Remote Sensing of Environment 328 (2025) 114820

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

On the generalization ability of probabilistic neural networks for hyperspectral remote sensing of absorption properties across optically complex waters

Mortimer Werther [a,*], Olivier Burggraaff [b], Daniela Gurlin [c], Arun M. Saranathan [d,e], Sundarabalan V. Balasubramanian [f], Claudia Giardino [g,h], Federica Braga [i], Mariano Bresciani [g], Andrea Pellegrino [g,j], Monica Pinardi [g], Stefan G.H. Simis [k], Moritz K. Lehmann [l,m], Kersti Kangro [n,o], Krista Alikas [n], Dariusz Ficek [p], Daniel Odermatt [a,q]

[a] Swiss Federal Institute of Aquatic Science and Technology, Department of Surface Waters - Research and Management, Dübendorf, Switzerland
[b] Institute of Environmental Sciences (CML), Leiden University, Leiden, Netherlands
[c] Wisconsin Department of Natural Resources, Madison, WI, United States of America
[d] Science Systems and Applications, Inc., Lanham, MD, United States
[e] Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, United States
[f] Geo-Sensing and Imaging Consultancy, Trivandrum, India
[g] Institute for Electromagnetic Sensing of the Environment, National Research Council (CNR-IREA), 20133 Milano, Italy
[h] National Bio-diversity Future Center (NBFC), 90133 Palermo, Italy
[i] Institute of Marine Sciences, National Research Council (CNR-ISMAR), 30122 Venice, Italy
[j] University of Sapienza, Department of Engineering, 00185 Rome, Italy
[k] Plymouth Marine Laboratory, Plymouth, United Kingdom
[l] Starboard Maritime Intelligence, Wellington, New Zealand
[m] The University of Waikato, Hamilton, New Zealand
[n] Department of Remote Sensing, Tartu Observatory, University of Tartu, Tõravere, Estonia
[o] Centre for Limnology, Estonian University of Life Sciences, Vehendi, Estonia
[p] Institute of Geography, Pomeranian University in Slupsk, 76-200 Slupsk, Poland
[q] Department of Geography, University of Zurich, 8057, Zurich, Switzerland

ARTICLE INFO

Edited by Menghua Wang

Keywords:
Generalization
Neural networks
Inherent optical properties
Hyperspectral remote sensing
Optically complex waters

ABSTRACT

Machine learning models have steadily improved in estimating inherent optical properties (IOPs) from remote sensing observations. Yet, their generalization ability when applied to new water bodies, beyond those they were trained on, is not well understood. We present a novel approach for assessing model generalization across various scenarios, including interpolation within *in situ* observation datasets, extrapolation beyond the training scope, and application to hyperspectral observations from the PRecursore IperSpettrale della Missione Applicativa (PRISMA) satellite involving atmospheric correction. We evaluate five probabilistic neural networks (PNNs), including novel architectures like recurrent neural networks, for their ability to estimate absorption at 443 and 675 nm from hyperspectral reflectance. The median symmetric accuracy (MdSA) worsens from ≥25% in interpolation scenarios to ≥50% in extrapolation scenarios, and reaches ≥80% when applied to PRISMA satellite imagery. Across all scenarios, models produce uncertainty estimates exceeding 40%, often reflecting systematic underconfidence. PNNs show better calibration during extrapolation, suggesting an intrinsic awareness of retrieval constraints. To address this miscalibration, we introduce an uncertainty recalibration method that only withholds 10% of the training dataset, but improves model calibration in 86% of PRISMA evaluations with minimal accuracy trade-offs. Resulting well-calibrated uncertainty estimates enable reliable uncertainty propagation for downstream applications. IOP retrieval uncertainty is predominantly aleatoric (inherent to the observations). Therefore, increasing the number of measurements from the same distribution or selecting a different neural network architecture trained on the same dataset does not enhance

* Corresponding author.
E-mail address: mortimer.werther@eawag.ch (M. Werther).

https://doi.org/10.1016/j.rse.2025.114820
Received 15 October 2024; Received in revised form 9 May 2025; Accepted 12 May 2025
Available online 2 June 2025

model accuracy. Our findings indicate that we have reached a predictability limit in retrieving IOPs using purely data-driven approaches. We therefore advocate embedding physical principles of IOPs into model architectures, creating physics-informed neural networks capable of surpassing current limitations.

1. Introduction

Reflectance measurements of surface waters in the visible-light spectrum, obtained through satellite and airborne sensors, enable the assessment of water quality (Bukata et al., 1974; Morel, 1980), detection of extreme ecological events (Irani Rahaghi et al., 2024; Köhler et al., 2024; Tang et al., 2021), and monitoring of long-term changes in aquatic environments (Meyer et al., 2024; Schaeffer et al., 2022b). Central to these applications is understanding how photons interact with water and its constituents, characterized by inherent optical properties (IOPs; Morel and Prieur (1977)). The spectral absorption coefficients of phytoplankton ($a_{ph}(\lambda)$), colored dissolved organic matter ($a_{CDOM}(\lambda)$), and non-algal particles ($a_{NAP}(\lambda)$) are key IOPs that provide insights into biogeochemical processes, forming the foundation for numerous remote sensing applications (Astuti et al., 2018; Behrenfeld et al., 2009; Effler et al., 2006; Hommersom et al., 2009; Silsbe et al., 2016).

Spaceborne remote sensing approaches for retrieving IOPs through reflectance inversion were initially developed for open ocean and coastal waters using multispectral sensors. Widely used inversion approaches include the quasi-analytical algorithm (QAA; Lee et al. (2002)), generalized inherent optical property model (GIOP; Werdell et al. (2013)), and three-component semi-analytical algorithm (3SAA; Jorge et al. (2021)). With the advent of hyperspectral missions like NASA's Plankton, Aerosol, Cloud, ocean Ecosystem (PACE), DLR's Environmental Mapping and Analysis Program (EnMAP), and ASI's PRecursore IperSpettrale della Missione Applicativa (PRISMA), there is a need for algorithms that fully exploit the enhanced spectral resolution available (Cael et al., 2023). Such high-dimensional observations call for methods capable of capturing the subtle spectral signatures of IOPs, an ability at which neural networks (NNs) excel.

NNs have become popular in aquatic remote sensing because of their ability as universal function approximators and the growing availability of extensive training datasets (Bricaud et al., 2007; Brockmann et al., 2016; González Vilas et al., 2011; Hieronymi et al., 2017; Keiner and Yan, 1998; Schiller and Doerffer, 1999). Their flexibility enables them to model complex relationships between hyperspectral reflectance and IOPs across optically complex waters. However, successful NN application depends on training datasets that accurately represent the full range of environmental conditions. Consequently, quantifying uncertainties in NN-based estimates is critical to ensure reliable deployment (Gray et al., 2024).

Most operational NNs in this domain provide only deterministic estimates, with only a few approaches incorporating explicit uncertainty quantification. For example, the C2RCC neural network (Brockmann et al., 2016) offers uncertainty estimates for water constituents and IOPs by leveraging forward and inverse models to flag inputs that deviate from its training scope (Doerffer and Schiller, 2007). However, this method has not yet been extended to hyperspectral sensors. More recently, probabilistic neural networks (PNNs) have emerged, producing probability distributions for outputs and thus enabling explicit uncertainty quantification (Werther et al., 2022; Saranathan et al., 2023). Despite their promise, relatively few studies have examined PNN architectures for hyperspectral IOP retrieval (O'Shea et al., 2023; Pahlevan et al., 2022; Saranathan et al., 2024), leaving important gaps in our understanding of their performance under varied optical conditions. A key dimension of this performance involves uncertainty estimation, which plays a major role in assessing model reliability.

Uncertainties in IOP retrieval stem from various sources and can be studied through the lens of aleatoric and epistemic components (Hüllermeier and Waegeman, 2021). *Aleatoric uncertainty* arises from inherent variability or error in observations and persists even with increased sampling of the same type. Examples include error in R_{rs} measurements (both *in situ* and atmospherically corrected), uncertainties in reference *in situ* IOP measurements (Leymarie et al., 2010; Wang et al., 2005), ambiguity due to different IOP combinations producing (near-)identical R_{rs} (Defoin-Platel and Chami, 2007; Zaneveld, 1994), and other observational processing errors (Burggraaff, 2020), as well as true variability in the measurands due to random effects. *Epistemic uncertainty*, conversely, reflects limitations in model knowledge or problem space coverage, often due to gaps within or at the limits of the distribution of training samples. In practice, aleatoric and epistemic uncertainty often coexist (Valdenegro-Toro and Mori, 2022).

Limited datasets, whether from *in situ* measurements (Lehmann et al., 2023) or simulations (Loisel et al., 2023), cannot capture the full bio-optical variability of natural and man-made waters. Consequently, predictive models are frequently required to extrapolate beyond their training IOP distribution, especially when applied to diverse satellite imagery. Ideally, the generalization ability of a model should be indicated by its uncertainty estimates, exhibiting low uncertainty for accurate estimates in familiar scenarios and high uncertainty when extrapolating to novel scenarios. However, good alignment between estimated uncertainties and actual errors – known as *calibration* – is not guaranteed (Guo et al., 2017; Minderer et al., 2021).

PNNs can simultaneously model aleatoric uncertainty and assess epistemic uncertainty through output variability, effectively quantifying and distinguishing different types of uncertainty. However, three critical gaps remain in understanding how these networks generalize.

First, most NN studies have presented results from individual models without rigorous comparisons between different architectures under standardized conditions (Brockmann et al., 2016; Hieronymi et al., 2017; O'Shea et al., 2023; Pahlevan et al., 2020; Werther et al., 2022). Although Saranathan et al. (2024) recently compared two PNNs using identical datasets and hyperparameters for estimating water constituents like chlorophyll-*a* concentration, and found that the models achieved similar retrieval accuracy but substantially differed in their uncertainty estimates, standardized evaluations using PNNs for IOP retrieval are still undocumented. It remains unresolved to what extent variations in PNN performance for IOP estimation are due to architectural choices, training datasets, the inherent complexity of IOPs, or other unidentified factors.

We address this gap through a comprehensive investigation of five distinct PNN architectures using consistent training and parameter configurations. Based on aforementioned literature, we expect minimal differences in IOP retrieval accuracy between architectures, but significant variability in uncertainty estimation.

Second, the generalization ability of PNN-based IOP retrieval methods to unfamiliar real-world scenarios is inadequately studied. Current literature predominantly employs either random splits or geographical leave-one-out (LOO) strategies for validation with large *in situ* datasets (Cao et al., 2020; Pahlevan et al., 2020, 2022; Saranathan et al., 2024; Smith et al., 2021; Werther et al., 2022). Random splits are subject to knowledge leakage (Stock et al., 2023) because observations from the same water body appear in both training and test sets, hindering the assessment of true generalization ability (Gray et al., 2024). The LOO approach divides the dataset into regions, training on all but one and testing on the excluded region. However, with large datasets like GLORIA (Lehmann et al., 2023), this approach can result in over 40 regions being used for training (Pahlevan et al., 2022), and the single left-out region may not adequately represent an extrapolation scenario. Importantly, both evaluation methods fail to explicitly account for the underlying distributions of water constituents or IOPs, making it unclear whether model generalization ability is being assessed effectively.

This gap in understanding is critical for satellite applications, where IOP distributions in new scenes may deviate significantly from training datasets, and additional layers of uncertainty, such as atmospheric correction, are introduced (Pahlevan et al., 2021).

In this study, we introduce a novel model assessment approach that tests models on water bodies not included in the training *in situ* dataset, distinguishing between similar IOP conditions (within-distribution, WD) and unseen IOP conditions (out-of-distribution, OOD). We anticipate significant disparities in performance between WD and OOD scenarios. Additionally, when incorporating a PRISMA match-up dataset for model evaluation, we expect to observe further differences between *in situ* and satellite applications, due to the uncertainties and errors associated with satellite imagery.

Finally, although PNNs quantify uncertainties, the calibration properties of these estimates are not well-characterized. Reliable uncertainty estimates are essential for model trustworthiness, especially in unknown conditions; mis-calibrated uncertainties can mislead assessments of reliability and affect downstream applications. Recent studies have begun addressing uncertainty calibration (Saranathan et al., 2024; Werther et al., 2022), but a comprehensive understanding of PNN calibration properties is lacking. Recalibration, which adjusts estimated uncertainties post hoc to better align with actual discrepancies between model estimates and *in situ* observations, can enhance reliability for models that provide distributional outputs amenable to such adjustments (Kuleshov et al., 2018). However, to date, the impact of post-hoc recalibration on the reliability of uncertainty estimates from PNNs remains unexplored in aquatic remote sensing applications.

We evaluate the uncertainty calibration properties of the five PNNs and the efficacy of applying recalibration. We expect calibration properties to vary between PNNs based on their different mechanisms for uncertainty estimation. Provided sufficient observations are available, we expect post-hoc recalibration to consistently improve the reliability of uncertainty estimates. Our exploration of these aspects advances the understanding of PNN generalization ability to estimate IOPs through hyperspectral remote sensing.

2. Datasets

2.1. In situ observations

The core datasets used in this study originate from GLORIA (Lehmann et al., 2023) and SeaBASS (Werdell et al., 2003). We extended the GLORIA dataset with observations from various sources (Appendix A). The resulting dataset comprises 2 066 *in situ* observations from 155 water bodies (lakes, coastal waters, and oceans) across 14 countries. Each observation includes spectral remote-sensing reflectance ($R_{rs}(\lambda)$) from 400 to 700 nm in 5 nm increments and absorption IOPs (a_{ph}, a_{CDOM}, a_{NAP}) at 443 and 675 nm (Fig. 1). Erroneous spectra with a Quality Water Index Polynomial (QWIP; Dierssen et al. (2022)) score exceeding ±0.2 were discarded. Both R_{rs} and IOPs were measured using diverse methodologies, which are documented in GLORIA and SeaBASS for the respective datasets, and in Appendix A for the rest.

The IOP wavelengths of 443 and 675 nm were selected because of their distinct absorption features and the greater measurement availability at these wavelengths in our dataset. Although our reflectance inputs cover the entire hyperspectral range (400–700 nm), we only estimate IOPs at 443 and 675 nm. At 443 nm, strong absorption by a_{CDOM} and a_{NAP} is observed, with these components decaying exponentially towards longer wavelengths, while a_{ph} exhibits peaks at both 443 and 675 nm. In our study, we also estimate a_{CDOM} and a_{NAP} at 675 nm to serve as a baseline for evaluating PNN uncertainty, particularly relative to the a_{ph} peak. Although 675 nm may exhibit relatively uncertain scattering components due to scattering from small particles, it offers valuable information for studying PNN generalization for absorption IOPs in the red region of the spectrum.

2.2. PRISMA match-up dataset

The PRISMA satellite, launched on March 22nd, 2019, contains a high-spectral-resolution imaging spectrometer and panchromatic camera. For this study, we selected 36 bands in the visible range from 406 to 694 nm, which contain the most relevant information content for obtaining IOPs at 443 and 675 nm. This range offers a spectral resolution of ≤12 nm (Full Width at Half Maximum) and a spatial resolution of 30 m with a swath width of 30 km (2.45° field of view). The signal-to-noise ratio (SNR) for coastal waters has been found to be 100–120 in the 450–600 nm range (Braga et al., 2022). The radiometric accuracy in this wavelength range is within 2%–7% compared to field/airborne spectroscopy (Cogliati et al., 2021).

59 *in situ* measurements from PRISMA validation campaigns including the combination of a_{ph}, a_{CDOM}, and a_{NAP} at 443 and 675 nm were available, of which 50 were match-ups with PRISMA overpasses. These measurements were made over the Venice Lagoon, Lake Garda, and Lake Trasimeno in Italy and the Curonian Lagoon in Lithuania. A 3 × 3 pixel window was extracted from the satellite scenes around each sampling station, and a spatial homogeneity check (Bailey and Werdell, 2006) was performed. Specifically, for each window, pixel values outside the median ±1.5 standard deviations were discarded. The mean of the remaining pixel values was then used for the match-up analysis. A ±1.5-hour time window aligned the field sampling with the satellite overpass (Guanter et al., 2010; Warren et al., 2019). Further details about the match-up protocols and processing details are described in Braga et al. (2022), Pellegrino et al. (2023).

PRISMA top-of-atmosphere radiance was atmospherically corrected using the standard PRISMA L2C processor (ASI, 2021) and ACOLITE (Vanhellemont and Ruddick, 2018). We chose to compare two atmospheric correction (AC) methods because of the impact of AC on downstream product quality (Braga et al., 2022; Pahlevan et al., 2021; Warren et al., 2019).

3. Methods

We designed six scenarios to assess the generalization ability of the PNNs (Section 3.1). These scenarios consist of both *in situ* and PRISMA-based applications. For each of these six scenarios, we trained and evaluated 25 independent model instances of each of the five distinct PNN architectures, thus probing both systematic differences between scenarios and architectures and variability due to random effects (Section 3.2).

3.1. Generalization scenarios

3.1.1. Random split

In the random split approach, the entire *in situ* dataset was randomly divided into two equally sized subsets for training and testing. While this method is prone to issues such as knowledge leakage (Stock et al., 2023) and spatial and temporal autocorrelation (Stock and Subramaniam, 2022), it is widely used in the literature to establish a best-case baseline for model performance (Cao et al., 2020; Pahlevan et al., 2020, 2022; Smith et al., 2021; Werther et al., 2022).

3.1.2. Within-distribution split

The within-distribution (WD) split explicitly considers the distribution of IOPs at 443 nm when dividing the entire *in situ* dataset, such that the IOP distributions in both training and test sets mirror each other closely (Fig. 2). We focused exclusively on the 443 nm wavelength for the dataset split for three reasons. Firstly, measurements at 443 nm are generally less prone to error and uncertainty. Additionally, IOPs exhibit larger optical variability at 443 nm than at 675 nm. Lastly, using three IOPs as variables instead of six reduced the complexity and computational runtime of the splitting process.

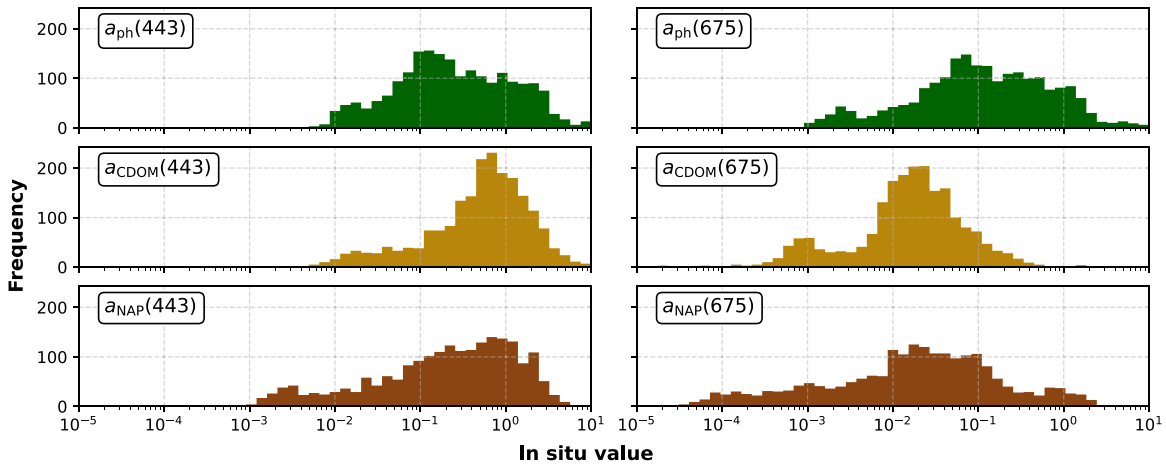

Fig. 1. Log-scaled distributions of a_{ph}, a_{CDOM}, and a_{NAP} at 443 and 675 nm in the *in situ* dataset. The number of measurements (2066) is equal across IOPs and wavelengths.

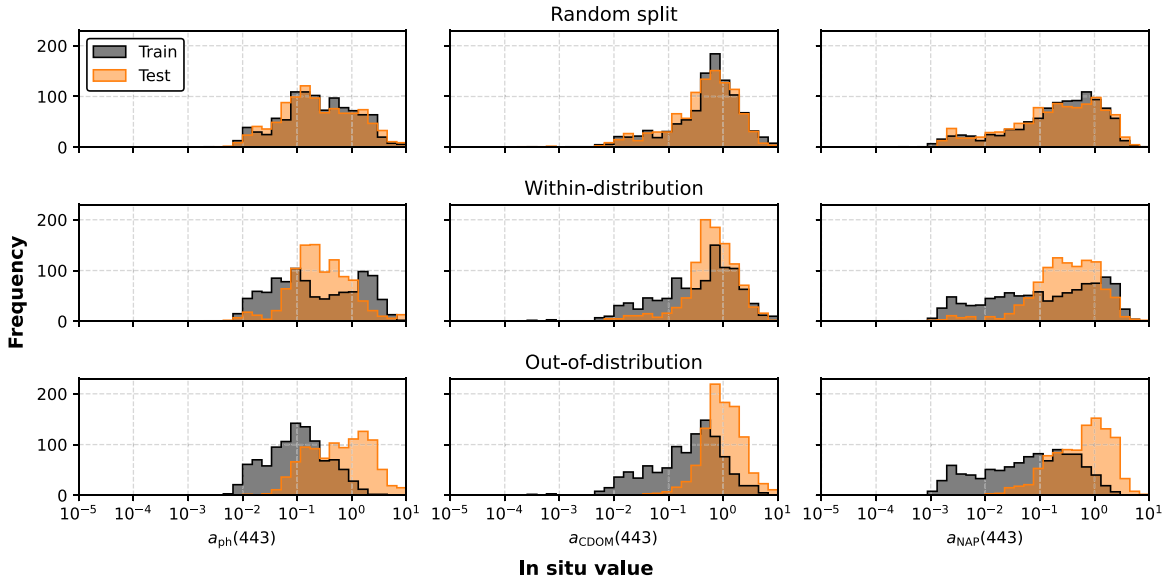

Fig. 2. Histograms illustrating the distributions of the three IOPs $a_{ph}(443)$, $a_{CDOM}(443)$, and $a_{NAP}(443)$ in the training and test sets under three splitting strategies. The top row shows a random split, while the middle and bottom rows respectively represent the within-distribution and out-of-distribution splits generated by the dataset splitting algorithm.

To prevent knowledge leakage, we ensured that all observations from a specific system (water body) were grouped together, residing entirely in either the training set or the test set, but not both. By training on observations from a set of systems and testing on completely separate systems, the models are evaluated on their interpolation ability to generalize to new waters that are similar to known waters.

3.1.3. Out-of-distribution split

The out-of-distribution (OOD) split evaluates the ability of a model to generalize when confronted with observations that differ substantially from those seen during training. Similar to the WD split, we used IOPs at 443 nm to partition the entire *in situ* dataset into training and test sets. However, the OOD split maximizes the dissimilarity between the IOPs in these sets. Consequently, the test set includes observational properties and IOP combinations that are either absent or underrepresented in the training set (Fig. 2). Such scenarios are common in satellite remote sensing. For example, a model might encounter entirely new phytoplankton species compositions (absent case) or optical properties affected by extreme weather or climate change (underrepresented case). Unlike the WD split, the OOD split challenges models to extrapolate their learned knowledge not just to independent waters, but also to entirely new or significantly underrepresented biogeochemical and optical conditions.

3.1.4. Dataset splitting algorithm

The primary goal of the dataset splitting algorithm is to automatically separate a dataset into training and testing subsets with distributions that are either highly similar (WD) or dissimilar (OOD), while ensuring that each water body is assigned entirely to one subset and that both subsets are as balanced in sample size as possible. To achieve this, we adopt a dual annealing optimization strategy, which combines global exploration with local search (Tsallis and Stariolo, 1996).

