., Yule, I.J., 2023a. A hybrid model to predict nitrogen concentration in heterogeneous grassland using field spectroscopy. Remote Sens. Environ. 285, 113385. http://dx.doi.org/10.1016/j.rse.2022.113385.

Dehghan-Shoar, M.H., Pullanagari, R.R., Kereszturi, G., Orsi, A.A., Yule, I.J., Hanly, J., 2023b. A unified physically based method for monitoring grassland nitrogen concentration with Landsat 7, Landsat 8, and sentinel-2 satellite data. Remote Sens. 15 (10), 2491. http://dx.doi.org/10.3390/rs15102491.

Dehghan-Shoar, M.H., Pullanagari, R.R., Orsi, A.A., Yule, I.J., 2023c. Simulating spaceborne imaging to retrieve grassland nitrogen concentration. Remote Sens. Appl.: Soc. Environ. 29, 100912. http://dx.doi.org/10.1016/j.rsase.2022.100912.

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. J. R. Stat. Soc. B 68 (3), 411–436. http://dx.doi.org/10.1111/j.1467-9868.2006.00553.x.

Dhakal, R., Maimaitijiang, M., Chang, J., Caffe, M., 2023. Utilizing spectral, structural and textural features for estimating oat above-ground biomass using UAV-based multispectral data and machine learning. Sensors 23 (24), 9708.

FAO, 2020. Global livestock production. http://dx.doi.org/10.1787/affaf7ec-en, OECD-FAO Agricultural Outlook, Food and Agriculture Organization of the United Nations.

Féret, J.-B., Berger, K., de Boissieu, F., Malenovský, Z., 2021. PROSPECT-PRO for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents. Remote Sens. Environ. 252, 112173. http://dx.doi.org/10.1016/j.rse.2020.112173.

Foreman-Mackey, D., Farr, W., Sinha, M., Archibald, A., Hogg, D., Sanders, J., Zuntz, J., Williams, P., Nelson, A., de Val-Borro, M., Erhardt, T., Pashchenko, I., Pla, O., 2019. emcee v3: A Python ensemble sampling toolkit for affine-invariant MCMC. J. Open Source Softw. http://dx.doi.org/10.21105/joss.01864.

Ge, Y., Atefi, A., Zhang, H., Miao, C., Ramamurthy, R.K., Sigmon, B., Yang, J., Schnable, J.C., 2019. High-throughput analysis of leaf physiological and chemical traits with VIS–NIR–SWIR spectroscopy: A case study with a maize diversity panel. Plant Methods 15 (1), 1–12. http://dx.doi.org/10.1186/s13007-019-0450-8.

Goetz, H., Whitehead, D.C., 1971. The role of nitrogen in grassland productivity. Ecology (48), http://dx.doi.org/10.2307/1934171.

grantbuster, Rossol, M., Bannister, M., Hettinger, D., 2021. NREL/phygnn: release for DOI. http://dx.doi.org/10.5281/zenodo.4498541, Zenodo.

Grzybowski, M., Wijewardane, N.K., Atefi, A., Ge, Y., Schnable, J.C., 2021. Hyperspectral reflectance-based phenotyping for quantitative genetics in crops: progress and challenges. Plant Commun. 2 (4), 100209.

Hodáňová, D., 1985. Leaf optical properties. In: Photosynthesis During Leaf Development. http://dx.doi.org/10.1007/978-94-009-5530-1_5.

Ishaq, R.A.F., Zhou, G., Tian, C., Tan, Y., Jing, G., Jiang, H., 2023. A systematic review of radiative transfer models for crop yield prediction and crop traits retrieval. Remote Sens. 16 (1), 121.

Jacquemoud, S., Baret, F., 1990. PROSPECT: A model of leaf optical properties spectra. Remote Sens. Environ. 34 (2), 75–91. http://dx.doi.org/10.1016/0034-4257(90)90100-z.

Jarvis, R.M., Goodacre, R., 2005. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. Bioinformatics 21 (7), 860–868. http://dx.doi.org/10.1093/bioinformatics/bti102.

Junhwan, C., Seokmin, O., Joongmoo, B., 2022. Uncertainty estimation in AVO inversion using Bayesian dropout based deep learning. In: International Conference on Machine Learning. J. Pet. Sci. Technol. 1050–1059. http://dx.doi.org/10.1016/j.petrol.2021.109288.

Kecman, V., 2005. Support vector machines–An introduction. In: Support Vector Machines: Theory and Applications. Springer, pp. 1–47.

Kokaly, R.F., 2001. Investigating a physical basis for spectroscopic estimates of leaf nitrogen concentration. Remote Sens. Environ. 75 (2), 153–161. http://dx.doi.org/10.1016/s0034-4257(00)00163-2.

Lanitis, A., Taylor, C., Cootes, T., 2017. A unified approach to coding and interpreting face images. In: Proc. IEEE Int. Conf. Comput. Vis., vol. 30, http://dx.doi.org/10.1109/iccv.1995.466919.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemom. 6 (5), 267–281. http://dx.doi.org/10.1002/cem.1180060506.

Liang, L., Di, L., Huang, T., Wang, J., Lin, L., Wang, L., Yang, M., 2018. Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. Remote Sens. 10 (12), 1940.