The algorithm proceeds in the following stages:

1. **Initialization:** Each water body (uniquely identified, e.g., by its name) is initially assigned at random to either the training or test set.
2. **Objective function:** The quality of the current split is evaluated with an objective function that includes:
   - For WD: A mean-based similarity score computed over selected summary columns (here a_{ph}, a_{CDOM}, a_{NAP}).

- For OOD: A percentile-based dissimilarity score computed using the same summary columns.

Here, the IOPs at 443 nm are used as summary columns; we excluded the 675 nm band to reduce complexity. In both cases, an additional penalty is imposed based on the imbalance between the subsets (i.e., the difference in number of observations), thus favoring splits with near-equal sample sizes.

3. **Global search via dual annealing:** Rather than performing explicit swaps of water bodies between the subsets, the optimization process searches over possible selections of water bodies for the training set by minimizing the objective function using dual annealing. This approach balances global search with local refinements.
4. **Convergence:** The optimization runs under a fixed time budget (here, 10 min), and the process terminates once this time limit is reached.

By integrating distribution-based objectives directly into the dual annealing optimization process, our method yields deterministic and reproducible splits that avoid knowledge leakage. The method can be flexibly adapted to other datasets and variables, provided that split variables are defined. Further details, including the full objective functions, are provided in Appendix B, and our Python implementation is described in the Code Availability section.

3.1.5. PRISMA application

We evaluated the PNNs using PRISMA to understand their generalization properties in the context of a spaceborne hyperspectral application. This evaluation was conducted through three different, PRISMA-specific scenarios:

1. *In situ* vs. *in situ*: Training on the *in situ* dataset ($n = 2034$), resampled to the spectral response function (SRF) of PRISMA at 406–694 nm (Section 2.2), and then applying the models to the *in situ* dataset accompanying the PRISMA match-ups ($n = 59$). We note that 32 of the *in situ* spectra could not be resampled due to spectral range limits. This approach assessed model performance on resampled reflectance aligned with the spectral characteristics of PRISMA, without introducing additional uncertainty through prior AC. This scenario serves as a baseline for the following scenarios.
2. General: Training on the *in situ* dataset, resampled to the PRISMA SRF, and then applying the models to the atmospherically corrected R_{rs} from PRISMA. The performance was evaluated against the corresponding match-up *in situ* IOPs. This scenario represents the general satellite application, where the model encounters hyperspectral imagery without prior specific knowledge.
3. Local knowledge: Training on a combination of the full *in situ* dataset ($n = 2034$) along with the local *in situ* dataset ($n = 59$) accompanying the PRISMA match-ups (in total $n = 2093$), and then applying the models to the atmospherically corrected R_{rs} from PRISMA. This scenario evaluates the impact of incorporating local knowledge on model generalization.

3.2. Probabilistic neural networks

A standard neural network computes an output y from an input x using a function f parameterized by weights θ:

$$y = f(x;\theta) \tag{1}$$

where y in this study is the vector of the six IOPs and x is the pre-processed input vector $R_{rs}(\lambda)$. The pre-processing of the input and target variables is described in Appendix C.

PNNs modify this approach by estimating a probability distribution over the possible outcomes for each output variable, here each IOP. This is achieved using Bayesian methods or other distributional techniques. Instead of yielding a single point estimate for each IOP, a PNN simultaneously estimates the mean μ and variance σ^2 of the estimated output distribution:

$$(\mu, \sigma^2) = f(x;\theta), \quad y \sim \mathcal{N}(\mu, \sigma^2). \tag{2}$$

The output variables may be estimated individually or simultaneously; here, we estimated the absorption IOPs simultaneously to account for their correlated nature. Simultaneous estimation enables the model to capture the relationships between IOPs, which has been shown to be advantageous compared to individual retrieval (Cao et al., 2022; Pahlevan et al., 2022; Saranathan et al., 2024).

We implemented five PNN architectures encompassing a broad spectrum of state-of-the-art methodologies (Fig. 3). The PNNs were constructed around the same core NN structure, defined as the shared foundational neural network comprising dense layers, activation functions, and hyperparameters such as the number of neurons and total layers. All five PNNs were trained to estimate the six IOPs from R_{rs} inputs. To ensure consistent comparison between PNNs, we standardized their architecture in terms of the number of neurons, layers, and other hyperparameters such as learning rates. This uniformity enables the individual assessment of model estimation capabilities, controlling for potential variation arising from architectural differences. Details, including model training, overfitting and hyperparameters are given in Appendix D.

For each of the generalization scenarios (Section 3.1), we trained 25 instances of each PNN to probe the effects of random initialization of network weights, variability in model training convergence, and effectiveness of regularization mechanisms (Smith et al., 2021). Results in Section 4 are presented for all 25 model instances, with the median value and $k = 1$ or 1-σ confidence interval (CI). When analyzing all 25 model instances would introduce excessive complexity, we focus on the **median-performing model**, defined as the model instance with the median composite score derived by summing the median symmetric accuracy (MdSA; Morley et al. (2018)) values for all IOPs at 443 nm (i.e., the 13th instance when sorted by this composite score). Standardizing comparisons to the central tendency (median behavior) of the 25 model instances per PNN architecture ensures a consistent and equitable evaluation of model performance across scenarios under identical statistical criteria, reducing potential bias introduced by outlier model instances.

3.2.1. Bayesian neural network with Monte Carlo dropout

The Bayesian Neural Network with Monte Carlo Dropout (BNN-MCD) employs dropout layers that randomly deactivate 25% of neurons during both training and application (Gal and Ghahramani, 2016a). This process enables Monte Carlo sampling by simply estimating each output multiple times (Section 3.3), approximating the posterior distribution of the model outputs and providing uncertainty estimates. BNN-MCDs have previously been applied to water quality parameter and IOP estimation (Saranathan et al., 2024; Werther et al., 2022).

3.2.2. Bayesian neural network with Monte Carlo DropConnect

The Bayesian Neural Network with Monte Carlo DropConnect (BNN-DC) applies a stochastic principle similar to BNN-MCD, but uses DropConnect instead of Dropout (Wan et al., 2013). DropConnect randomly sets weights between neurons to zero, allowing for a finer-grained exploration of the neural configuration space compared to BNN-MCD. To our knowledge, the application of a BNN-DC for IOP retrieval is not documented in the aquatic remote sensing literature.

3.2.3. Mixture density network

The Mixture Density Network (MDN) uses a deterministic core NN to estimate parameters for a number of mixture distributions (here five), thereby forming a Gaussian mixture model (GMM). This GMM is trained using maximum likelihood estimation (MLE) where the network

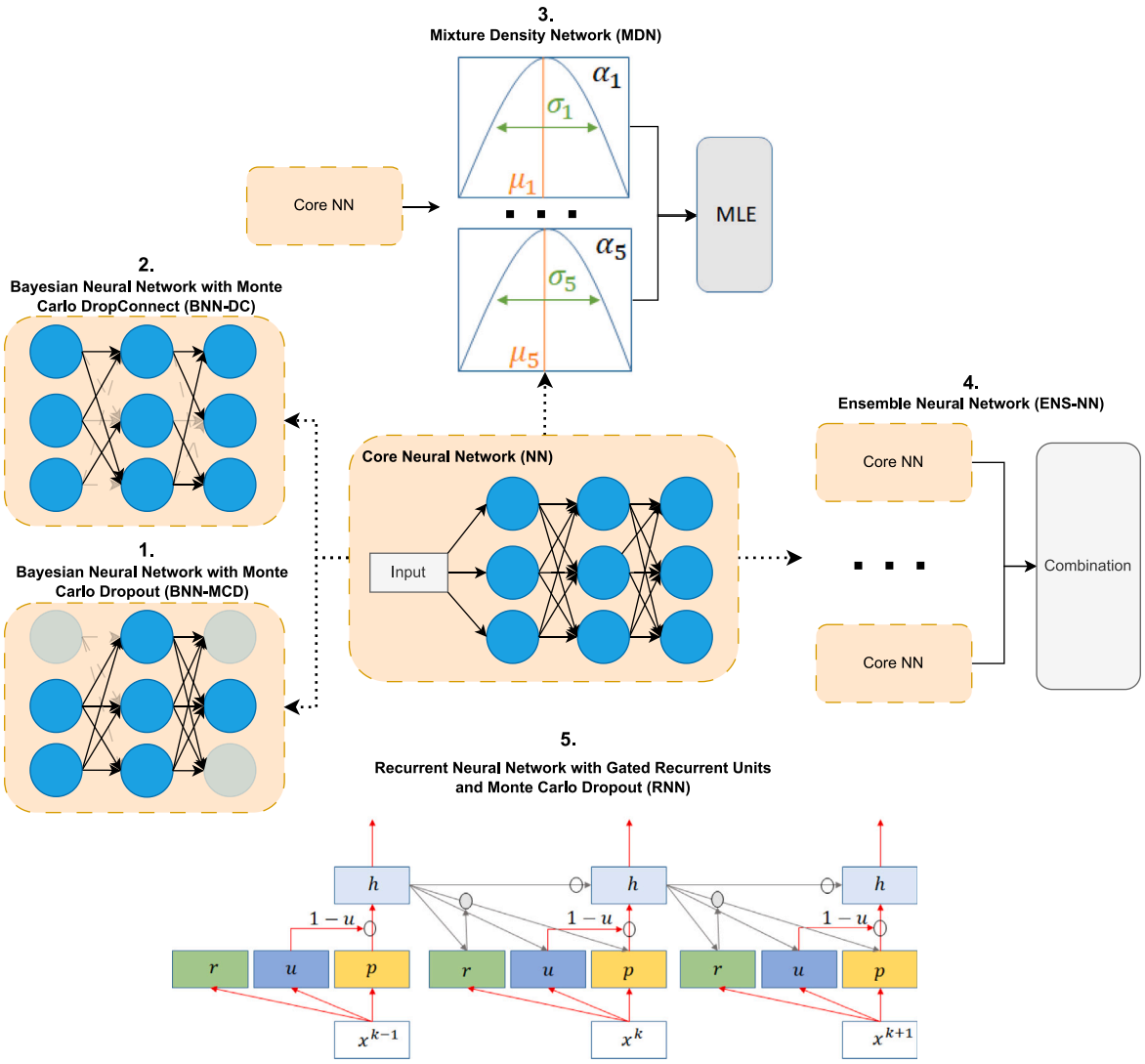

Fig. 3. Schematic illustration of the PNN architectures used in this work. For the RNN, a spectral band k is used in reset and update gates denoted by r and u, respectively, and p and h are the resultant proposal and final activations. For simplicity, the RNN scheme does not depict multiple GRU layers and Dropout in between, as in the code implementation (see Code Availability section).

parameters are optimized to maximize the likelihood of the observed target variables (Bishop, 1994; Pahlevan et al., 2020; Saranathan et al., 2023). During inference, a point estimate is approximated by taking the mean of the Gaussian component with the largest weight. The MDN approach captures multi-modal characteristics of reflectance spectra and models correlations between target IOPs using a full covariance matrix constructed via Cholesky decomposition (Pahlevan et al., 2022; O'Shea et al., 2023).

3.2.4. Ensemble neural network

The Ensemble Neural Network (ENS-NN) aggregates outputs from several (here 10) individual core neural networks. This architecture, similar to earlier works (Bricaud et al., 2007; Brockmann et al., 2016; Hieronymi et al., 2017), aims to enhance estimation accuracy and reliability by overcoming the limitations of single-model estimates (Lakshminarayanan et al., 2017; Schaeffer et al., 2022a; Werther et al., 2021).

3.2.5. Recurrent neural network with gated recurrent units and Monte Carlo dropout

As the final PNN, we introduce the Recurrent Neural Network (RNN) equipped with Gated Recurrent Units (GRUs) and Monte Carlo Dropout. Although RNNs have been widely developed for hyperspectral remote sensing (Mou et al., 2017; Li et al., 2019), their use for IOP estimation in aquatic remote sensing has not been explored.

RNNs are particularly suited to modeling the sequential nature of hyperspectral reflectance spectra, where strong correlations exist between adjacent spectral bands (Cael et al., 2023). This sequential modeling capability became particularly relevant in aquatic remote sensing after the advent of hyperspectral satellite sensors in 2019.

GRUs address challenges that traditional RNNs face, such as the vanishing gradient problem, by implementing gating mechanisms (Cho et al., 2014). The reset gate evaluates how much information from the previous spectral band should be forgotten, while the update gate balances information carried over from previous bands with the current input. This process ensures that each state in the sequence is a well-balanced representation of past and present information, allowing the network to effectively capture and model the complex dependencies in hyperspectral sequences (Chung et al., 2014). Within the GRU architecture, we implemented MCD with a 25% chance to enable uncertainty estimation (Gal and Ghahramani, 2016b), similar to the BNN-MCD and BNN-DC.

3.3. Uncertainty estimation

Using a negative log-likelihood loss function (Appendix E), the PNNs were trained to estimate mean values μ and associated variances σ^2 for the IOPs at 443 and 675 nm (Section 3.2). For the BNN-MCD, BNN-DC, and RNN, the estimation was performed 100 times per output, with Dropout or DropConnect providing a different network configuration each time. This process produces different estimates for μ and σ^2, sampling the posterior distribution of the model (Valdenegro-Toro and Mori, 2022). The mean output $\bar{\mu}$ is the mean of the 100 sample means μ_i:

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mu_i \tag{3}$$

The total estimated variance σ^2_{tot} consists of two terms, namely the aleatoric and epistemic variance (Section 1). The aleatoric variance is what the individual networks estimate, and is thus calculated from the mean of the individual estimates σ_i^2:

$$\sigma^2_{\text{alea}} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 \tag{4}$$

The epistemic variance represents the uncertainty due to the model configuration, and is calculated from the variance in the individual sample means μ_i:

$$\sigma^2_{\text{epi}} = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2 \tag{5}$$

The total variance is the sum of the aleatoric and epistemic variances (Valdenegro-Toro and Mori, 2022), as in Eq. (6). The uncertainty on individual measurements is the square root of the total variance, σ_{tot}:

$$\sigma^2_{\text{tot}} = \sigma^2_{\text{alea}} + \sigma^2_{\text{epi}} \tag{6}$$

A similar process is performed with the ENS-NN, but using its 10 component NNs to produce 10 samples of μ and σ^2. The MDN instead employs a GMM to estimate the mean and variance (Section 3.2.3).

3.4. Evaluation metrics

3.4.1. IOP estimation accuracy

We assessed PNN IOP estimation accuracy through three common metrics (Morley et al., 2018), namely the median symmetric accuracy (MdSA, [%]), symmetric signed percentage bias (SSPB, [%]) and coefficient of determination (R^2). MdSA and SSPB were calculated using a direct comparison between *in situ* observations and PNN estimates. R^2 was calculated on the 10-logs of the observations and estimates.

3.4.2. Uncertainty calibration

We used two metrics to evaluate the uncertainty calibration of the PNNs, namely coverage and miscalibration area.

Coverage is the probability that an estimation interval $\mu \pm k\sigma$ contains the observed value $y_{in\ situ}$ (Stoudt et al., 2021), as in Eq. (7) with $\mathbb{1}$ the indicator function. A well-calibrated normal distribution has a coverage of 68%, corresponding to $k = 1$ or a 1-σ uncertainty. Coverage >68% indicates underconfidence, meaning the estimated uncertainties are systematically larger than the actual errors, while <68% indicates overconfidence:

$$C = \frac{100}{n}\sum_{i=1}^{n}\mathbb{1}\left(y_{in\ situ,i} \in [\mu_i - k\sigma_i,\ \mu_i + k\sigma_i]\right) \quad [\%] \tag{7}$$

Miscalibration area (MA) quantifies the agreement between accuracy and estimated uncertainty on varying scales. The observed fraction of errors within quantiles $\alpha \in [0, 1]$ is compared to the fraction expected from the estimated uncertainty, assuming a normal distribution (Rasmussen et al., 2023). This is essentially a generalization of coverage.

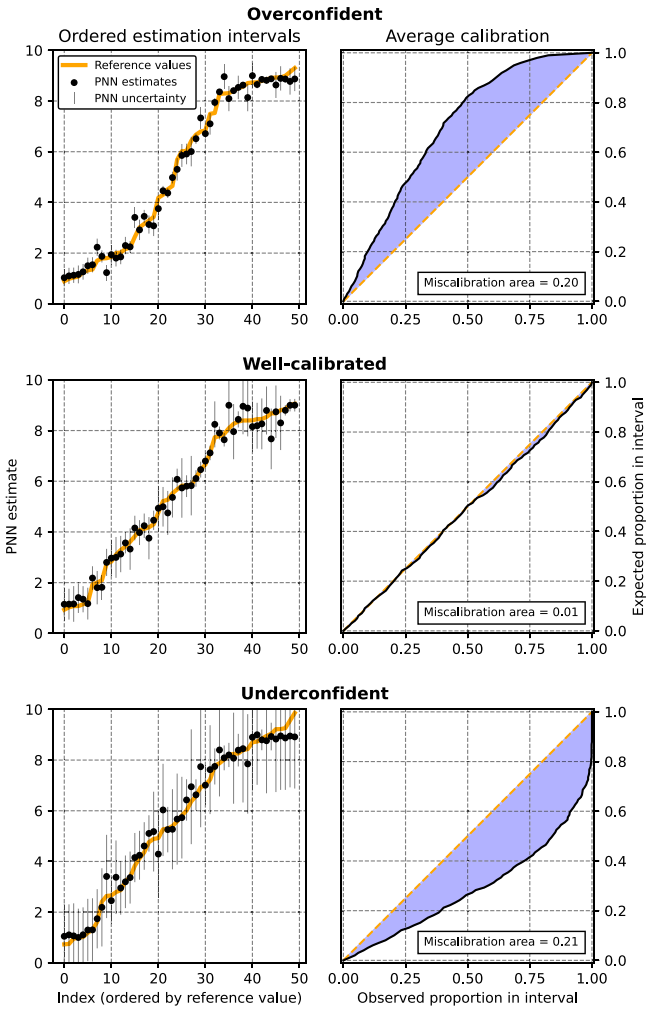

Fig. 4. Comparison between calibration curves using overconfident, well-calibrated, and underconfident toy models with synthetic observations. Left: Ordered estimation intervals with reference values and PNN estimates. Right: Calibration curves with miscalibration area shaded in blue.

Plotting the observed vs. expected fraction of errors produces a calibration curve (Fig. 4), which is diagonal for perfect uncertainty calibration and curved otherwise. MA is the area between the calibration curve and the diagonal, ranging from 0 (perfect calibration) to 0.5 (extremely over- or underconfident). We used the Uncertainty Toolbox (Chung et al., 2021) implementation to calculate calibration curves and MA.

3.5. Uncertainty recalibration

Recalibration consists of first estimating IOPs and uncertainties using a machine learning model capable of uncertainty quantification and then adjusting these uncertainties with fitted recalibration functions. The scenario-specific training set was randomly partitioned into 80% training and 20% recalibration parts. PNNs were trained on the former and then applied to the latter to obtain calibration curves, to which isotonic functions – one for each IOP for each PNN instance – were fit. Isotonic functions are non-parametric and monotonically increasing, allowing them to approximate calibration curves of arbitrary shapes (Kuleshov et al., 2018). The efficacy of recalibration was assessed by comparing the IOP and uncertainty estimates from the 25 regular PNNs and the 25 recalibrated PNNs on the test set.

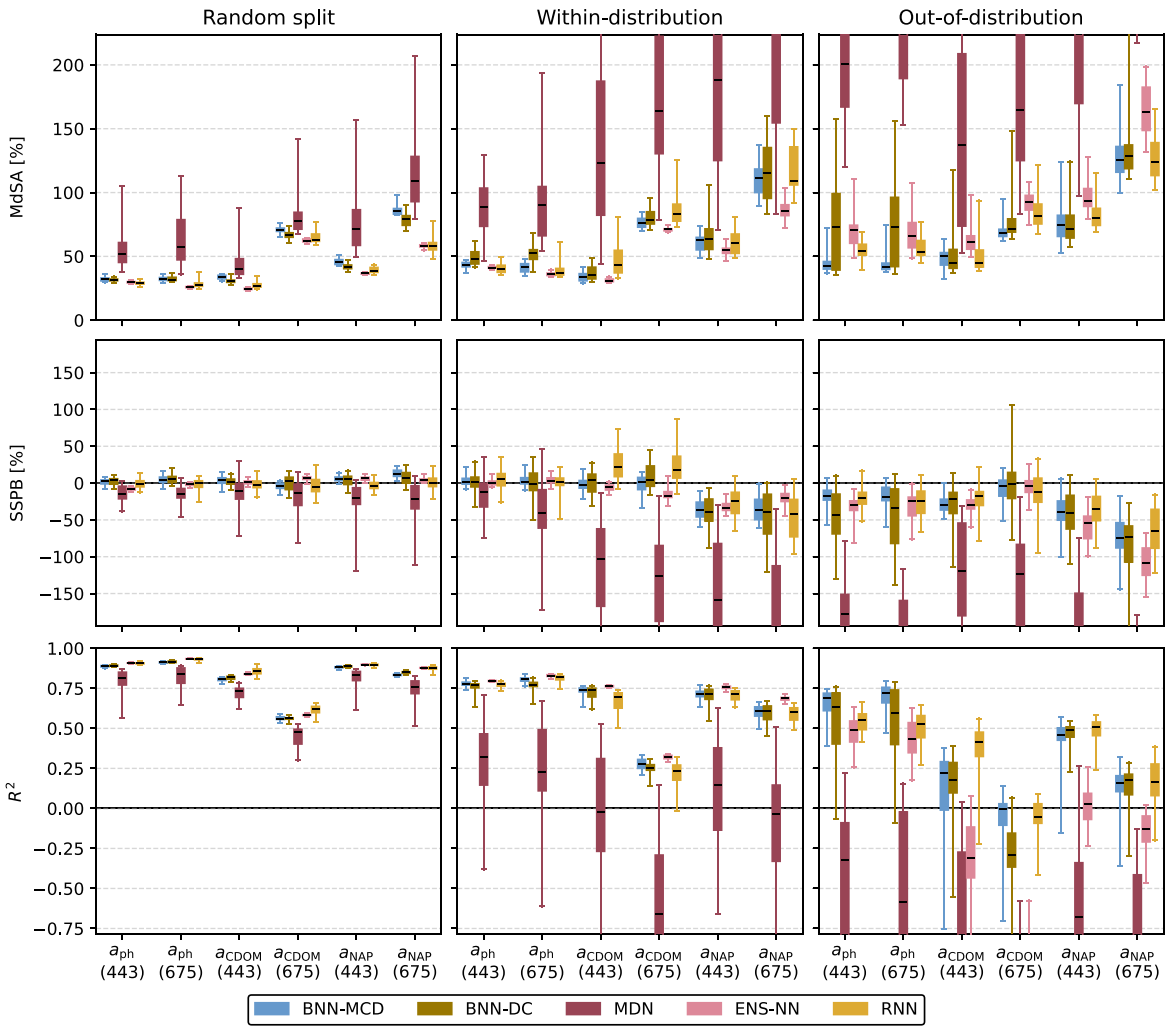

Fig. 5. Accuracy metrics for the PNNs in the random, within-distribution, and out-of-distribution split scenarios, grouped by IOP. The boxplots represent the range across the 25 instances of each PNN, with the box spanning the interquartile range and the whiskers spanning the full range.

4. Results

4.1. *In situ* scenarios

4.1.1. Estimation accuracy

The accuracy of the PNNs in estimating IOPs varied across the three *in situ* scenarios (Fig. 5). In the random split scenario, the accuracy showed minor variations across all models. For all 25 instances of each architecture, the MdSA was $\leq$ 38% for $a_{ph}(443)$, $a_{ph}(675)$, and $a_{CDOM}(443)$, with the median MdSA ranging from 25%–34%. The MDN consistently exhibited larger errors than the other models, prompting a more detailed analysis in Section 4.1.2. Errors from all PNNs for $a_{CDOM}(675)$ and a_{NAP} at 443 and 675 nm were notably larger, with errors in a_{CDOM} and a_{NAP} at 675 nm approximately double the error in a_{ph}. Despite these variations, all models demonstrated the ability to estimate the six IOPs without significant positive or negative bias, evidenced by the SSPB values clustering around zero.

The WD split scenario was characterized by elevated inaccuracy across all models and IOPs compared to the random split. MdSA ranged from 31%–53% for $a_{ph}(443)$, $a_{ph}(675)$, and $a_{CDOM}(443)$ and from 55%–64% for $a_{NAP}(443)$, increasing substantially to 71%–115% for a_{CDOM} and a_{NAP} at 675 nm. Again, the MDN performed worse than the other model types (Section 4.1.2). The similarity in patterns between random split and WD results establishes a baseline for expected performance when the models interpolate within familiar IOP ranges.

The median MdSA values in the OOD split scenario were up to 30 percent point (%pt.) larger than in the WD split for the IOPs with relatively low MdSA (a_{ph} and $a_{CDOM}(443)$). For $a_{CDOM}(675)$ and $a_{NAP}(675)$, the difference in MdSA between WD and OOD was generally smaller. Importantly, the OOD scenario had consistent negative bias across most variables and models. The SSPB matches the discrepancy between the training and test IOP distributions, the training set containing smaller IOP values than the test set (Fig. 2). Consequently, the models failed to generalize effectively, leading to underestimation of larger IOP values. This outcome demonstrates a limitation in model ability to generalize to truly OOD conditions. These findings were corroborated by R^2 showing a clear degradation from consistently good fits in the random split scenario to more scattered and sometimes negative values (as explained in Kvålseth (1985)) in the OOD scenario.

Variations in accuracy among the 25 instances of each PNN were small for the random split scenario (standard deviations of 2%–11% of the median MdSA for each IOP-architecture pair) but became more pronounced in the WD (2%–31%) and OOD (10%–52%) scenarios. Variability in median MdSA between the four PNN architectures (all but MDN, Section 4.1.2) increased from the random split (3%–27%) through WD (8%–29%) to the OOD scenario (17%–39%), being similar for a_{ph} and a_{CDOM} but higher for a_{NAP}.

4.1.2. MDN analysis

A sensitivity study was conducted to explore the causes behind the larger errors and variability exhibited by the MDN compared to the

other PNNs (Appendix F). The training and evaluation process was repeated without the variables $a_{CDOM}(675)$ and $a_{NAP}(675)$, which are naturally more prone to uncertainty and error in both observations and model application. The 4-IOP MDN was significantly more accurate than the 6-IOP one. For example, the random split $a_{ph}(443)$ MdSA decreased from 52% (CI 41%–75%) to 36% (CI 33%–42%) and the WD $a_{ph}(443)$ MdSA decreased from 89% (CI 65%–111%) to 55% (CI 45%–74%). When the sensitivity analysis was repeated for the other four PNNs in the 4-IOP configuration, no comparable changes were observed, indicating that the MDN was uniquely sensitive to the choice of output variables.

We also investigated the impact of Dropout regularization on the MDNs, following Saranathan et al. (2024). Unlike in the BNN-MCD and RNN models, Dropout was applied solely as a regularizer during training, not for inference, as MDNs estimate uncertainty through their mixture of probability distributions. 25 new MDN instances were trained with Dropout layers inserted after each dense layer of neurons. In the random split scenario, MDNs with Dropout demonstrated more consistent and marginally improved accuracy metrics. For $a_{ph}(443)$, the MdSA reached 45% (CI 43%–52%) with Dropout, compared to 52% (CI 41%–75%) without. However, Dropout had insignificant or even adverse effects in the WD and OOD scenarios. For $a_{ph}(443)$, the MdSA increased from 89% (CI 65%–111%) to 114% (CI 91%–122%) in the WD scenario, and from 201% (CI 150%–404%) to 417% (CI 301%–533%) in the OOD scenario. Analogous trends were observed across the other IOPs, except $a_{CDOM}(443)$.

These outcomes can be attributed to the characteristics of each scenario. The random split scenario suffers from knowledge leakage leading to overfitting (Section 3.1.1). Dropout regularization mitigated MDN overfitting, thus reducing intra-model variability and slightly increasing accuracy. Conversely, the WD and OOD scenarios inherently prevent knowledge leakage, rendering additional model regularization potentially counterproductive. The large epistemic uncertainty exhibited by the MDN in these scenarios (Section 4.1.3, Fig. 6) suggests that it was operating at the limits of its available model knowledge and would benefit from additional training observations. Consequently, the introduction of Dropout layers in the WD and OOD scenarios resulted in over-regularization (causing underfitting), thereby increasing intra-model variability and decreasing accuracy.

The following sections include further results describing the sources of error and variability specific to the MDN.

4.1.3. Estimated uncertainty

The total estimated uncertainty varied considerably across scenarios, PNN architectures, and IOPs (Fig. 6). Generally, IOPs at 443 nm and a_{ph} at 675 nm exhibited the smallest uncertainties, reflecting model sensitivity to these variables. For the average-performing model, the total relative uncertainties in the WD scenario were typically 1.2–2.6× larger than in the random split scenario, while OOD uncertainties were typically 1.2–6.2× those of the random split. ENS-NN uncertainties were typically similar to those from the other networks, but showed extreme spikes in some cases, such as $a_{CDOM}(675)$ in Fig. 6. MDN estimates differed significantly from the others, as in Section 4.1.1.

Across all scenarios, aleatoric uncertainty dominated the total uncertainty for most models, consistently exceeding 89%. This predominance indicates that the estimated uncertainty primarily stemmed from inherent variability in the input reflectance spectra, rather than insufficient training observations (Hüllermeier and Waegeman, 2021). Consequently, for these models, additional training observations from the same distribution would not reduce total uncertainty. Only the MDN exhibited substantial epistemic uncertainty for several IOPs and scenarios, indicating that the model lacked sufficient knowledge about the test set conditions. This finding partially explains the variability observed among MDN instances (Fig. 5).

The coverage, which expresses the alignment between accuracy and uncertainty (Section 3.4.2), was greater than 68% ($k = 1$) for nearly all models and IOPs in the random split scenario (Fig. 7), with a median coverage of 88% (CI 81%–94%), indicating significant underconfidence. The same was true for all but a handful of models in the WD scenario (median 90%; CI 80%–95%). While the OOD scenario resulted in a wide spread from extremely underconfident (100%) to extremely overconfident model instances (45%), it also resulted in the most instances falling near the optimal value of 68%, with a median coverage of 88% (CI 71%–98%). In conclusion, while the OOD scenario yielded the poorest accuracy across all models, the PNNs demonstrated some awareness of the reduced accuracy and adjusted their uncertainty estimates accordingly.

4.1.4. Uncertainty recalibration

The calibration curves (Fig. 8) confirmed the findings from Section 4.1.3, namely that the models tended towards underconfidence. This was again especially true for the random split and WD scenarios, as well as for $a_{CDOM}(675)$ and $a_{NAP}(675)$. Consequently, most PNNs, notably except the RNN, displayed large (≥ 0.1) miscalibration areas (MA; Fig. 9).

Recalibration (Section 3.5) proved beneficial in the random split and WD scenarios, yielding coverage values close to 68% ($k = 1$) across the IOPs, albeit with residual underconfidence and occasional overconfidence (Fig. 7). Measured by MA (Fig. 9), the random split scenario showed the highest percentage of beneficial recalibrations (96.7%), with a median improvement of -0.110 (CI -0.189 to -0.035). Similarly, the median coverage decreased from extremely underconfident (88%) to mildly underconfident (median 74%; CI 68%–79%). Recalibration was similarly effective for the WD split, decreasing the MA in 90.7% of the cases, with a median decrease of -0.145 (CI -0.253 to -0.026), and lowering the median coverage from 90% to 61% (CI 54%–69%), more closely matching the desired 68%.

Already well-calibrated uncertainties rendered recalibration less effective in the OOD scenario. As a result, only 70.8% of the OOD model instances showed improvement, with a median difference in MA of -0.094 (CI -0.272 to 0.064). These values suggest that recalibration was detrimental to many PNN instances. The median coverage shifted from underconfident (88%; CI 71%–98%) to mildly overconfident (53%; CI 44%–67%), in both cases with much wider CIs than the random split and WD scenarios.

To determine a threshold for recalibration efficacy, we analyzed the MA difference as a function of the MA without recalibration. For simplicity, we compared non-recalibrated and recalibrated models in a 1-to-1 manner (e.g., the first OOD RNN instance $a_{ph}(443)$ MA vs. the first recalibrated OOD RNN instance $a_{ph}(443)$ MA). Although the model instances were trained independently, meaning there was no true 1-to-1 relationship, this comparison effectively samples the underlying populations randomly 25 times. A more rigorous analysis would involve comparing all 25 × 25 pairs of model instances, but the present approach yielded an adequate estimate.

Analysis of the MA difference, binned in 0.01 MA intervals and aggregated across all scenarios, PNNs, and IOPs, revealed a clear relationship (Fig. 10). The binned CI upper limit was < 0, indicating recalibration benefited a clear majority of comparisons, when the initial MA was ≥ 0.13.

Despite recalibration reducing the training set size by a fifth, no notable decrease in accuracy was observed (Table 1). The difference in MdSA between non-recalibrated and recalibrated models generally fell within the range of variation observed among the 25 instances of each model (Section 4.1.1). This finding aligns with the large aleatoric uncertainty fraction, indicating that the quantity of available measurements is not the primary source of inaccuracy. The MDN, exhibiting higher epistemic uncertainty, showed larger changes when the training dataset was reduced.

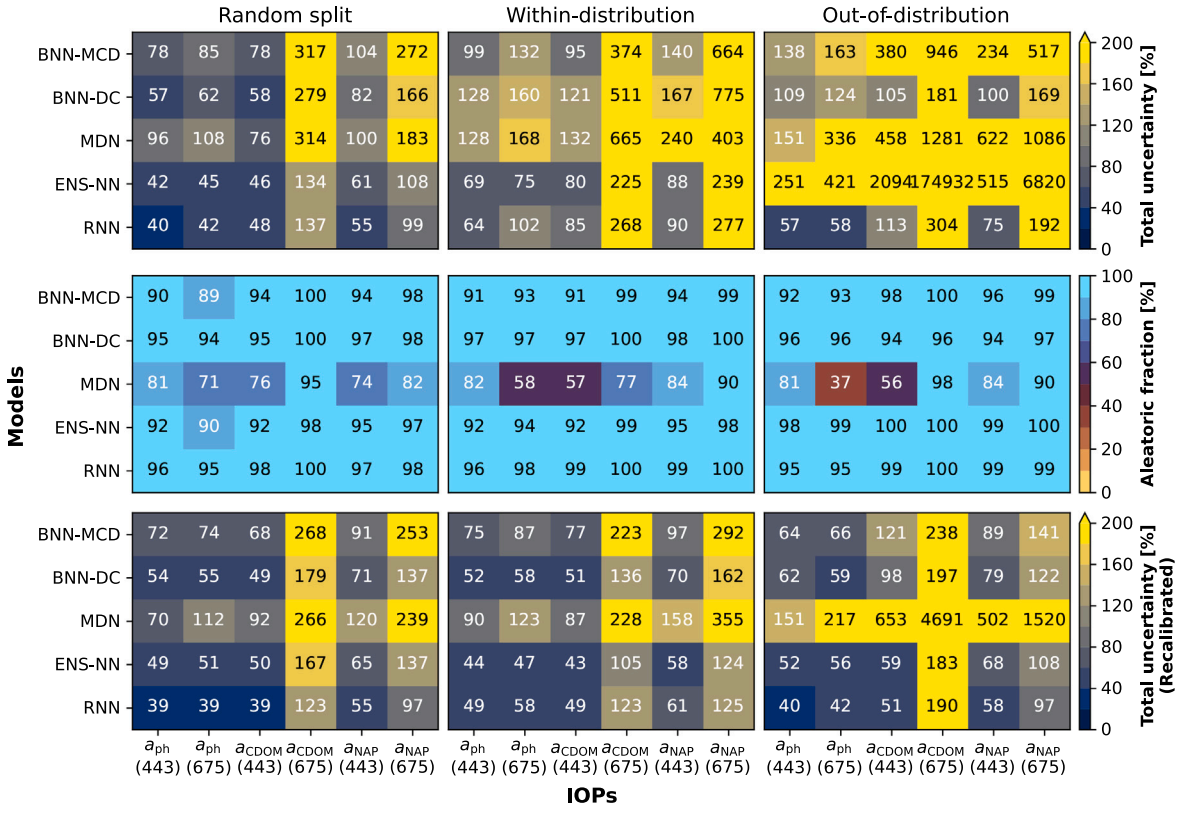

Fig. 6. Median uncertainty in PNN estimates from the average-performing models, for the *in situ* dataset. Top row: total uncertainty as the sum of epistemic and aleatoric uncertainty. Middle row: aleatoric fraction of the total uncertainty. Bottom row: total uncertainty for recalibrated models. The aleatoric fraction for the recalibrated models may differ, but is not shown here for brevity.

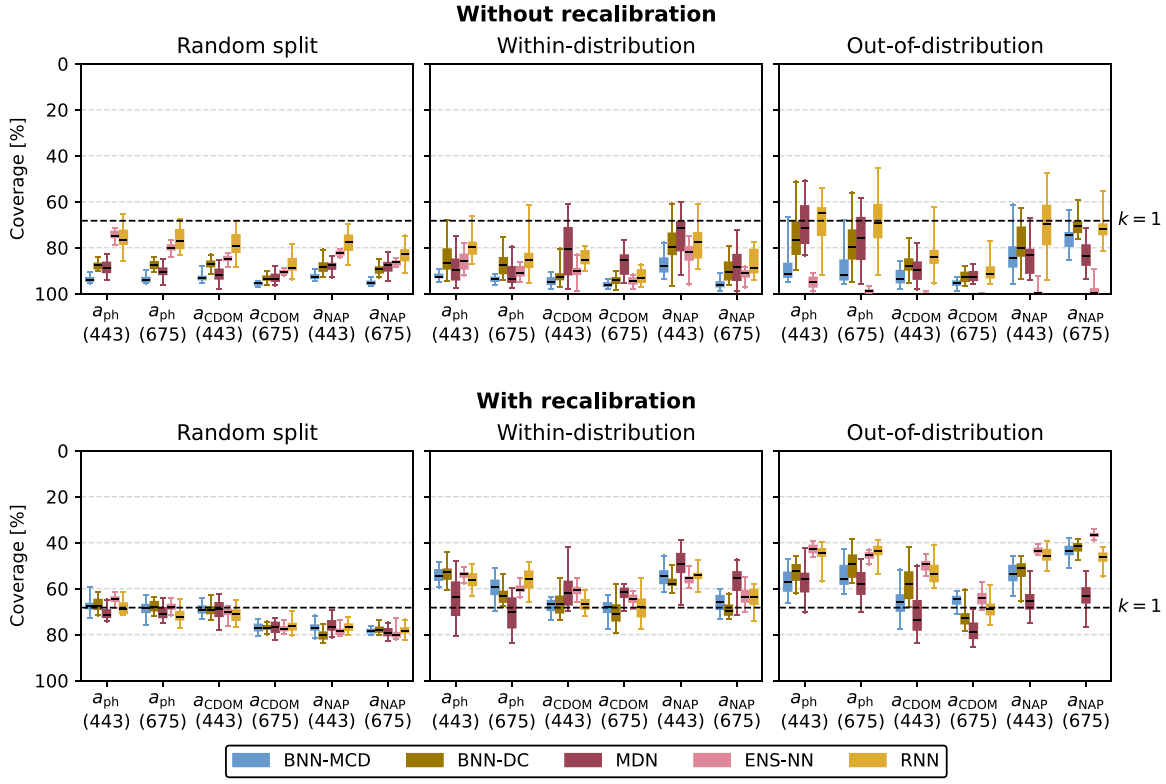

Fig. 7. Coverage of the uncertainty estimates for each PNN, scenario, and IOP. The boxplots show the observed coverage, while the dashed line indicates $k = 1$ or 1-σ coverage (68%), corresponding to a well-calibrated normal distribution (Section 3.4.2). Overconfidence (low coverage) is at the top, while underconfidence (high coverage) is at the bottom. The boxplots represent the range across the 25 PNN instances, as in Fig. 5.

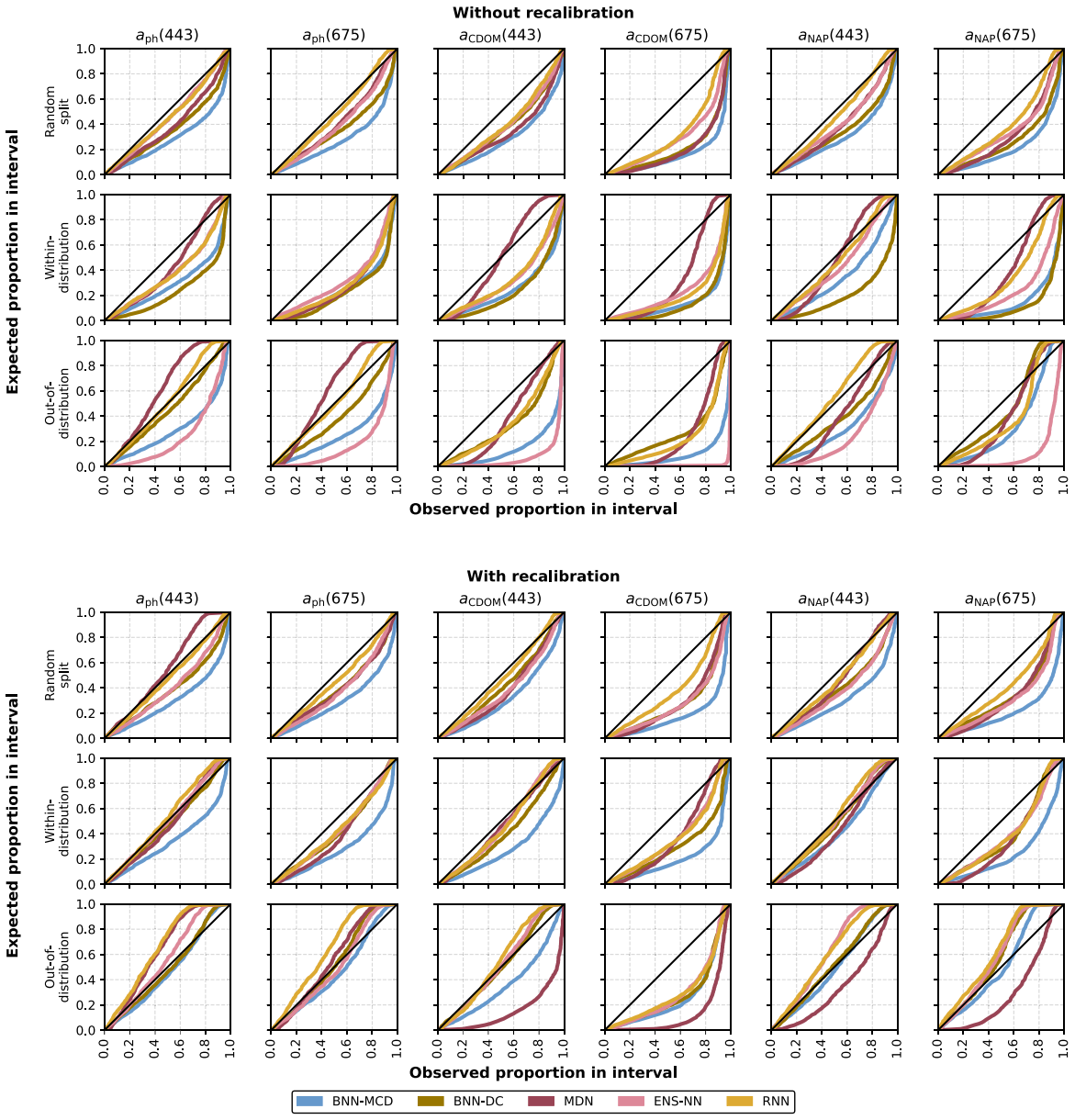

Fig. 8. Calibration curves for the *in situ* scenarios without (top) and with (bottom) recalibration. The average-performing model for each combination of scenario and architecture is displayed. We note that the axes are swapped compared to the standard orientation, so that underconfident models fall below the diagonal line and overconfident ones above it.

Table 1
Difference in MdSA [%] between the average-performing PNNs (see Section 3.2 for an explanation) without and with recalibration. A positive number, meaning an increase in MdSA, indicates a decrease in accuracy.

Scenario	PNN	$a_{ph}443$	$a_{ph}675$	$a_{CDOM}443$	$a_{CDOM}675$	$a_{NAP}443$	$a_{NAP}675$
Random	BNN-MCD	0	0	−2	0	−3	8
Random	BNN-DC	−1	−4	−1	−3	−1	−9
Random	MDN	27	−2	8	11	29	15
Random	ENS-NN	1	1	1	1	2	4
Random	RNN	3	−2	1	9	2	11
WD	BNN-MCD	−6	−13	−2	−3	−6	−17
WD	BNN-DC	−16	−30	−14	4	9	7
WD	MDN	−16	7	−59	−76	−84	−44
WD	ENS-NN	−2	−5	−1	−3	−9	4
WD	RNN	6	3	−7	0	−15	−57
OOD	BNN-MCD	−2	−5	8	−5	−9	−16
OOD	BNN-DC	−35	−25	24	−17	11	25
OOD	MDN	26	−249	−69	−26	−97	−150
OOD	ENS-NN	−14	−16	−18	−9	−12	−28
OOD	RNN	7	3	5	−5	3	36

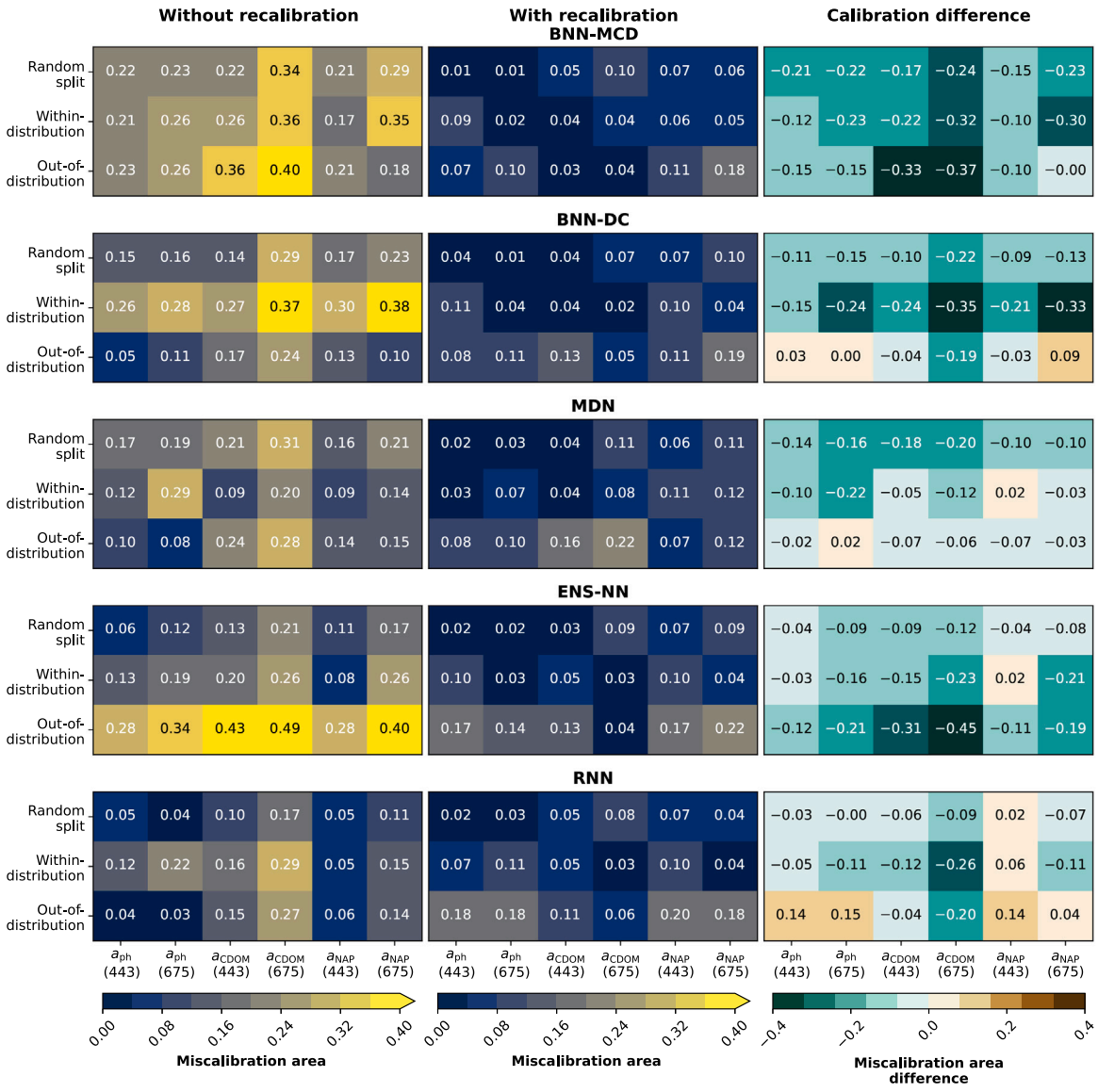

Fig. 9. Miscalibration area for the average-performing models without recalibration (left), with recalibration applied (middle), and the calibration difference (right) for the *in situ* scenarios. A negative calibration difference represents a beneficial recalibration outcome (Section 3.4.2).