Liu, S., Yan, Z., Wang, Z., Serbin, S., Visser, M., Zeng, Y., Ryu, Y., Su, Y., Guo, Z., Song, G., Wu, Q., Zhang, H., Cheng, K., Dong, J., Hau, B.C.H., Zhao, P., Yang, X., Liu, L., Rogers, A., Wu, J., 2023. Mapping foliar photosynthetic capacity in subtropical and tropical forests with UAS-based imaging spectroscopy: Scaling from leaf to canopy. Remote Sens. Environ. 293, 113612. http://dx.doi.org/10.1016/j.rse.2023.113612.

Mohammed, G.H., Colombo, R., Middleton, E.M., Rascher, U., van der Tol, C., Nedbal, L., Goulas, Y., Pérez-Priego, O., Damm, A., Meroni, M., Joiner, J., Cogliati, S., Verhoef, W., Malenovský, Z., Gastellu-Etchegorry, J.-P., Miller, J.R., Guanter, L., Moreno, J., Moya, I., Berry, J.A., Frankenberg, C., Zarco-Tejada, P.J., 2019. Remote sensing of Solar-Induced chlorophyll Fluorescence (SIF) in vegetation: 50 years of progress. Remote Sens. Environ. 231, 111177. http://dx.doi.org/10.1016/j.rse.2019.04.030.

Mutanga, O., 2004. Hyperspectral Remote Sensing of Tropical Grass Quality and Quantity (Ph.D. thesis) ITC Dissertation. Wageningen University, Netherlands, ITC Dissertation; 111.

Petermann, J.S., Buzhdygan, O.Y., 2021. Grassland biodiversity. Curr. Biol. 31 (19), R1195–R1201. http://dx.doi.org/10.1016/j.cub.2021.06.060.

Pullanagari, R., Dehghan-Shoar, M., Yule, I.J., Bhatia, N., 2021. Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network. Remote Sens. Environ. 257, 112353. http://dx.doi.org/10.1016/j.rse.2021.112353.

Pullanagari, R.R., Kereszturi, G., Yule, I., 2018. Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. Remote Sens. 10 (7), 1117.

Pullanagari, R.R., Yule, I.J., Tuohy, M.P., Hedley, M.J., Dynes, R.A., King, W.M., 2012. In-field hyperspectral proximal sensing for estimating quality parameters of mixed pasture. Precis. Agric. 13 (3), 351–369. http://dx.doi.org/10.1007/s11119-011-9251-4.

Queally, N., Ye, Z., Zheng, T., Chlus, A., Schneider, F., Pavlick, R.P., Townsend, P.A., 2022. FlexBRDF: A Flexible BRDF correction for grouped processing of airborne imaging spectroscopy flightlines. J. Geophys. Res.: Biogeosci. 127 (1), http://dx.doi.org/10.1029/2021jg006622, e2021JG006622.

Raissi, M., Perdikaris, P., Karniadakis, G., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707. http://dx.doi.org/10.1016/j.jcp.2018.10.045.

Ramankutty, N., Evan, A.T., Monfreda, C., Foley, J.A., 2008. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. Global Biogeochem. Cycles 22 (1), http://dx.doi.org/10.1029/2007gb002952.

Rodriguez-Gomez, C., Kereszturi, G., Jeyakumar, P., Pullanagari, R., Reeves, R., Rae, A., Procter, J.N., 2023. Remote exploration and monitoring of geothermal sources: A novel method for foliar element mapping using hyperspectral (VNIR-SWIR) remote sensing. Geothermics 111, 102716. http://dx.doi.org/10.1016/j.geothermics.2023.102716.

Sanches, I.D., Tuohy, M.P., Hedley, M.J., Bretherton, M.R., 2009. Large, durable and low-cost reflectance standard for field remote sensing applications. Int. J. Remote Sens. 30 (9), 2309–2319. http://dx.doi.org/10.1080/01431160802549377.

Schell, U., 2022. Keras/Tensorflow. Maschinelles Lernen Mit R, http://dx.doi.org/10.3139/9783446472440.010.

Taubert, F., Frank, K., Huth, A., 2012. A review of grassland models in the biofuel context. Ecol. Modell. 245, 84–93. http://dx.doi.org/10.1016/j.ecolmodel.2012.04.007.

Verhoef, W., 1984. Light scattering by leaf layers with application to canopy reflectance modeling: The SAIL model. Remote Sens. Environ. 16 (2), 125–141. http://dx.doi.org/10.1016/0034-4257(84)90057-9.

Virgolin, M., Alderliesten, T., Witteveen, C., Bosman, P.A.N., 2021. Improving model-based genetic programming for symbolic regression of small expressions. Evol. Comput. 29 (2), 211–237. http://dx.doi.org/10.1162/evco_a_00278.

Virgolin, M., De Lorenzo, A., Medvet, E., Randone, F., 2020. Learning a formula of interpretability to learn interpretable formulas. In: Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, the Netherlands, September 5-9, 2020, Proceedings, Part II 16. Springer, pp. 79–93.

Wan, L., Zhou, W., He, Y., Wanger, T.C., Cen, H., 2022. Combining transfer learning and hyperspectral reflectance analysis to assess leaf nitrogen concentration across different plant species datasets. Remote Sens. Environ. 269, 112826. http://dx.doi.org/10.1016/j.rse.2021.112826.