4.2. PRISMA application

4.2.1. Estimation accuracy

Model accuracy was evaluated across three PRISMA scenarios. In the general and local knowledge scenarios, models were tested using both L2C and ACOLITE AC methods, resulting in five distinct scenario outcomes (Table 2). In the *in situ* vs. *in situ* baseline scenario, the combined median MdSA for non-MDN models was 67% (CI 56%–84%) in $a_{ph}(443)$ and 60% (CI 44%–94%) in $a_{ph}(675)$, comparable to the WD and OOD *in situ* scenarios (Section 4.1.1).

Estimation accuracy was worse in the general scenario using reflectance spectra obtained through AC, although the outliers overlapped (Fig. 11). The median MdSA for $a_{ph}(443)$ was similar between PRISMA imagery processed with L2C (82%; CI 67%–109%) and ACOLITE (86%; CI 65%–115%). The difference was more noticeable for $a_{ph}(675)$, at 104% (CI 76%–142%) vs. 75% (CI 53%–132%). This trend can also be seen in Fig. 11. The MdSA for a_{CDOM} and a_{NAP} in the PRISMA general AC scenario was often more than twice the values observed in the *in situ* scenarios.

Incorporating local *in situ* measurements into the training set had mixed effects on the estimation accuracy. The median MdSA in $a_{ph}(443)$ for the L2C (82%; CI 61%–99%) and ACOLITE (80%; CI 57%–106%) satellite-derived datasets were comparable to the general scenario. In $a_{ph}(675)$, the MdSA decreased to 95% (CI 63%–125%) and 66% (CI 46%–98%), respectively, the latter (ACOLITE) being comparable to the baseline scenario.

Compared to the *in situ* scenarios (Section 4.1.1), the MDN was more in line with the other four PNNs in terms of accuracy (Table 2) and variability (Fig. 11). Based on the large epistemic uncertainty in the *in situ* scenarios, the improvement in accuracy can be attributed to the fact that the training set was much larger – incorporating the entire *in situ* dataset rather than only half – in the PRISMA application. This need for more observations can in turn be explained by the MDN methodology, since it estimates covariance matrices rather than mean–variance pairs. The discrepancy in accuracy in the *in situ* scenarios can therefore be attributed to the splitting properties. This result underlines the need for a large dataset when training MDNs for simultaneous IOP estimation.

4.2.2. Estimated uncertainty

The estimated uncertainties for a_{ph} varied widely across the PNN architectures and PRISMA scenarios (Fig. 12), typically larger than 100% but with some outliers in either direction. The MDN estimated

Table 2
MdSA [%] of the average-performing PNNs (see Section 3.2 for an explanation) for the PRISMA scenarios.

Model	Scenario	AC	$a_{ph}443$	$a_{ph}675$	$a_{CDOM}443$	$a_{CDOM}675$	$a_{NAP}443$	$a_{NAP}675$
BNN-MCD	*in situ* vs. *in situ*	–	68	53	68	136	110	325
BNN-MCD	General	L2C	87	56	125	175	74	286
BNN-MCD	General	ACOLITE	81	67	82	179	132	511
BNN-MCD	Local knowledge	L2C	73	101	110	147	83	243
BNN-MCD	Local knowledge	ACOLITE	88	41	99	152	98	326
BNN-DC	*in situ* vs. *in situ*	–	69	64	78	134	126	549
BNN-DC	General	L2C	89	105	97	197	108	367
BNN-DC	General	ACOLITE	78	69	84	174	137	533
BNN-DC	Local knowledge	L2C	89	81	82	180	103	304
BNN-DC	Local knowledge	ACOLITE	106	43	79	159	67	205
MDN	*in situ* vs. *in situ*	–	111	85	89	127	92	492
MDN	General	L2C	90	117	92	180	170	1219
MDN	General	ACOLITE	98	222	139	430	89	160
MDN	Local knowledge	L2C	101	70	112	158	139	223
MDN	Local knowledge	ACOLITE	89	68	115	165	86	458
ENS-NN	*in situ* vs. *in situ*	–	68	67	59	120	101	321
ENS-NN	General	L2C	90	103	149	158	90	249
ENS-NN	General	ACOLITE	57	64	105	179	94	468
ENS-NN	Local knowledge	L2C	91	114	160	153	64	245
ENS-NN	Local knowledge	ACOLITE	71	35	89	143	71	201
RNN	*in situ* vs. *in situ*	–	83	159	49	123	156	825
RNN	General	L2C	84	69	114	104	127	523
RNN	General	ACOLITE	148	92	73	138	171	768
RNN	Local knowledge	L2C	50	53	121	123	116	584
RNN	Local knowledge	ACOLITE	81	138	94	162	156	1043

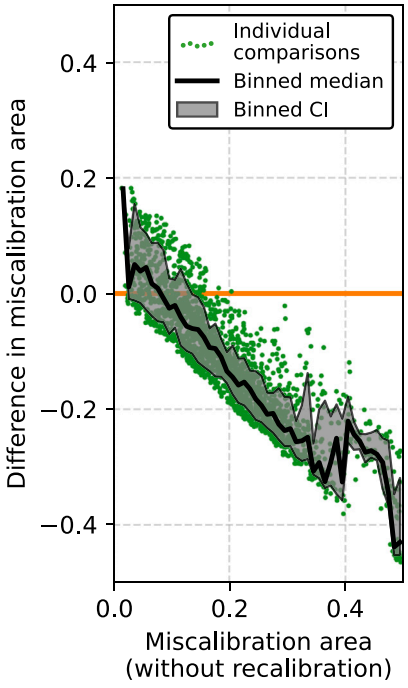

Fig. 10. Difference in miscalibration area due to recalibration, as a function of miscalibration area. Negative values on the vertical axis indicate improvement.

extremely large uncertainties (> 3000%) in the general case for ACOLITE but not in the *in situ* vs. *in situ* or local knowledge scenarios. Both of these patterns resemble the random split, WD, and OOD scenarios (Section 4.1.3), with increasing uncertainty moving from interpolation to extrapolation. A third similarity is the dominance of aleatoric uncertainty, representing ≥ 83% of the total uncertainty with only two exceptions from the MDN.

The coverage (Fig. 13) again indicated a large degree of underconfidence. Aggregated over both a_{ph} wavelengths and all five PNN architectures, the median coverage in each of the five scenarios was 96%–98%. The RNN was closest to, but still far from, $k = 1$ at 90% (CI 83%–94%); some outlier RNN instances were well-calibrated or even overconfident.

4.2.3. Uncertainty recalibration

Recalibration substantially reduced the estimated uncertainty for most models in all scenarios (Fig. 12). Correspondingly, the calibration curves for recalibrated models were much closer to and more symmetrically distributed around the diagonal (Fig. 14).

The coverage was drastically reduced in most cases (Fig. 13). For the *in situ* vs. *in situ* (58%; CI 46%–68%), general L2C (50%, CI 38%–68%) and ACOLITE (54%, CI 38%–68%) scenarios, this reduction resulted in slight overconfidence. The recalibrated models in the local knowledge scenarios with L2C (62%; CI 52%–74%) and ACOLITE (70%; CI 54%–80%) were on average close to $k = 1$. In all cases, the coverage remained widely dispersed between the 25 instances of each PNN architecture.

The percentage of beneficial recalibrations was 80%–92% across the five scenarios (Appendix G). This percentage was highest for the MDN (97%), followed by the ENS-NN and BNN-MCD (both 92%), BNN-DC (90%), and distantly the RNN (61%), mirroring the coverage factor trend. The range of changes in MA was similar across the five scenarios, with an overall median of −0.178 (CI −0.306 to −0.017). Lastly, the threshold in miscalibration area at which recalibration was beneficial for the majority of model instances was MA ≥ 0.20 and ≥ 0.15 in the general and local knowledge scenarios (Fig. 15), respectively, higher than the *in situ* threshold of 0.13. The difference is likely caused by the increase in observations available to train the PNNs and to fit the recalibration functions.

4.2.4. Spatial variability

We assessed spatial variability in model accuracy, uncertainty, and recalibration efficacy using two PRISMA scenes of the Venetian Lagoon and surrounding waters. These scenes were selected both for their optically complex water conditions and for the availability of multiple *in situ* measurement locations, with seven sampling points for a scene in May 2023 and thirteen for September 2023. The September dataset's larger sample size enabled us to evaluate not only PNN spatial accuracy but also model calibration through calculation of coverage (Eq. (7)) with and without recalibration of model uncertainties.

The Venetian Lagoon is characterized by shallow depths and dynamic sediment loads (Braga et al., 2020, 2022). Although benthic

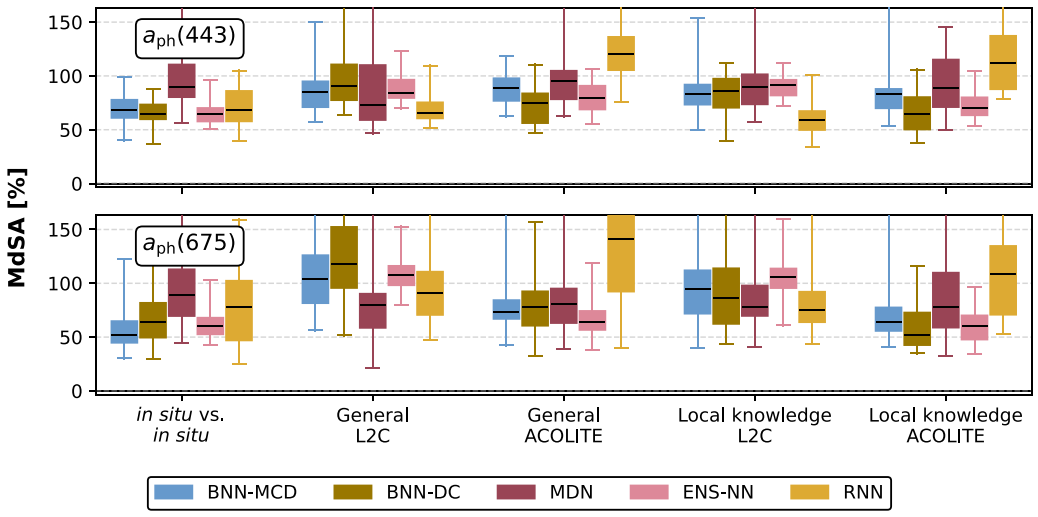

Fig. 11. Median symmetric accuracy (MdSA) of the $a_{ph}(443$, top) and $a_{ph}(675$, bottom) estimates from the 25 instances of each PNN, by PRISMA scenario.

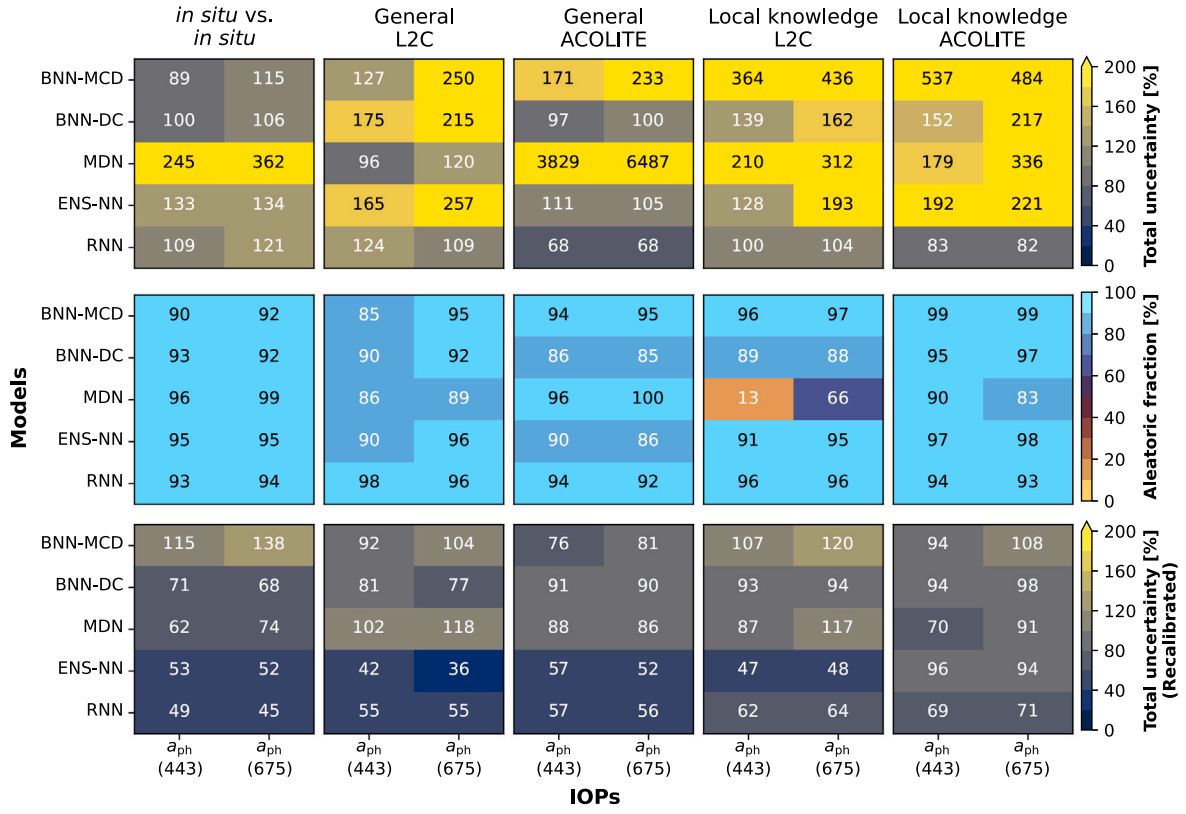

Fig. 12. Median uncertainty in PNN estimates from the average-performing models, for the PRISMA scenarios. Top row: total uncertainty as the sum of epistemic and aleatoric uncertainty. Middle row: aleatoric fraction of the total uncertainty. Bottom row: total uncertainty for recalibrated models.

albedo effects on both *in situ* and atmospheric correction-derived R_{rs} cannot be completely excluded, we explicitly assessed their influence on PNN-derived IOP retrievals. Specifically, in the September scene, four locations were identified as optically shallow as the Secchi depth remained visible down to the bottom, and we quantified how these areas affected model accuracy and uncertainty estimates.

The average-performing general-case models ENS-NN and BNN-MCD were applied under notably turbid conditions, as indicated by $R_{rs}(446)$ values above 0.020 sr^{-1} in many pixels of the May 2023 scene (Fig. 16a). For this first scene (Fig. 16), the atmospherically corrected and *in situ* match-up R_{rs} measurements agreed well, with MdSA typically 8%–22% by wavelength, 5% at 446 nm (Fig. 16a), and 17% overall. The ENS-NN and BNN-MCD estimates of $a_{ph}(443)$ agreed to 31% and 49% MdSA, respectively, being more accurate for this subset of match-ups than the overall match-up accuracy reported in Section 4.2.1. Estimates of $a_{ph}(675)$ and other IOPs were much less accurate, with MdSA $\gg 100\%$ particularly for a_{NAP}, matching Appendix G. In all cases, however, there was clear visual agreement between spatial patterns in R_{rs} and IOPs.

Similar to the first scene, the second scene from September 2023 exhibited strong agreement between atmospherically corrected and *in situ* R_{rs}, with an MdSA of 7% at 674 nm (Fig. 17a) and 12% overall. Across all match-ups in this scene, $a_{ph}(675)$ was retrieved with an MdSA of 37% by the MDN model and 44% by the RNN model (Fig.

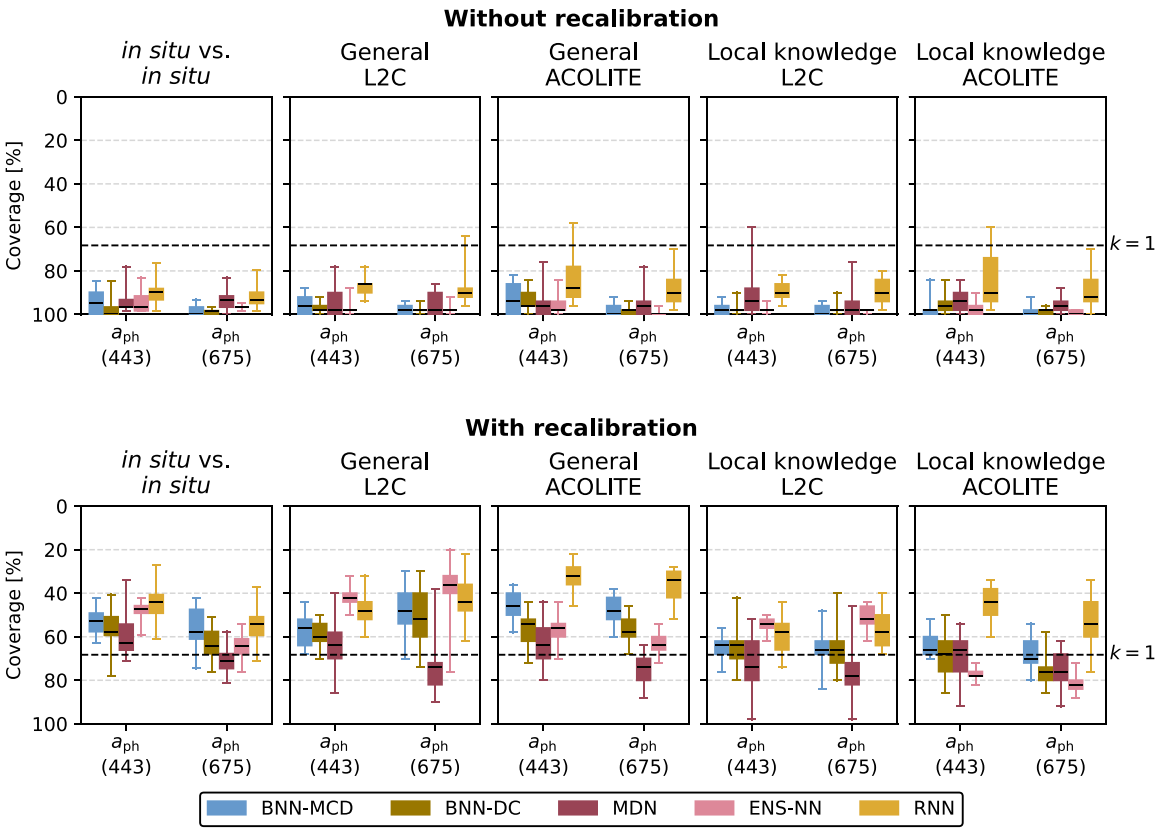

Fig. 13. Coverage of the uncertainty estimates for the PRISMA scenarios, analogous to Fig. 7.

17). To quantify the influence of optically shallow water on retrieval accuracy, we used median absolute error (MAE; Seegers et al. (2018)), a robust linear metric suitable for small sample sizes, where outliers could disproportionately impact logarithmic metrics such as MdSA.

MAE values for $a_{ph}(675)$ were lower in optically deep waters (MDN: $MAE_{deep} = 0.008\ m^{-1}$; RNN: $MAE_{deep} = 0.015\ m^{-1}$) compared to shallow waters (MDN: $MAE_{shallow} = 0.024\ m^{-1}$; RNN: $MAE_{shallow} = 0.107\ m^{-1}$). Both models consistently overestimated $a_{ph}(675)$ at the four optically shallow stations. Excluding these shallow-water observations significantly improved MdSA from 37% to 29% (MDN) and from 44% to 26% (RNN). In contrast, MdSA values for $a_{CDOM}(675)$ and $a_{NAP}(675)$ exceeded 200%. For $a_{CDOM}(675)$, both models showed large MAE values, approximately $MAE_{deep} = 0.028\ m^{-1}$ and $MAE_{shallow} = 0.048\ m^{-1}$, consistently underestimating the *in situ* measurements. For $a_{NAP}(675)$, MAE values remained around $0.010\ m^{-1}$ for both models, with negligible differences between optically shallow and deep waters. These observations for $a_{CDOM}(675)$ and $a_{NAP}(675)$ align well with broader accuracy patterns detailed in Section 4.2.1 and Appendix G.

In this Venetian Lagoon scene, benthic reflectance impacts the IOP retrieval, with $a_{ph}(675)$ showing the highest sensitivity. Notably, after excluding optically shallow water stations, model performance approached interpolation accuracy levels documented for random splits and within-distribution *in situ* scenarios (Section 4.1.1).

Uncertainty estimates were generally high (≥100%; Figs. 16, 17), comparable to Section 4.2.2. We used the thirteen match-ups to calculate coverage, which revealed that the MDN was highly underconfident for a_{ph} (92%) and a_{NAP} (100%), but overconfident for a_{CDOM} (38%) compared to the 68% expected from a well-calibrated model. The RNN performed similarly for a_{ph} (85%) and a_{NAP} (100%), but was relatively well-calibrated for a_{CDOM} (62%). Spatial patterns in uncertainty corresponded to patterns in R_{rs} and IOP, as well as to physical features, and in some cases (e.g., Figs. 16e, 17m) appeared to show banding parallel to the sensor geometry. Since the models estimated each pixel independently, these patterns propagating into the uncertainty estimates showcases the ability of PNNs to recognize, with limitations, their own domain knowledge.

Recalibration significantly reduced estimated uncertainties while preserving spatial patterns (Fig. 18). The recalibrated MDN had 69% coverage for $a_{ph}(675)$ and 62% for $a_{NAP}(675)$, both well-calibrated. However, at 23%, it was even more overconfident in its $a_{CDOM}(675)$ estimates than the non-recalibrated MDN. Results for the recalibrated RNN were more varied, at 54% (a_{ph}), 46% (a_{CDOM}), and 85% (a_{NAP}). The retrieval accuracy for $a_{ph}(675)$ was 44% for the recalibrated MDN and 54% for the RNN, and this difference relative to the non-recalibrated models was smaller than the typical variations between model instances (Fig. 11). Consistent with Section 4.2.3, we conclude that recalibration can substantially improve uncertainty estimation, with some limitations, and does not compromise retrieval accuracy.

5. Discussion

5.1. Generalization ability

Random split estimates for a_{ph}, a_{CDOM}, and a_{NAP} at 443 nm, with MdSA values of 25%–34%, aligned well with previous studies (O'Shea et al., 2023; Pahlevan et al., 2022; Saranathan et al., 2024). Similarly, our PRISMA scenario results matched the decline in accuracy observed by O'Shea et al. (2023). Said study reported an increase in MdSA of 37–63%pt., partially overlapping with the 15–44%pt. increase in $a_{ph}(443, 675)$ MdSA we found in the general AC PRISMA scenario for the two AC processors. Retrieval errors in a_{CDOM} and a_{NAP} were similarly elevated.

Retrieval of a_{CDOM} and a_{NAP} is inherently more sensitive to input perturbations than a_{ph}, as their absorption spectra are dominated by features at 443 nm with an exponential decay towards the red. As established in the literature, uncertainties from AC disproportionately affect the blue region (Braga et al., 2022; Gilerson et al., 2022; Warren

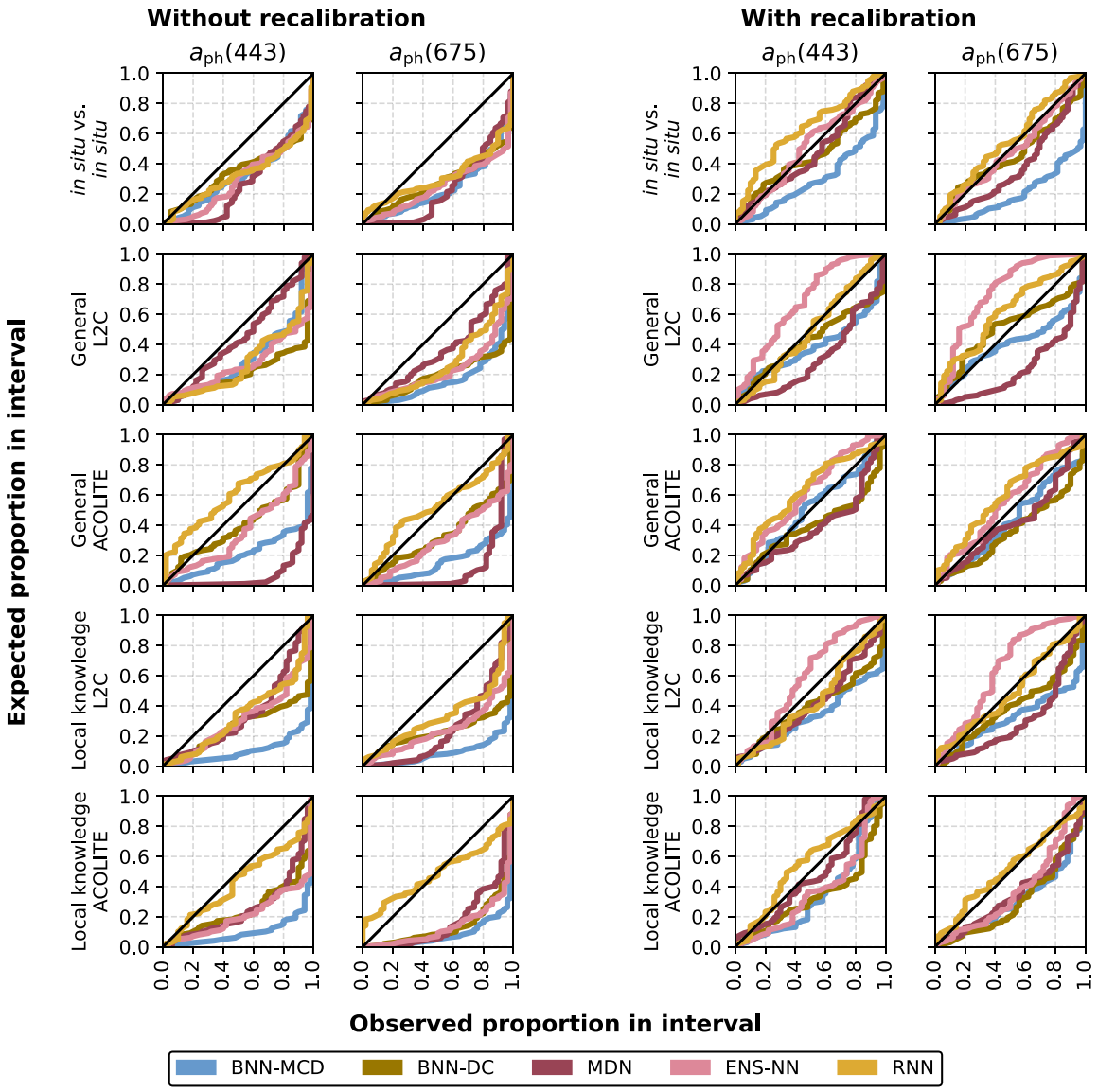

Fig. 14. Calibration curves for the PRISMA scenarios without (left) and with (right) recalibration. The average-performing model for each combination of scenario and architecture is displayed. As in Fig. 8, underconfident models fall below the diagonal line and overconfident ones above it.

et al., 2019), thereby amplifying IOP retrieval errors. Furthermore, because the PNNs were trained exclusively on *in situ* measurements, which represent a distinct domain compared to satellite-derived R_{rs}, the models were faced with a domain shift when applied to PRISMA observations, forcing them to generalize to conditions (e.g., spectral noise, sensor artifacts, AC-induced uncertainties) not encountered during training. Notably, adding local knowledge to the training set did not substantially reduce the impact of this domain shift on retrieval accuracy. After recalibration, however, the models were well-calibrated (close to $k = 1$), suggesting that regional measurements can effectively produce well-calibrated models for local applications. For a comprehensive analysis of all PRISMA scenario results, including those for a_{CDOM} and a_{NAP}, see Appendix G.

PNNs trained on *in situ* datasets showed a clear deterioration in estimation accuracy from the random split to the WD and OOD scenarios. The median MdSA per architecture in $a_{ph}(443)$, $a_{ph}(675)$, and $a_{CDOM}(443)$ increased by up to 20%pt. and 40%pt. in the WD and OOD scenarios, respectively. This decline highlights the challenges PNNs face when encountering unknown conditions, as models underestimated IOPs due to extrapolation in OOD scenarios. These findings were corroborated by the PRISMA *in situ* vs. *in situ* comparison (Section 4.2.1), where the PRISMA *in situ* measurements were partially WD and partially OOD relative to the primary *in situ* dataset, resulting in similar accuracy metrics.

Our findings underscore the critical role of independent dataset splitting for assessing model generalization effectively. Evaluation methods must account for both the variability across independent water bodies and the underlying water constituent or IOP distribution scenarios, including spatial autocorrelations, which are often overlooked in conventional approaches (Stock and Subramaniam, 2022). Our WD and OOD dataset splitting approaches (Section 3.1, Appendix B) satisfy these criteria. The PNNs exhibited high consistency in random split conditions, with minimal variation between model instances. In contrast, WD and OOD scenarios caused pronounced fluctuations in performance across PNN architectures and IOPs. Since real-world applications are generally WD or OOD, these results demonstrate major limitations in the capability of random splits and LOO cross-validation to assess model performance.

The MDN sensitivity analysis (Section 4.1.1, Appendix F) reveals a further advantage of the 50/50 WD and OOD split approach, showing that the MDNs were knowledge-limited – an observation that would be undetectable in LOO splits, where the training set vastly outnumbers

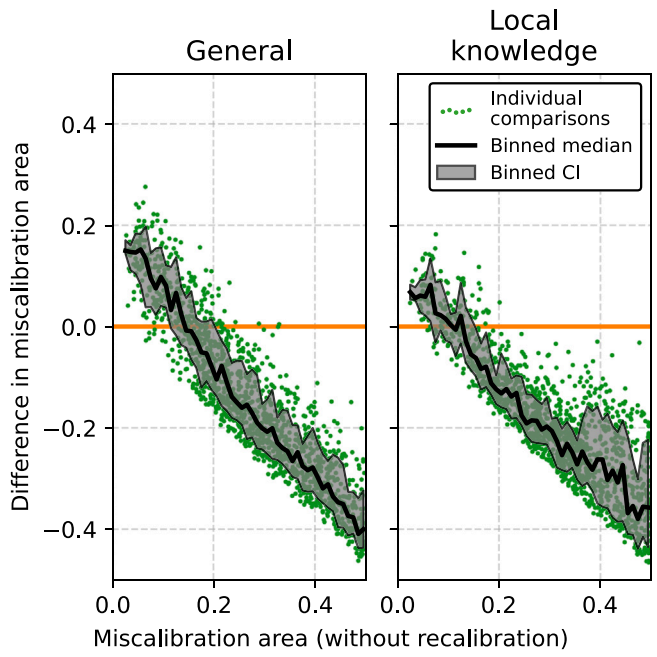

Fig. 15. Difference in miscalibration area due to recalibration, as a function of miscalibration area for the PRISMA AC general (left) and local knowledge (right) scenarios. Negative values on the vertical axis indicate improvement.

the test set. Random splits suffer from knowledge leakage (Stock et al., 2023), making the amount of training observations a poor indicator of epistemic uncertainty (lack of knowledge). Therefore, we recommend using random splits and LOO evaluations in their current form only for representing optimal-scenario performance, initial experimentation with model architectures and hyperparameter tuning.

The scarcity of quality-controlled match-up datasets linking satellite R_{rs} with reference IOPs often restricts model evaluations to WD scenarios, meaning models are only evaluated for their interpolation capability. This limitation helps explain the erratic behavior of (P)NNs applied to unknown conditions requiring extrapolation (Mouw et al., 2013; Neil et al., 2019; Saranathan et al., 2024; Werther et al., 2022). These issues can be mitigated if studies employing (P)NNs for IOP and water constituent retrieval more explicitly delineate the application limits of their models through domain checks (D'Alimonte et al., 2003) or uncertainty quantification (Sections Section 5.2, 5.3).

The consistency between BNN-MCD, BNN-DC, ENS-NN, and RNN suggests that selecting a different NN architecture, given similar access to training observations, is unlikely to substantially improve retrieval accuracy. Variability among instances of the same PNN architecture under identical scenarios is common in NNs and arises from factors such as random initialization, the sequence of training observations, model convergence issues, and the varying effectiveness of regularization techniques (Smith et al., 2021). An ensemble approach can help mitigate some of these inconsistencies (Pahlevan et al., 2022; Werther et al., 2021).

The MDNs initially showed significant retrieval errors in the *in situ* scenarios, which were explained through sensitivity analysis (Section 4.1.2) and application to PRISMA using a larger training dataset. It is unclear to what extent the difference was caused by the specific choice of IOPs left out versus the difference in dimensionality in general. Although estimating fewer IOPs brought the MDN closer to the other PNNs, it remained an outlier in the WD and OOD scenarios.

Assessing the relative importance of each wavelength band for IOP retrieval in the PNN architectures could advance the understanding of how PNNs leverage hyperspectral information from sensors like PRISMA. Such an analysis may highlight bands that mainly contribute noise or are strongly impacted by prior AC, potentially justifying their removal. Additionally, mapping the most influential bands to known absorption and scattering features could improve the physical interpretability of the results. However, multiple practical barriers exist. First, there is no universally accepted framework for determining feature importance in NNs, especially for PNNs. Second, because multiple methods exist, such as SHapley Additive exPlanations (SHAP; Lundberg and Lee (2017)), Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al. (2016)), permutation importance, and gradient-based approaches, their outcomes must be compared to obtain reliable insights. Third, feature importance methods often disagree in practice (Krishna et al., 2024), implying that a comprehensive, multi-method analysis across our 25 PNN instances would require novel methodological development to yield meaningful outcomes.

5.2. Extrapolation indication

Recognizing when NNs are forced to extrapolate beyond their training domain is critical for ensuring their reliability in remote sensing applications. While recent efforts have emphasized the importance of determining retrieval uncertainties for assessing the applicability of an approach (IOCCG, 2019; Saranathan et al., 2023; Werther and Burggraaff, 2023), there remains a gap in methodologies specifically designed to detect and handle OOD scenarios, especially in the context of PNNs and IOP estimations. Our results showed that uncertainty estimates may lack proper calibration (Sections Section 4.1.4, 4.2.3, 4.2.4), potentially leading to misplaced confidence – or lack thereof – when models are forced to extrapolate.

The ability to recognize extrapolation has been a subject of interest since the earliest research on NNs in aquatic remote sensing (Doerffer and Schiller, 1998; Schiller and Doerffer, 1999). These early studies employed a combination of forward and inverse NNs to determine whether encountered conditions fell within or outside the scope of the training distribution (Schiller and Krasnopolsky, 2001). This dual network approach not only aimed to make it possible to restrict NN application to interpolation conditions, but also provided a quality metric δ. This metric was defined as $\delta = \|y - f_{NN}(f_{NN}^{-1}(y))\|$, with y the observed values and f_{NN} and f_{NN}^{-1} NNs emulating the forward and inverse models, respectively. The value of δ quantifies the consistency between observed and reconstructed values, serving as an indicator of the inversion quality. This methodology was integrated into the ENVISAT Medium Resolution Imaging Spectrometer (MERIS) Case 2 processor (Doerffer and Schiller, 2007) and evolved into the C2RCC processor (Brockmann et al., 2016), now part of EUMETSAT's Sentinel-3 Ocean and Land Color Instrument (OLCI) operational processing chain. However, many complex NN processors do not implement domain checks. Without these checks, users may not know whether the underlying models are operating under interpolation or extrapolation, which, as our results demonstrate, can severely impact performance.

The implications of our findings extend to water constituent retrieval approaches that use Optical Water Type (OWT) classification schemes. OWTs represent clusters of similar water bodies based on their optical properties (Moore et al., 2001, 2014) and OWT classification is commonly used for estimating variables like chlorophyll-*a* (Liu et al., 2021) and total suspended matter concentration (Jiang et al., 2023). Since OWT classifications are designed from limited datasets (Bi and Hieronymi, 2024; Spyrakos et al., 2018), the underlying retrieval models used within each OWT are likely to encounter extrapolation conditions when applied to satellite imagery. Although recent schemes have begun to address the extrapolation scenario for chlorophyll-*a* (Liu et al., 2021), their performance stills need to be systematically evaluated. Furthermore, classifying R_{rs} spectra into OWTs does not inherently resolve the optical ambiguities that affect all retrieval algorithms (Defoin-Platel and Chami, 2007), including PNNs. However, OWTs can assist

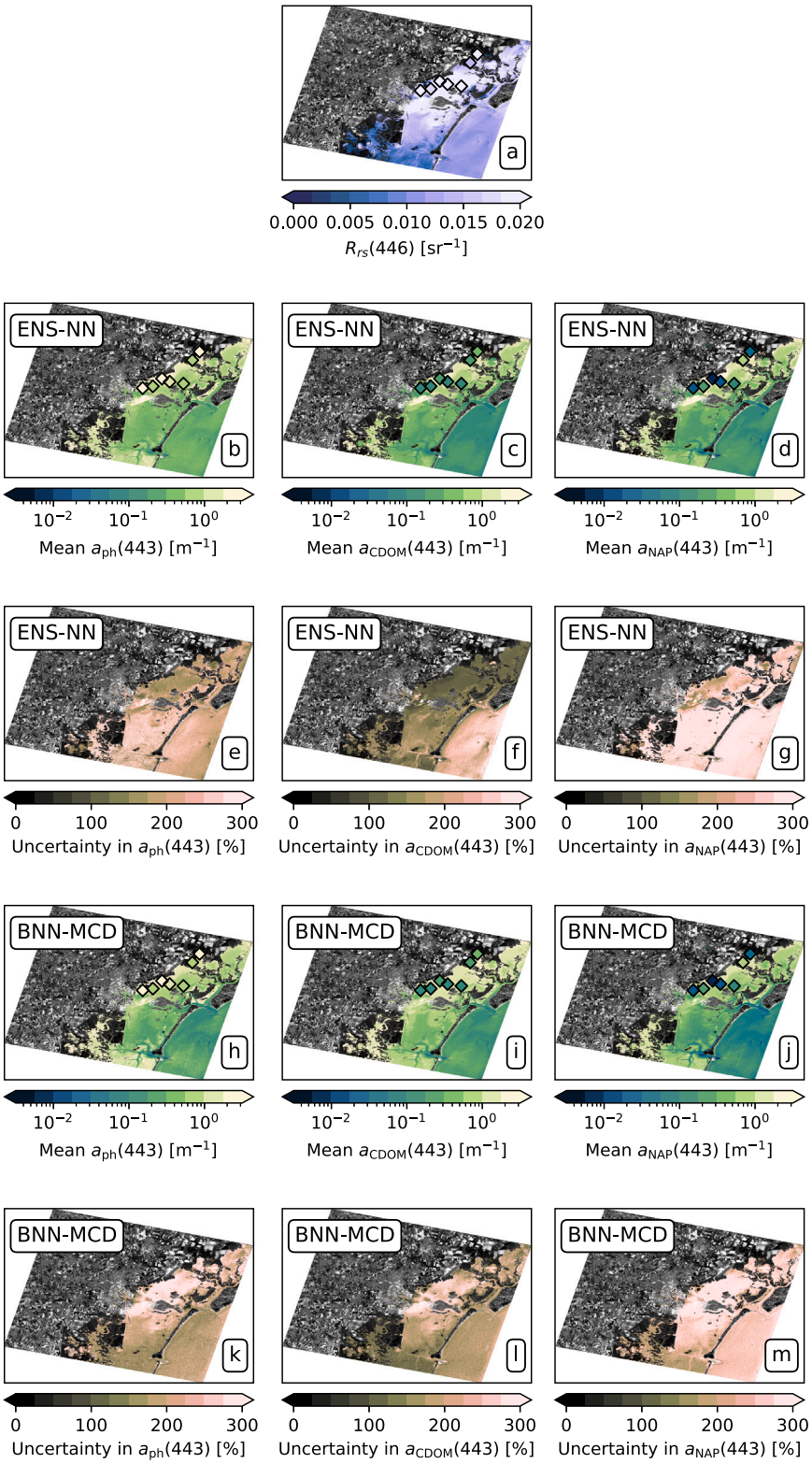

Fig. 16. PNN application to a PRISMA scene of the Venetian Lagoon on 2023-05-24, atmospherically corrected using the standard L2C processor. Panel (a) shows the input R_{rs}; panels (b–d) and (h–j) display the estimated IOPs; panels (e–g) and (k–m) present the corresponding estimated uncertainties. Diamonds indicate locations of *in situ* match-up observations. Land pixels were masked using an ad-hoc normalized difference water index (NDWI) based on $R_{rs}(559)$ and $R_{rs}(860)$ with a threshold of NDWI ≥ 0 for water, and are shown in grayscale based on an approximate luminance.

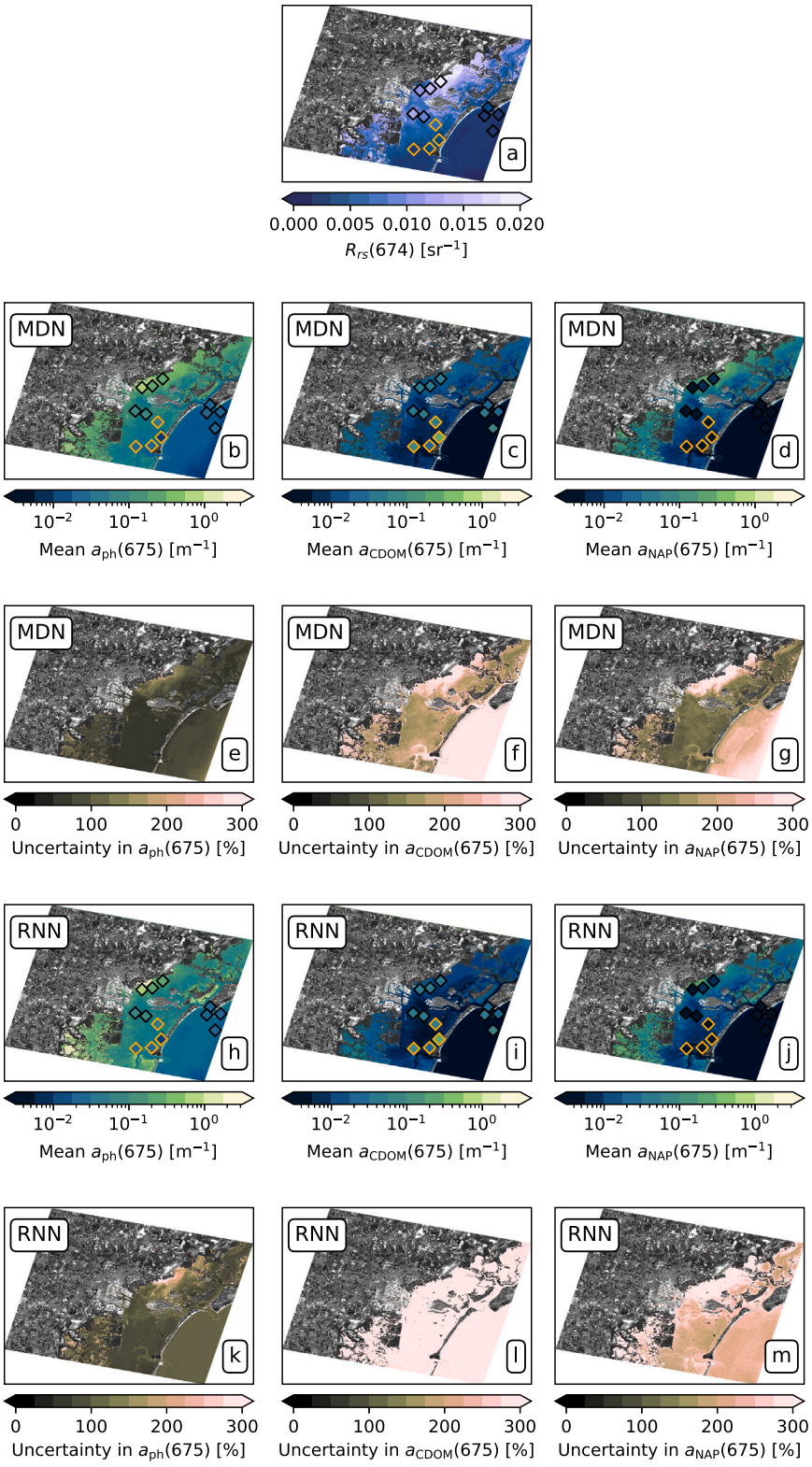

Fig. 17. PNN application to a PRISMA scene of the Venetian Lagoon on 2023-09-11, atmospherically corrected using ACOLITE. Similar to Fig. 16, with optically shallow match-up locations indicated by orange-outlined diamonds.

in selecting optimal AC methods (Pahlevan et al., 2021), potentially mitigating associated uncertainties.

Therefore, our study underscores the critical need for more robust methodologies capable of simultaneously indicating extrapolation conditions and addressing optical ambiguities between R_{rs} and IOPs, which vary in degree and can occur concurrently.

5.3. Uncertainty estimation

The predominance (> 80%) of aleatoric uncertainty – inherent to the observations – suggests that we may have reached the predictability limit for IOPs, imposed by fundamental characteristics (variability and error) of the measurements and the longstanding issue that no unique

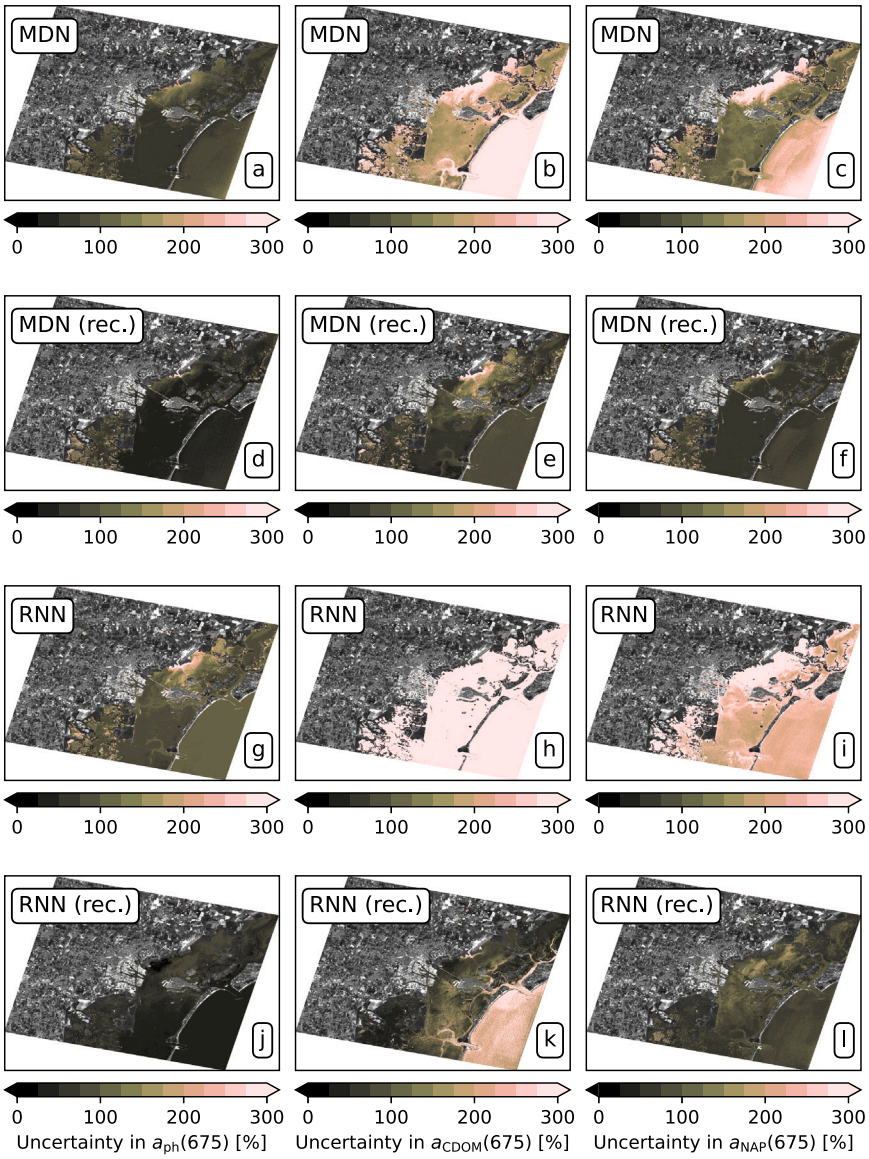

Fig. 18. Uncertainty estimates for the scene displayed in Fig. 17 for the average-performing (a–c, g–i) and average-performing recalibrated (d–f, j–l) PNN instances. Each column corresponds to an IOP, as shown at the bottom.

relationship exists between IOPs and R_{rs} (Zaneveld, 1994; Defoin-Platel and Chami, 2007). While Defoin-Platel and Chami (2007) advocated for Bayesian approaches and MDNs to address ambiguity, our findings suggest that data-driven models alone do not overcome this inherent limitation. This underscores the need to incorporate physical constraints or prior knowledge into retrieval methods to enhance their accuracy.

Epistemic uncertainty – arising from a lack of knowledge – can be reduced by expanding the training set with more diverse measurements, particularly in under-sampled or OOD scenarios. Notably, MDN uncertainty estimates improved in the PRISMA application owing to the larger training set (Section 4.2.2), which also improved model accuracy. However, there are practical constraints, as acquiring new *in situ* samples, particularly in remote or inaccessible regions, remains logistically and financially challenging (Jha and Chowdary, 2007; Meyer et al., 2024; Palmer et al., 2015). This persistent limitation implies that OOD scenarios will continue to present challenges for model generalization. Therefore, it is essential to adopt uncertainty quantification methods that adequately account for epistemic uncertainty in the absence of comprehensive training sets. Techniques like active learning, which allow models to identify and prioritize areas of high uncertainty for further collection of training observations, offer promising avenues for future research.

Recalibration improved the quality of uncertainty estimates, thereby increasing model trustworthiness (Section 4.1.4, 4.2.3, 4.2.4). Importantly, the effectiveness of recalibration depended on the degree of initial miscalibration, proving beneficial when the miscalibration area exceeded certain thresholds, namely ≥ 0.13 for *in situ* data and ≥ 0.15 or ≥ 0.20 for PRISMA. These thresholds offer practical guidelines for determining when recalibration is useful.

While adding more measurements will not linearly improve IOP retrieval accuracy, it can increase model trustworthiness through recalibration. For medium-to-large datasets, re-allocating part of the training set to recalibration can enhance model reliability without sacrificing accuracy (Table 1, Section 4.2.4). This observation highlights the value of explicitly considering sources of uncertainty when developing strategies for model improvement (Werther and Burggraaff, 2023).

6. Conclusions & future work

This study rigorously assessed PNN performance for hyperspectral estimation of absorption IOPs in optically complex waters. While PNNs

achieved favorable accuracy (MdSA as low as 25%) when *in situ* training and test sets shared similar IOP distributions, accuracy substantially degraded for independent water bodies and satellite applications. In the PRISMA general application scenario, MdSA exceeded 80% for $a_{ph}(443)$ through both L2C and ACOLITE ACs, while for $a_{ph}(675)$, it exceeded 100% through L2C and reached approximately 75% with ACOLITE (see Section 4.2.1). We therefore conclude that the worsened IOP retrieval inaccuracy from PRISMA imagery is systematic, reflecting inherent limitations associated with data-driven NN-based approaches rather than being attributable solely to a particular PNN architecture, training setup, or the choice of the AC method. These results demonstrate the limited generalization capacity of PNNs to estimate IOPs in novel conditions.

To systematically evaluate generalization ability, we introduced a novel dataset splitting strategy that distinguishes between interpolation (*in situ*, within-distribution) and extrapolation (out-of-distribution) settings. This approach addresses a critical gap in previous methodologies that conflated interpolation with true extrapolation, revealing that standard assessment approaches (such as random split) tend to overestimate PNN generalization.

Our systematic comparison of various PNN architectures further uncovered significant variations in model calibration. Models with miscalibration beyond a defined threshold (0.13) benefited consistently from post-hoc uncertainty recalibration, while well-calibrated models did not. This represents the first systematic demonstration in aquatic remote sensing of how miscalibration analysis and recalibration techniques can improve model reliability.

Crucially, our findings reveal that aleatoric uncertainty dominates in IOP retrieval, implying fundamental limitations in resolving the reflectance-IOP relationship solely through data-driven approaches. Simply expanding the training dataset with similar observations is unlikely to overcome this inherent uncertainty, and alternative machine learning methods (including decision trees or other neural network variants) are unlikely to offer substantial improvements over the evaluated PNNs.

Real advancement in applying machine learning in our field will have to come from novel methods that integrate physical principles governing the relationship between IOPs and reflectance into neural network architectures, thereby creating *physics-informed neural networks* (PINNs). PINNs have already been applied across different disciplines of aquatic research, such as lake temperature profiling (Jia et al., 2019), underwater imaging polarimetry (Hu et al., 2022), and reservoir pressure management (Donnelly et al., 2024). However, the use of PINNs to retrieve IOPs or other variables from remote sensing is thus far undocumented in literature. We recommend three strategies to be explored for developing PINNs for IOP estimation:

1. **Physics-constrained loss functions:** One approach involves incorporating physical constraints directly into the loss function of the neural network (Raissi et al., 2019). In the context of IOP estimation, this could mean penalizing the network for violating known physical relationships between IOPs and apparent optical properties like R_{rs}. Such an approach ensures that the network produces estimates grounded in fundamental optical principles, potentially reducing errors in OOD scenarios by improving its ability to generalize across diverse water conditions.
2. **Physics-inspired architecture design:** The neural network architecture itself can be designed to reflect the underlying physical processes of light propagation through water. For instance, different layers or sub-networks could be structured to represent various components of the radiative transfer equation, with their interactions guided by established physical principles (Chattopadhyay et al., 2022). This physics-inspired architecture could provide insights into the physical processes during IOP inference. Although not explicitly physics-informed, the forward-inverse framework developed by Schiller and Doerffer in the late 1990s, which forms the basis for C2RCC, was clearly physics-inspired and should be regarded as an early attempt to incorporate physical principles into model architecture.
3. **Hybrid physics-ML integration**: A highly ambitious approach involves the integration of established radiative transfer models for optical oceanography, such as HydroLight (Hedley and Mobley, 2021) or WASI (Gege, 2014), directly into the PINN framework. This could be achieved by creating differentiable versions of these models, either through neural network surrogates/emulators (Raissi et al., 2019) or automatic differentiation techniques (Baydin et al., 2018). The resulting hybrid architecture would allow for seamless integration of physics-based simulations within the neural network, enabling end-to-end training that leverages both data-driven learning and well-established physical principles. This approach could not only enhance generalization but also enable both forward and inverse modeling within a unified framework, increasing the versatility of PINNs across diverse IOP estimation scenarios.

By pursuing these strategies to integrate physical principles into PNNs, we can address current limitations and achieve more robust and accurate IOP estimation across optically complex waters, thereby enhancing the use of hyperspectral remote sensing for aquatic research.

CRediT authorship contribution statement

Mortimer Werther: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Olivier Burggraaff:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Daniela Gurlin:** Writing – review & editing, Data curation. **Arun M. Saranathan:** Writing – review & editing, Methodology, Investigation. **Sundarabalan V. Balasubramanian:** Writing – review & editing, Data curation. **Claudia Giardino:** Writing – review & editing, Data curation. **Federica Braga:** Writing – review & editing, Data curation. **Mariano Bresciani:** Writing – review & editing, Data curation. **Andrea Pellegrino:** Writing – review & editing, Data curation. **Monica Pinardi:** Writing – review & editing, Data curation. **Stefan G.H. Simis:** Writing – review & editing, Data curation. **Moritz K. Lehmann:** Writing – review & editing, Data curation. **Kersti Kangro:** Writing – review & editing, Data curation. **Krista Alikas:** Writing – review & editing, Data curation. **Dariusz Ficek:** Writing – review & editing, Data curation. **Daniel Odermatt:** Writing – review & editing, Resources, Project administration, Investigation.

Code availability

Following FAIR research principles, all code used to generate the scenarios, train and apply the PNNs, produce the manuscript figures, and auxiliary functions is available under: https://github.com/mowerther/iop_nns.

All manuscript datasets, trained PNNs, and PNN IOP estimates are available on Zenodo under: https://doi.org/10.5281/zenodo.14893798.

The code makes use of the following packages: CMCrameri (Crameri et al., 2020; Rollo, 2024), Dill (McKerns et al., 2011), Matplotlib (Hunter, 2007), NumPy (Harris et al., 2020), Pandas (McKinney, 2010), Scikit-learn (Pedregosa et al., 2011), SciPy (Virtanen et al., 2020), TensorFlow (Abadi et al., 2016), Uncertainty Toolbox (Chung et al., 2021), xarray (Hoyer and Hamman, 2017).

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.

Acknowledgments

We are grateful to all providers for sharing or granting open access to their measurements, including Janet Anstee, Zhigang Cao, Kendall L. Carder, Kimberly A. Casey, Glenn Cota, Arnold G. Dekker, Nathan Drayson, Dalin Jiang, Anatoly A. Gitelson, Rafael Gonçalves-Araujo, Steven R. Greb, Chuanmin Hu, Gary Kirkpatrick, Lin Li, Ronghua Ma, Tim J. Malthus, Bunkei Matsushita, Deepak R. Mishra, Sachidananda Mishra, Greg Mitchell, Wesley J. Moses, Frank Muller-Karger, David M. O'Donnell, Antonio Ruiz Verdú, and Richard Zimmerman. We thank Salvatore Mangano for his field work support activities in Italian lakes. Finally, we would like to thank the four anonymous reviewers for their thoughtful and supportive reviews.

This work was supported by the Swiss National Science Foundation (SNSF) under the Lake3P project, grant no. 204783. Partial funding for this study was also provided through the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.4 – Call for Tender No. 3138 of December 16, 2021, as rectified by Decree No. 3175 of December 18, 2021, issued by the Italian Ministry of University and Research. This initiative is funded by the European Union – NextGenerationEU (project code CN_00000033, Concession Decree No. 1034 of June 17, 2022, adopted by the Italian Ministry of University and Research, CUP B83C22002930006, project title "National Biodiversity Future Center - NBFC"). Additional support was provided by the ASI-CNR project PANDA-WATER (PRISMA Products and Applications for Inland and Coastal Water, 2022-15-U.0) and the Space It Up project, funded by ASI and MUR, Contract No. 2024-5-E.0, CUP I53D24000060005.

Appendix A. Datasets

The GLORIA dataset was expanded with additional IOPs for existing samples (n = 1 605) and new samples (n = 461) that included hyperspectral R_{rs}, IOPs, and water constituent concentrations (in total: n = 2 066) (Table A.3). These additions came from various contributors and repositories:

- SeaBASS repository (n = 50): Observations from Cota and Zimmerman (2000), Carder and Mitchell (1999), Muller-Karger (2015), Hu and Muller-Karger (2012), Carder (1998) extending sampling locations in Bahamian and U.S. coastal waters.
- PANGAEA repository (n = 364): Observations from Casey et al. (2020) and Gonçalves-Araujo et al. (2018), introducing new sampling locations in:
  - U.S. coastal waters of the Gulf of Mexico (n = 356)
  - International oceanic waters of the Arctic Ocean (n = 8)
- Additional lake observations (n = 47): Complementing existing GLORIA datasets for Chinese and Japanese lakes (Lehmann et al., 2023) with observations for:
  - Taihu Lake, China (n = 39)
  - Lake Suwa, Japan (n = 8)

For more information on the dataset extension and additional IOPs, please refer to O'Shea et al. (2023), Pahlevan et al. (2022).

Appendix B. Dataset splitting algorithm

The developed algorithm partitions a dataset with splitting variables (e.g., IOPs, water constituents) into within-distribution (WD) and out-of-distribution (OOD) training and test sets. The algorithm ensures unique water bodies in each set while optimizing for different distribution characteristics.

The algorithm uses dual annealing optimization to find the best split by minimizing an objective function over multiple iterations. For WD datasets, the aim is to maximize similarity between train and test sets, while for OOD the algorithm maximizes their dissimilarity.

The WD objective function is defined as:

$$f_{WD}(x) = S_{WD}(D_{\text{train}}, D_{\text{test}}) + P(D_{\text{train}}, D_{\text{test}}). \tag{8}$$

The OOD objective function is:

$$f_{OOD}(x) = -S_{OOD}(D_{\text{train}}, D_{\text{test}}) + P(D_{\text{train}}, D_{\text{test}}). \tag{9}$$

Here, S_{WD} and S_{OOD} are similarity and dissimilarity scores, respectively, and P is a penalty term for dataset size imbalance to obtain equal training and test sets.

The WD similarity score is calculated as:

$$S_{WD}(D_1, D_2) = \sum_{i=1}^{n} |\mu_i(D_1) - \mu_i(D_2)|. \tag{10}$$

The OOD dissimilarity score is:

$$S_{OOD}(D_1, D_2) = \sum_{i=1}^{n} \sum_{p \in P} |Q_i^p(D_1) - Q_i^p(D_2)|. \tag{11}$$

In these equations, $\mu_i(D)$ is the mean of the ith IOP column at 443 nm in dataset D, $Q_i^p(D)$ is the pth percentile of the ith IOP column at 443 nm, n is the number of IOP columns, and P is a set of specified percentiles (10%–90%).

The balance penalty is defined as:

$$P(D_{\text{train}}, D_{\text{test}}) = |N_{\text{train}} - N_{\text{test}}|. \tag{12}$$

where N_{train} and N_{test} are the numbers of observations in the train and test sets, respectively.

Please see the Code Availability section for the implementation in Python.

Appendix C. Scaling of input and target variables

C.1. Input variables: R_{rs}

We used Sklearn's RobustScaler to scale the R_{rs} input values. The RobustScaler subtracts the median and scales R_{rs} according to the interquartile range. The RobustScaler is independently trained on and applied to each input feature, i.e. each wavelength of R_{rs}.

C.2. Target variables: IOPs

Because the target variables (a_{ph}, a_{CDOM}, a_{NAP}) span several orders of magnitude (10^{-4} to 10^{1} m^{-1}), they are first log-transformed as in Eq. (13). Here $\mathbf{y}_i$ is the original vector of 6 target values for sample i, with $\mathbf{y}_i^{\ell}$ its log-transformed counterpart; the natural logarithm is applied element-wise. The log transformation ensures that the network treats observations at different orders of magnitude with equal weight, rather than optimizing for the highest values only:

$$\mathbf{y}_i^{\ell} = \log(\mathbf{y}_i). \tag{13}$$

All logarithms in this work are natural logarithms, i.e. $\log = \log_e$, unless otherwise specified. The log-transformed observations are then scaled using Sklearn's MinMax scaler. This scaler transforms the observations to a fixed range, in this case between −1 and 1. While not strictly necessary, this scaling is beneficial to the training process. First, it ensures numerical stability and accelerates the gradient descent process by presenting the network with a consistent range of values. Second, for simultaneous estimation of multiple target variables, it ensures that each variable is treated equally and prevents any single target from dominating the learning process. Scaling is especially critical in the context of PNNs, as it directly influences the estimation of variance, and thus uncertainty.

The MinMax scaling to the range $[-1, 1]$ is defined in Eq. (14). Here $\mathbf{y}_i^{\ell}$ is the vector of log-transformed target observations for sample i,

Table A.3
GLORIA subsets including additional $a_{ph}(443)$, $a_{ph}(675)$, $a_{CDOM}(443)$, $a_{CDOM}(675)$, $a_{NAP}(443)$ and $a_{NAP}(675)$ observations from individual contributors and SeaBASS used in this study.

Dataset ID	Water system(s)	Number of		Author(s)
		unique water system(s) (n = 151)	**observation(s)** (n = 1605)	
AlikasK_EE_UT-TO	Estonian and Finnish lakes	27	57	Alikas, K.; Kangro, K.; Ligi M.
AnsteeJ_AU_CSIRO	Australian lakes and water treatment plant ponds	13	104	Anstee, J.; Drayson, N.
DekkerAG_NL_VU	Dutch lakes and rivers	16	20	Dekker, A. G.; Malthus, T. J.
FicekD_PL_APSL	Polish lakes	13	97	Ficek, D.
GiardinoC_IT_CNR-IREA	Italian lakes	6	60	Giardino, C.; Bresciani, M.
GitelsonAA_US_UNL	U.S. lakes	12	178	Gitelson, A. A.; Gurlin, D.; Moses, W. J.
GrebSR_US_WDNR	U.S. lakes and rivers	34	195	Greb, S. R.; Gurlin, D.
LehmannMK_NZ_UOW_NZ_LK	New Zealand lakes	3	12	Lehmann, M. K.; Reed, L.
LiL_US_IUPUI	U.S. lakes	3	141	Li, L.
MatsushitaB_JP_Tsukuba	Japanese lakes	1	26	Matsushita, B.; Jiang, D.
MishraDR_US_MSU	U.S. aquaculture ponds	1	41	Mishra, D. R.; Mishra, S.
O'DonnellDM_US_UFI	Canadian and U.S. lakes	3	41	O'Donnell, D. M.
Ruiz-VerduA_ES_UVEG-CEDEX	Spanish lakes	4	16	Ruiz Verdú, A.
SeaBASS_US_ODU	U.S. coastal waters	1	45	Cota and Zimmerman (2000)
SeaBASS_US_USF	U.S. coastal waters	1	97	Hu (2010a)
SeaBASS_US_USF	U.S. coastal waters	1	38	Hooker et al. (2011)
SeaBASS_US_USF	U.S. coastal waters	1	201	Carder and Mitchell (1999)
SeaBASS_US_USF	U.S. lakes	1	10	Carder (1997)
SeaBASS_US_USF	U.S. coastal waters	1	11	Carder and Kirkpatrick (1998)
SeaBASS_US_USF	U.S. coastal waters	1	16	Hu (2010b)
SeaBASS_US_USF	U.S. coastal waters	1	75	Hu (2008)
SeaBASS_US_USF	Bahamian and U.S. coastal waters	2	13	Carder (1998)
SeaBASS_US_USF	U.S. coastal waters	1	67	Carder and Hu (2005)
SeaBASS_US_USF	U.S. coastal waters	1	1	Hu and Muller-Karger (2012)
SeaBASS_US_USF	U.S. coastal waters	1	12	Muller-Karger (2015)
SimisSGH_NL_NIOO-KNAW	Dutch lakes	2	31	Simis, S. G. H.

$\mathbf{y}_i^s$ its rescaled counterpart, and $\mathbf{y}_{\min}^{\ell}$, $\mathbf{y}_{\max}^{\ell}$ are the vectors with the minimum and maximum values across all samples, again evaluated independently for each of the 6 targets:

$$\mathbf{y}_i^s = 2\frac{\mathbf{y}_i^{\ell} - \mathbf{y}_{\min}^{\ell}}{\mathbf{y}_{\max}^{\ell} - \mathbf{y}_{\min}^{\ell}} - 1. \tag{14}$$

C.3. Inverse scaling of the output means and variances

As discussed above, the model estimates are in log-transformed and MinMax-scaled units. To obtain the estimates in real units, the estimates must pass through the inverse transformations.

C.4. MinMax scaling

The estimated means and variances describe a normal distribution in MinMax-scaled units, so that $\hat{y}_i^s \sim \mathcal{N}(\hat{\mu}_i^s, (\hat{\sigma}_i^s)^2)$. Since this scaling is a linear transformation, so is its inverse, meaning the resulting estimates in log-transformed units are also normally distributed. Recall Eq. (14):

$$\mathbf{y}_i^s = 2\frac{\mathbf{y}_i^{\ell} - \mathbf{y}_{\min}^{\ell}}{\mathbf{y}_{\max}^{\ell} - \mathbf{y}_{\min}^{\ell}} - 1.$$

The mean of the new distribution is obtained by inverting Eq. (14):

$$\hat{\mu}_i^{\ell} = \frac{1}{2}(\mathbf{y}_{\max}^{\ell} - \mathbf{y}_{\min}^{\ell})(\hat{\mu}_i^s + 1) + \mathbf{y}_{\min}^{\ell}. \tag{15}$$

The variance follows from uncertainty propagation:

$$(\hat{\sigma}_i^{\ell})^2 = \left(\frac{\partial y^{\ell}}{\partial y^s}\right)^2 (\hat{\sigma}_i^s)^2 \tag{16}$$

$$= \frac{1}{4}(\mathbf{y}_{\max}^{\ell} - \mathbf{y}_{\min}^{\ell})^2 (\hat{\sigma}_i^s)^2. \tag{17}$$

C.5. Log transform

Since the estimates are normally distributed in log-transformed units, i.e. $\hat{y}_i^{\ell} \sim \mathcal{N}(\hat{\mu}_i^{\ell}, (\hat{\sigma}_i^{\ell})^2)$, the corresponding estimates in real units follow a lognormal distribution: $\hat{y}_i \sim \text{Lognormal}(\hat{\mu}_i^{\ell}, (\hat{\sigma}_i^{\ell})^2)$. Since the lognormal distribution is naturally skewed, there are multiple options for the mean and variance. We use the geometric mean of the lognormal distribution, which equals the median, to transform $\hat{\mu}_i^{\ell}$ back to real units, as in Eq. (18). This is the direct inverse of the original log transformation:

$$\hat{\mu}_i = \exp(\hat{\mu}_i^{\ell}). \tag{18}$$

For the variance $\hat{\sigma}_i^2$, we use the arithmetic variance, as in Eq. (19). This is not strictly the most accurate choice, since it represents the typical variation around the arithmetic mean, rather than around the geometric mean. The geometric variance would more fairly represent the variation around the geometric mean as well as the asymmetry of the distribution. However, asymmetric uncertainties are notoriously difficult to propagate, and the arithmetic variance is more similar to typical uncertainty metrics reported in the literature. In practice, this choice will lead to a slight overestimation in the uncertainty in real units, but Monte Carlo simulations for various IOP samples used in this study showed that the difference is very minor, on the order of a few percent points:

$$\hat{\sigma}_i^2 = \exp(2\hat{\mu}_i^{\ell} + (\hat{\sigma}_i^s)^2) \times (\exp((\hat{\sigma}_i^s)^2) - 1). \tag{19}$$

Appendix D. PNN hyperparameters

The architectural parameters of the models, including the number of neurons and hidden layers, were informed by prior studies that used R_{rs} as the input variable for estimating aquatic variables (Pahlevan et al., 2020, 2022; O'Shea et al., 2023; Saranathan et al., 2024).

Table D.4
PNN hyperparameters and their values as used in this study.

Hyperparameter	Value	Comment
PNN instances per scenario	25	
Neurons input layer	61 (*in situ*), 36 (PRISMA)	
Hidden layers	5	
Neurons in hidden layers	100 - 100 - 100 - 100 - 100	
Neurons in output layer	12	6 means and 6 variances
Batch size	32	
Activation function	Rectified Linear Unit (ReLU)	
Drop rate	25%	BNN-MCD, BNN-DC, RNN
Monte Carlo samples	100	BNN-MCD, BNN-DC, RNN
Ensemble members	10	ENS-NN only
L2 regularization	10^{-3}	
Learning rate	10^{-4}	
Optimizer	Adam	
Loss function	Negative log-likelihood (NLL)	
Mixture components	5	MDN only

During training, each model was monitored for convergence using an internal validation set comprising 10% of the training observations in each scenario, which was randomly partitioned and reserved exclusively for early stopping. Training was terminated when the validation loss ceased to improve, thereby preventing overfitting to the training dataset. Since a validation set was part of a training set, it was distinct from the independent test set in each scenario used for final model evaluation.

To further mitigate overfitting, L2 regularization (weight decay) was applied to all PNNs (Krogh and Hertz, 1991). For BNN-DC, BNN-MCD and RNN, the Dropout or DropConnect layers provide additional regularization (Wager et al., 2013).

Appendix E. Negative log-likelihood loss function

The negative log-likelihood (NLL) loss function is used to estimate both the mean and variance of the target variables in the PNNs (except for MDN). The NLL for a target with an observed value y and an estimate $\hat{y}$ described by a Gaussian distribution $\hat{y} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ is given in Eq. (20). This derivation of this equation from the normal distribution is provided below. In this work, the NLL is applied to the scaled data y_i^s, meaning the model estimates are also in scaled units: $\hat{\mu}_i^s$ and $(\hat{\sigma}_i^s)^2$:

$$\mathrm{NLL} = \frac{1}{2}\left[\log(\hat{\sigma}^2) + \frac{(y-\hat{\mu})^2}{\hat{\sigma}^2} + \log(2\pi)\right]. \tag{20}$$

The specific choice for the NLL loss function serves several important purposes in the context of PNNs:

(1) **Quantifying uncertainty:** Unlike traditional loss functions that only estimate the mean, the NLL also estimates a variance, from which an uncertainty estimate can be derived.
(2) **Robustness to noise:** By explicitly modeling the variance, the NLL loss function can account for noise in the observations. Higher estimated variance in noisier regions reduces the penalization of estimation errors, which should lead to more robust models.
(3) **Improving estimation quality:** The NLL loss encourages the model to provide both accurate and confident estimates. It penalizes estimates that are both far from the true values and have low estimated variance, improving overall estimate quality.

In this work, the NNs are trained in batches of 32 samples (Table D.4), with 6 features each, so the average NLL across those $32 \times 6 = 192$ samples is used as the loss function, which the training process aims to minimize. The softplus function $\mathrm{softplus}(x) = \log(1 + \exp(x))$ ensures that the estimated variance is positive.

E.1. Derivation

The probability density function (PDF) of a Gaussian distribution with mean μ and variance σ^2 is:

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right). \tag{21}$$

The log-likelihood is the natural logarithm of this PDF:

$$\log\left[p(y \mid \mu, \sigma^2)\right] = \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)\right] \tag{22}$$

$$= \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] + \log\left[\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)\right] \tag{23}$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \tag{24}$$

$$= -\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{(y-\mu)^2}{\sigma^2}\right] \tag{25}$$

$$= -\frac{1}{2}\left[\log(\sigma^2) + \frac{(y-\mu)^2}{\sigma^2} + \log(2\pi)\right]. \tag{26}$$

The negative log-likelihood is therefore:

$$\mathrm{NLL} = -\log\left[p(y \mid \mu, \sigma^2)\right] = \frac{1}{2}\left[\log(\sigma^2) + \frac{(y-\mu)^2}{\sigma^2} + \log(2\pi)\right]. \tag{27}$$

The NLL has a scaling factor $\frac{1}{2}$ and a constant term $\log(2\pi)$, which follow from the normalization of the Gaussian PDF. Being constant, these terms do not affect the network optimization; we choose to keep them for consistency with the definition. The $(y-\mu)^2$ term is the classic mean square error (MSE) metric, which penalizes models that diverge too far from the 'true' observed values. Dividing the MSE by the variance σ^2 allows the model to account for uncertainty, lowering the weight of estimates that are far from y but also uncertain. By itself, this could lead to a runaway effect where the model simply increases σ^2 arbitrarily to lower the loss function; however, the additive $\log(\sigma^2)$ prevents this by penalizing large uncertainty estimates, and conversely rewarding models with high confidence.

Appendix F. Model sensitivity to $a_{cdom}(675)$ and $a_{nap}(675)$

A model sensitivity study was conducted to investigate the large differences between estimates from the MDN vs. the other architectures (Section 4.1.1), as well as to evaluate the influence of the relatively error-prone variables $a_{CDOM}(675)$ and $a_{NAP}(675)$ on the overall results. 25 new instances of each PNN architecture were trained on the *in situ* dataset, this time only estimating $a_{ph}(443)$, $a_{ph}(675)$, $a_{CDOM}(443)$ and $a_{NAP}(443)$.

The accuracy of the BNN-MCD, BNN-DC, ENS-NN, and RNN estimates differed little between the 6-IOP (Fig. 5) and 4-IOP (Fig. 19) studies. For example, the random split BNN-MCD $a_{ph}(443)$ MdSA

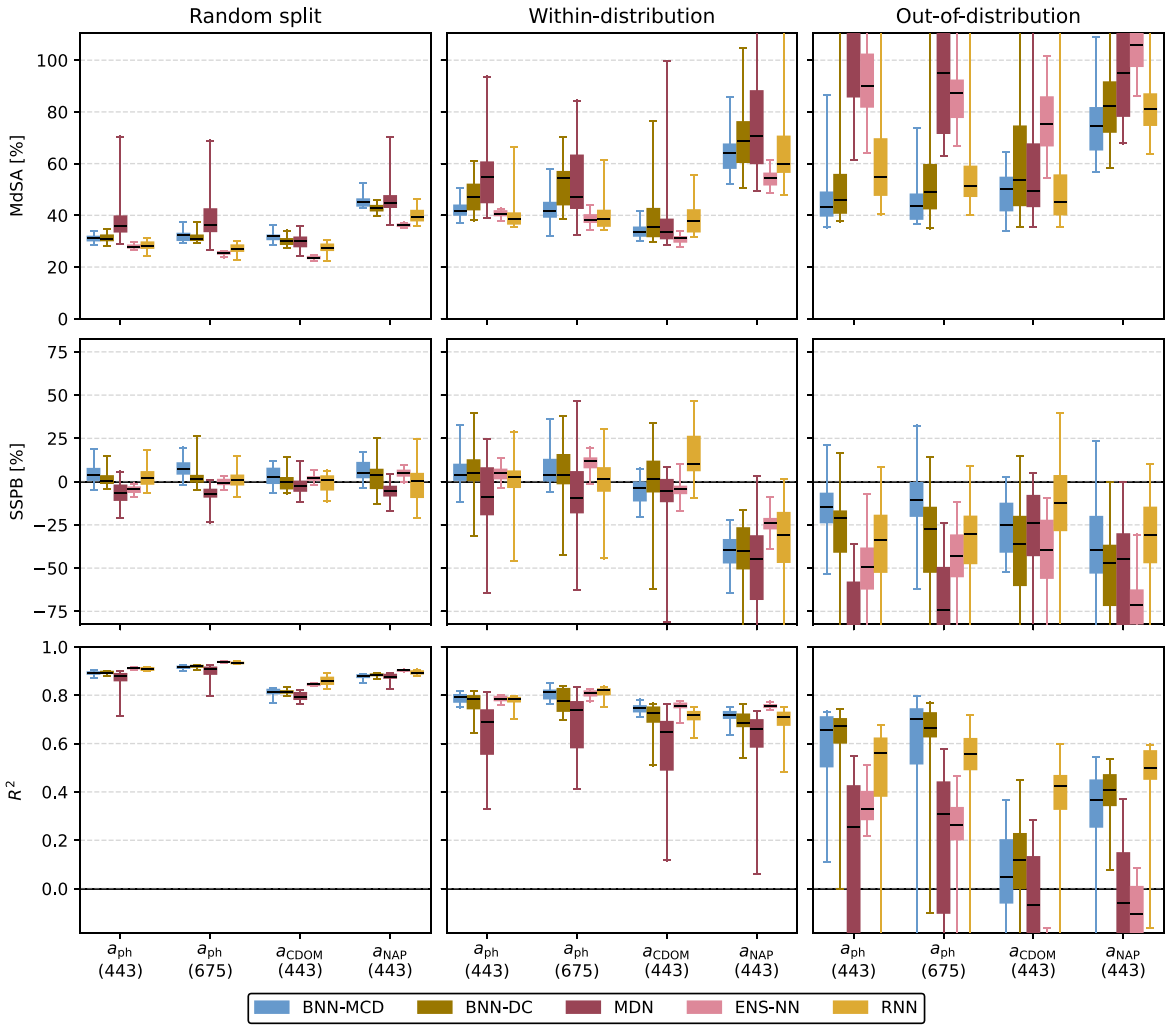

Fig. 19. Accuracy metrics for the PNNs trained without $a_{CDOM}(675)$ and $a_{NAP}(675)$, analogous to Fig. 5.

was 32% (CI 31%–33%) when trained on 6 target IOPs and 31% (CI 30%–32%) when trained on 4 target IOPs. The differences were similarly negligible for the other scenarios, networks, IOPs, and metrics.

The MDN, on the contrary, showed large differences. Notably, the MDN accuracy metrics were much more in line with those from the other architectures when trained on 4 IOPs (Fig. 19). As a numerical example, the random split MDN $a_{ph}(443)$ MdSA was 52% (CI 41%–75%) with 6 IOPs and 36% (CI 33%–42%) with 4 IOPs; the WD MDN $a_{ph}(443)$ MdSA was 89% (CI 65%–111%) with 6 IOPs and 55% (CI 45%–74%) with 4 IOPs. Some results changed more extremely, such as the WD MDN $a_{CDOM}(443)$ MdSA decreasing from 123% (CI 63%–196%) to 34% (CI 31%–42%), a significant difference. The 4-IOP MDN generally displayed much less variability between model instances than the 6-IOP one.

The estimated uncertainty (Fig. 20) displayed patterns similar to the 6-IOP study (Fig. 6). Notably, the MDN aleatoric fraction for OOD $a_{ph}(443)$ estimates was only 63%, corresponding to relatively high variability between model instances. In the other combinations of scenario and IOP, the 4-IOP MDN was in line with the other architectures. There were no systematic differences in coverage between the 6-IOP and 4-IOP studies.

We conclude from this sensitivity study that only the MDN is significantly affected by the inclusion of $a_{CDOM}(675)$ and $a_{NAP}(675)$ as outputs. The difference with the other four architectures is likely due to the different principle of operation and loss function (Section 3.2); while all five perform simultaneous estimation of the IOPs, only the MDN estimates the full covariance matrix through Cholesky decomposition. Further work should be done to determine to what extent the MDN accuracy is sensitive to the inclusion of additional variables in general (increased dimensionality) vs. the inclusion of $a_{CDOM}(675)$ and $a_{NAP}(675)$ specifically. Notably, the MDN accuracy differed and varied most in the WD and OOD scenarios, indicating a lower generalization ability that has not been adequately studied or accounted for in previous studies (Section 1).

Appendix G. PRISMA: additional results for a_{cdom} and a_{nap}

This appendix includes additional PRISMA estimates for a_{CDOM} and a_{NAP} (see Figs. 21–26), omitted in the results due to consistently worse PNN accuracy compared to a_{ph} at 443 and 675 nm.

Data availability

The dataset is made available on ZENODO. Please see the Code availability section of the manuscript.

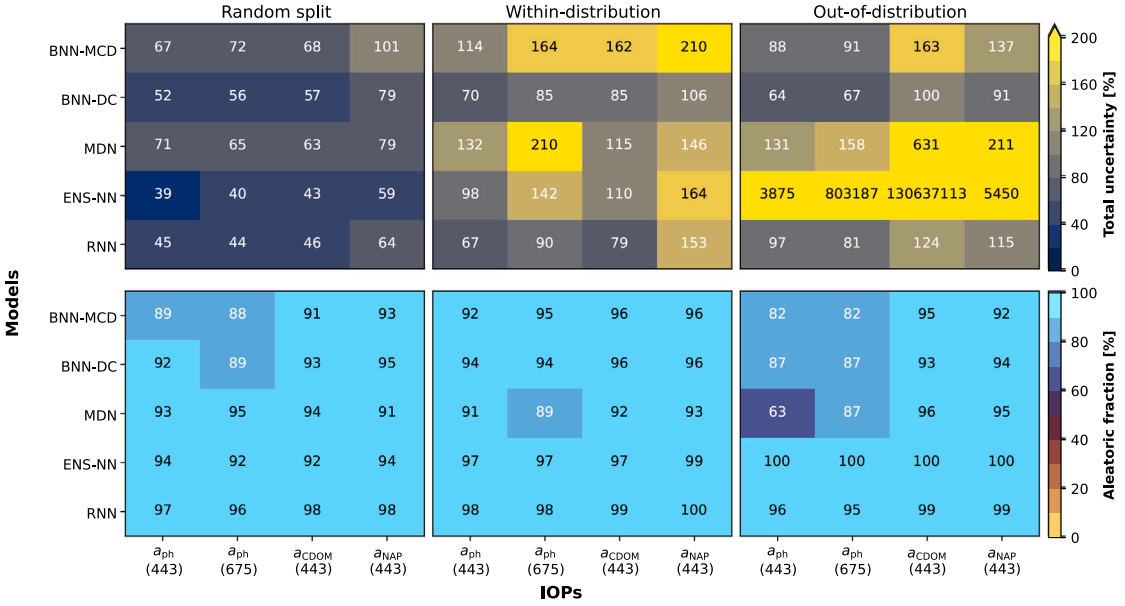

Fig. 20. Median uncertainty in PNN estimates from the average-performing models trained without $a_{CDOM}(675)$ and $a_{NAP}(675)$, analogous to Fig. 6.

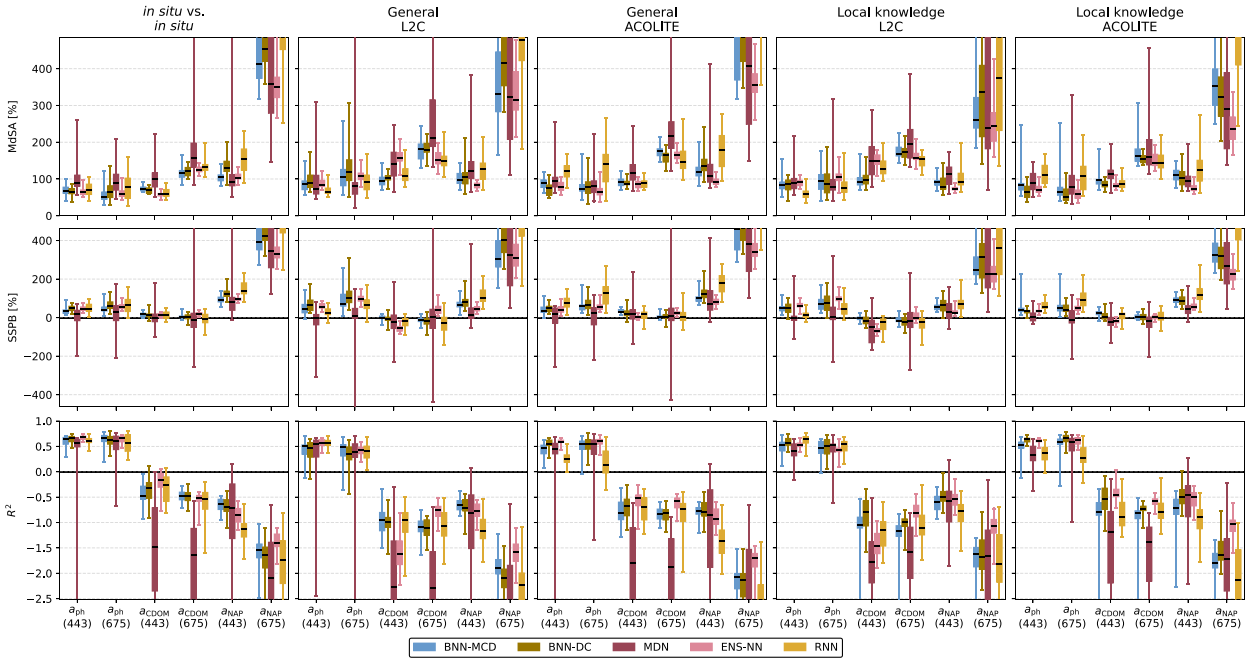

Fig. 21. Accuracy metrics for the full PRISMA dataset, analogous to Fig. 11.

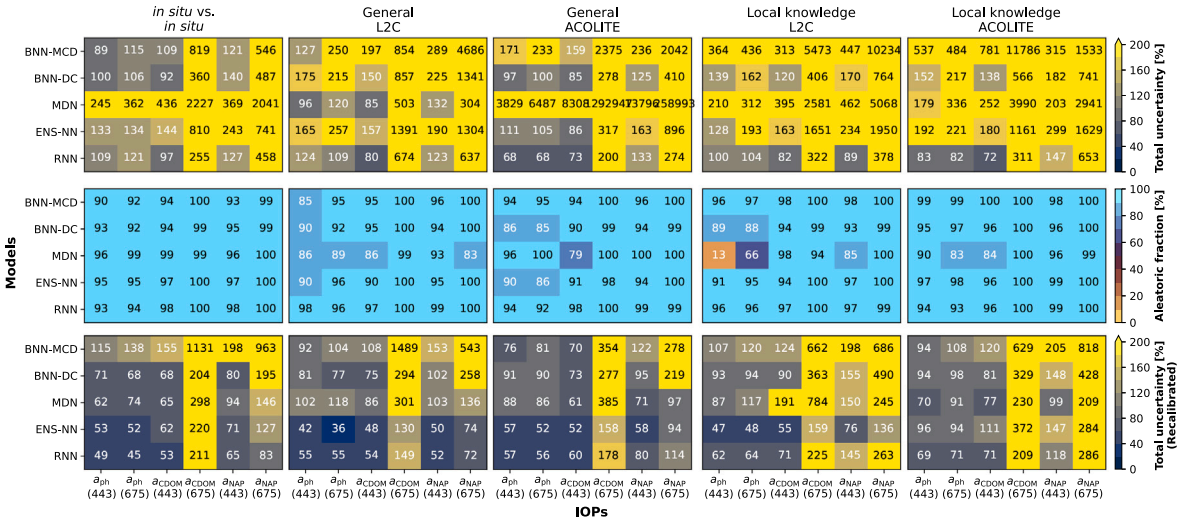

Fig. 22. Median uncertainty in PNN estimates for the full PRISMA dataset, analogous to Fig. 12.

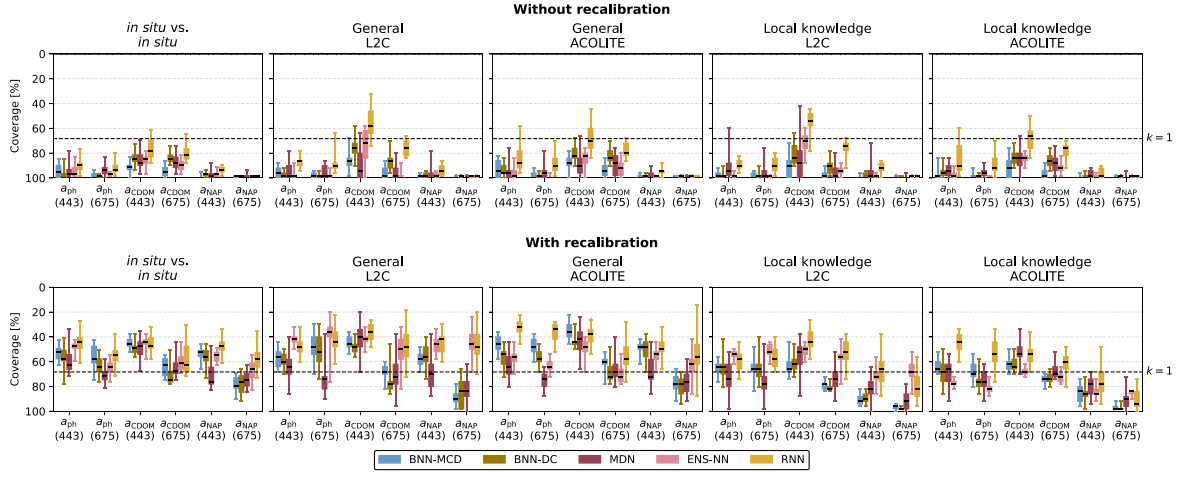

Fig. 23. Coverage of the uncertainty estimates for the full PRISMA dataset, analogous to Fig. 13.

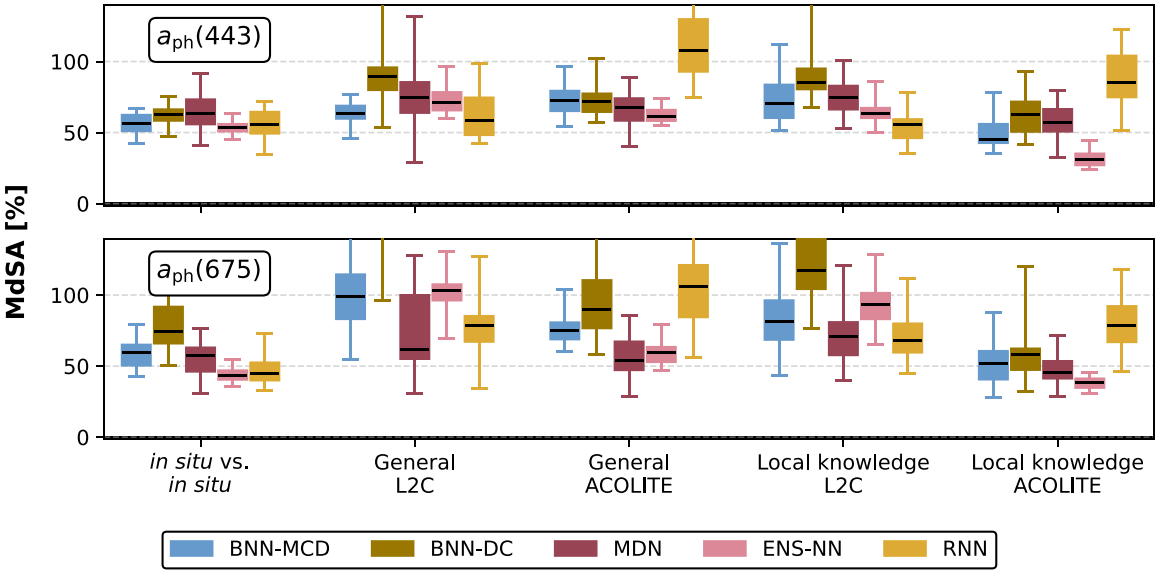

Fig. 24. Median symmetric accuracy (MdSA) with recalibration, analogous to Fig. 11.

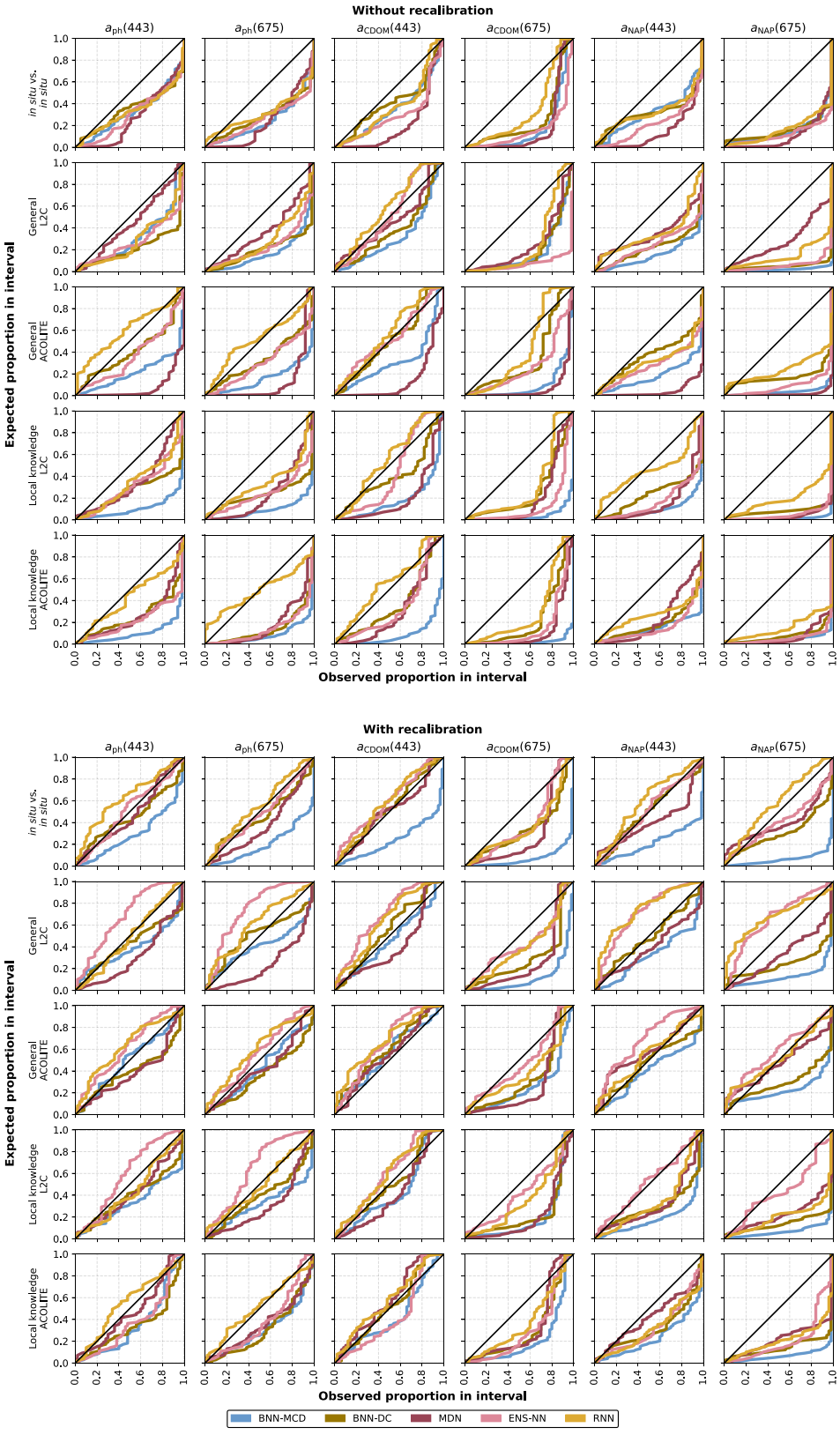

Fig. 25. Calibration curves for the full PRISMA dataset, analogous to Fig. 14.

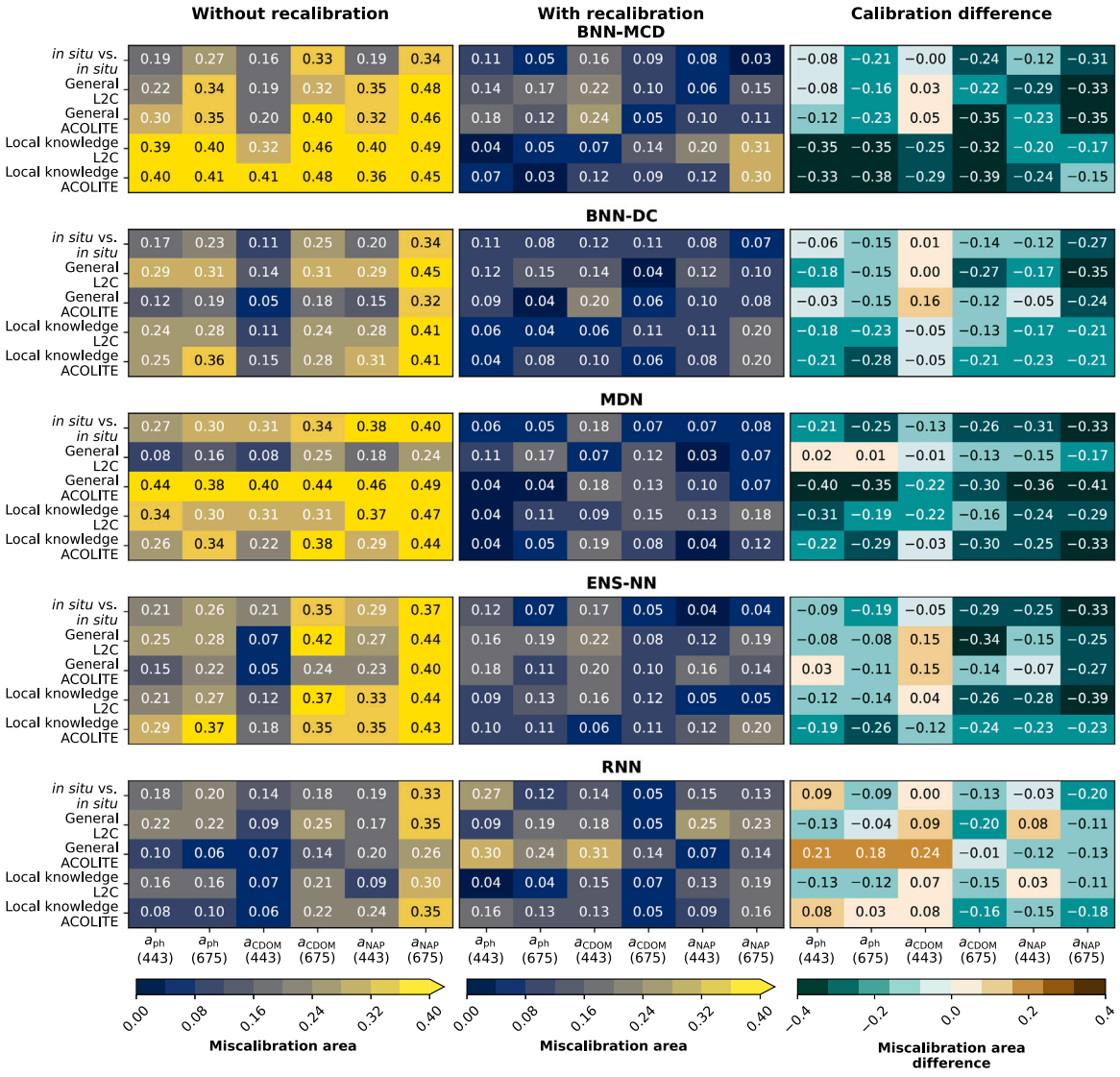

Fig. 26. Miscalibration area for the average-performing models without recalibration (left), with recalibration applied (middle) and the calibration difference (right) for the PRISMA scenarios. A negative calibration difference represents a beneficial recalibration outcome (Section 3.4.2).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. In: OSDI'16, USENIX Association, Savannah, Georgia, USA, pp. 265–283.

ASI, 2021. PRISMA Algorithm Theoretical Basis Document (ATBD). Technical Report 1, ASI – Italian Space Agency, p. 121.

Astuti, I.S., Mishra, D.R., Mishra, S., Schaeffer, B., 2018. Spatio-temporal dynamics of inherent optical properties in oligotrophic northern Gulf of Mexico estuaries. Cont. Shelf Res. 166, 92–107. http://dx.doi.org/10.1016/j.csr.2018.06.016.

Bailey, S.W., Werdell, P.J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sens. Environ. 102 (1–2), 12–23. http://dx.doi.org/10.1016/j.rse.2006.01.015.

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M., 2018. Automatic differentiation in machine learning: A survey. J. Mach. Learn. Res. 18 (153), 1–43.

Behrenfeld, M.J., Westberry, T.K., Boss, E.S., O'Malley, R.T., Siegel, D.A., Wiggert, J.D., Franz, B.A., McClain, C.R., Feldman, G.C., Doney, S.C., Moore, J.K., Dall'Olmo, G., Milligan, A.J., Lima, I., Mahowald, N., 2009. Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. Biogeosciences 6 (5), 779–794. http://dx.doi.org/10.5194/bg-6-779-2009.

Bi, S., Hieronymi, M., 2024. Holistic optical water type classification for ocean, coastal, and inland waters. Limnol. Oceanogr. 69 (7), 1547–1561. http://dx.doi.org/10.1002/lno.12606.

Bishop, C.M., 1994. Mixture Density Networks. Technical Report NCRG/94/004, Aston University, Birmingham, UK, p. 26.

Braga, F., Fabbretto, A., Vanhellemont, Q., Bresciani, M., Giardino, C., Scarpa, G.M., Manfè, G., Concha, J.A., Brando, V.E., 2022. Assessment of PRISMA water reflectance using autonomous hyperspectral radiometry. ISPRS J. Photogramm. Remote Sens. 192, 99–114. http://dx.doi.org/10.1016/j.isprsjprs.2022.08.009.

Braga, F., Scarpa, G.M., Brando, V.E., Manfè, G., Zaggia, L., 2020. COVID-19 lockdown measures reveal human impact on water transparency in the Venice Lagoon. Sci. Total Environ. 736, 139612. http://dx.doi.org/10.1016/j.scitotenv.2020.139612.

Bricaud, A., Mejia, C., Blondeau-Patissier, D., Claustre, H., Crepon, M., Thiria, S., 2007. Retrieval of pigment concentrations and size structure of algal populations from their absorption spectra using multilayered perceptrons. Appl. Opt. 46 (8), 1251–1260. http://dx.doi.org/10.1364/AO.46.001251.

Brockmann, C., Doerffer, R., Peters, M., Stelzer, K., Embacher, S., Ruescas, A., 2016. Evolution of the C2RCC neural network for sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In: ESA Living Planet Symposium (LPS2016), vol. 740, ESA, Prague, Czechia, p. 54.

Bukata, R.P., Harris, G.P., Bruton, J.E., 1974. The detection of suspended solids and chlorophyll a utilizing multispectral ERTS-1 data. In: Proceedings of the Second Canadian Symposium on Remote Sensing. Canadian Remote Sensing Society, Guelph, Ontario, Canada, pp. 551–564.

Burggraaff, O., 2020. Biases from incorrect reflectance convolution. Opt. Express 28 (9), 13801–13816. http://dx.doi.org/10.1364/OE.391470.

Cael, B.B., Bisson, K., Boss, E., Erickson, Z.K., 2023. How many independent quantities can be extracted from ocean color? Limnol. Ocean. Lett. 8 (4), 603–610. http://dx.doi.org/10.1002/lol2.10319.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., Xue, K., 2020. A machine learning approach to estimate chlorophyll-a from landsat-8 measurements in inland lakes. Remote Sens. Environ. 248, 111974. http://dx.doi.org/10.1016/j.rse.2020.111974.

Cao, Z., Ma, R., Pahlevan, N., Liu, M., Melack, J.M., Duan, H., Xue, K., Shen, M., 2022. Evaluating and optimizing VIIRS retrievals of chlorophyll-a and suspended particulate matter in turbid lakes using a machine learning approach. IEEE Trans. Geosci. Remote Sens. 60, 4211417. http://dx.doi.org/10.1109/TGRS.2022.3220529.

Carder, K., 1997. Okeechobee. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/OKEECHOBEE/DATA001.

Carder, K., 1998. TOTO. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/TOTO/DATA001.

Carder, K., Hu, C., 2005. West_Florida_Shelf. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/WEST_FLORIDA_SHELF/DATA001.

Carder, K., Kirkpatrick, G., 1998. RED_TIDE. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/RED_TIDE/DATA001.

Carder, K., Mitchell, G., 1999. ECOHAB. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/ECOHAB/DATA001.

Casey, K.A., Rousseaux, C.S., Gregg, W.W., Boss, E., Chase, A.P., Craig, S.E., Mouw, C.B., Reynolds, R.A., Stramski, D., Ackleson, S.G., Bricaud, A., Schaeffer, B., Lewis, M.R., Maritorena, S., 2020. A global compilation of in situ aquatic high spectral resolution inherent and apparent optical property data for remote sensing applications. Earth Syst. Sci. Data 12 (2), 1123–1139. http://dx.doi.org/10.5194/essd-12-1123-2020.

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., Kashinath, K., 2022. Towards physics-inspired data-driven weather forecasting: Integrating data assimilation with a deep spatial-transformer-based u-NET in a case study with ERA5. Geosci. Model. Dev. 15 (5), 2221–2237. http://dx.doi.org/10.5194/gmd-15-2221-2022.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: encoder–decoder approaches. In: Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M. (Eds.), Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics, Doha, Qatar, pp. 103–111. http://dx.doi.org/10.3115/v1/W14-4012.

Chung, Y., Char, I., Guo, H., Schneider, J., Neiswanger, W., 2021. Uncertainty toolbox: An open-source library for assessing, visualizing, and improving uncertainty quantification. http://dx.doi.org/10.48550/arXiv.2109.10254, arXiv:2109.10254.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning. http://dx.doi.org/10.48550/arXiv.1412.3555.

Cogliati, S., Sarti, F., Chiarantini, L., Cosi, M., Lorusso, R., Lopinto, E., Miglietta, F., Genesio, L., Guanter, L., Damm, A., Pérez-López, S., Scheffler, D., Tagliabue, G., Panigada, C., Rascher, U., Dowling, T.P.F., Giardino, C., Colombo, R., 2021. The PRISMA imaging spectroscopy mission: Overview and first performance analysis. Remote Sens. Environ. 262, 112499. http://dx.doi.org/10.1016/j.rse.2021.112499.

Cota, G., Zimmerman, R., 2000. Chesapeake_Light_Tower. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/CHESAPEAKE_LIGHT_TOWER/DATA001.

Crameri, F., Shephard, G.E., Heron, P.J., 2020. The misuse of colour in science communication. Nat. Commun. 11, 5444. http://dx.doi.org/10.1038/s41467-020-19160-7.

D'Alimonte, D., Melin, F., Zibordi, G., Berthon, J.-F., 2003. Use of the novelty detection technique to identify the range of applicability of empirical ocean color algorithms. IEEE Trans. Geosci. Remote Sens. 41 (12), 2833–2843. http://dx.doi.org/10.1109/TGRS.2003.818020.

Defoin-Platel, M., Chami, M., 2007. How ambiguous is the inverse problem of ocean color in coastal waters? J. Geophys. Res.: Ocean. 112 (C3), http://dx.doi.org/10.1029/2006JC003847.

Dierssen, H.M., Vandermeulen, R.A., Barnes, B.B., Castagna, A., Knaeps, E., Vanhellemont, Q., 2022. QWIP: a quantitative metric for quality control of aquatic reflectance spectral shape using the apparent visible wavelength. Front. Remote. Sens. 3, 869611. http://dx.doi.org/10.3389/frsen.2022.869611.

Doerffer, R., Schiller, H., 1998. Pigment Index, Sediment and Gelbstoff Retrieval from Directional Water Leaving Radiance Reflectances Using Inverse Modelling Technique. Technical Report Algorithm Theoretical Basis Document ATBD 2.12, ESA Coc. No. PO-TN-MEL-GS-0005, European Space Agency.

Doerffer, R., Schiller, H., 2007. The MERIS case 2 water algorithm. Int. J. Remote Sens. 28 (3–4), 517–535. http://dx.doi.org/10.1080/01431160600821127.

Donnelly, J., Daneshkhah, A., Abolfathi, S., 2024. Physics-informed neural networks as surrogate models of hydrodynamic simulators. Sci. Total Environ. 912, 168814. http://dx.doi.org/10.1016/j.scitotenv.2023.168814.

Effler, S.W., Prestigiacomo, A.R., Peng, F., Bulygina, K.B., Smith, D.G., 2006. Resolution of turbidity patterns from runoff events in a water supply reservoir, and the advantages of in situ beam attenuation measurements. Lake Reserv. Manag. 22 (1), 79–93. http://dx.doi.org/10.1080/07438140609353886.

Gal, Y., Ghahramani, Z., 2016a. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gal, Y., Ghahramani, Z., 2016b. A theoretically grounded application of dropout in recurrent neural networks. In: Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc, Barcelona, Spain, http://dx.doi.org/10.48550/arXiv.1512.05287.

Gege, P., 2014. WASI-2D: A software tool for regionally optimized analysis of imaging spectrometer data from deep and shallow waters. Comput. Geosci. 62, 208–215. http://dx.doi.org/10.1016/j.cageo.2013.07.022.

Gilerson, A., Herrera-Estrella, E., Foster, R., Agagliate, J., Hu, C., Ibrahim, A., Franz, B., 2022. Determining the primary sources of uncertainty in retrieval of marine remote sensing reflectance from satellite ocean color sensors. Front. Remote. Sens. 3, 857530. http://dx.doi.org/10.3389/frsen.2022.857530.

Gonçalves-Araujo, R., Rabe, B., Peeken, I., Bracher, A., 2018. High colored dissolved organic matter (CDOM) absorption in surface waters of the central-eastern arctic ocean: implications for biogeochemistry and ocean color algorithms. PLOS ONE 13 (1), e0190838. http://dx.doi.org/10.1371/journal.pone.0190838.

González Vilas, L., Spyrakos, E., Torres Palenzuela, J.M., 2011. Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician Rias (NW Spain). Remote Sens. Environ. 115 (2), 524–535. http://dx.doi.org/10.1016/j.rse.2010.09.021.

Gray, P.C., Boss, E., Prochaska, J.X., Kerner, H., Begouen Demeaux, C., Lehahn, Y., 2024. The promise and Pitfalls of machine learning in ocean remote sensing. Oceanography 37 (3), 52–63. http://dx.doi.org/10.5670/oceanog.2024.511.

Guanter, L., Ruiz-Verdú, A., Odermatt, D., Giardino, C., Simis, S., Estellés, V., Heege, T., Domínguez-Gómez, J.A., Moreno, J., 2010. Atmospheric correction of ENVISAT/MERIS data over inland waters: validation for European lakes. Remote Sens. Environ. 114 (3), 467–480. http://dx.doi.org/10.1016/j.rse.2009.10.004.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. ICML'17, In: Proceedings of the 34th International Conference on Machine Learning, vol.70, JMLR.org, Sydney, NSW, Australia, pp. 1321–1330. http://dx.doi.org/10.48550/arXiv.1706.04599.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2, arXiv:2006.10256.

Hedley, J.D., Mobley, C.D., 2021. HydroLight 6.0 Ecolight 6.0 Technical Documentation. Technical Report, Numerical Optics Ltd, Tiverton, United Kingdom, p. 139.

Hieronymi, M., Müller, D., Doerffer, R., 2017. The OLCI neural network swarm (ONNS): a bio-geo-optical algorithm for open ocean and coastal waters. Front. Mar. Sci. 4, 140. http://dx.doi.org/10.3389/fmars.2017.00140.

Hommersom, A., Peters, S., Wernand, M.R., de Boer, J., 2009. Spatial and temporal variability in bio-optical properties of the wadden sea. Estuar. Coast. Shelf Sci. 83 (3), 360–370. http://dx.doi.org/10.1016/j.ecss.2009.03.042.

Hooker, S., Hu, C., Jordan, C., Lee, Z., Mannino, A., Miller, R., Muller-Karger, F., Ondrusek, M., Salisbury, J., 2011. GEO-CAPE. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/GEO-CAPE/DATA001.

Hoyer, S., Hamman, J., 2017. Xarray: N-D labeled arrays and datasets in Python. J. Open Res. Softw. 5 (1), 10. http://dx.doi.org/10.5334/jors.148.

Hu, C., 2008. Tampa_Bay. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/TAMPA_BAY/DATA001.

Hu, C., 2010a. Big_Bend. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/BIG_BEND/DATA001.

Hu, C., 2010b. SWFL. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/SWFL/DATA001.

Hu, H., Han, Y., Li, X., Jiang, L., Che, L., Liu, T., Zhai, J., 2022. Physics-informed neural network for polarimetric underwater imaging. Opt. Express 30 (13), 22512–22522. http://dx.doi.org/10.1364/OE.461074.

Hu, C., Muller-Karger, F., 2012. SFP. SeaBASS, http://dx.doi.org/10.5067/SeaBASS/SFP/DATA001.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. 110 (3), 457–506. http://dx.doi.org/10.1007/s10994-021-05946-3.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95. http://dx.doi.org/10.1109/MCSE.2007.55.

IOCCG, 2019. Uncertainties in Ocean Colour Remote Sensing, vol. 18, International Ocean Colour Coordinating Group (IOCCG), Dartmouth, Canada, http://dx.doi.org/10.25607/OBP-696.

Irani Rahaghi, A., Odermatt, D., Anneville, O., Sepúlveda Steiner, O., Reiss, R.S., Amadori, M., Toffolon, M., Jacquet, S., Harmel, T., Werther, M., Soulignac, F., Dambrine, E., Jézéquel, D., Hatté, C., Tran-Khac, V., Rasconi, S., Rimet, F., Bouffard, D., 2024. Combined earth observations reveal the sequence of conditions leading to a large algal bloom in lake geneva. Commun. Earth Environ. 5, 229. http://dx.doi.org/10.1038/s43247-024-01351-5.

Jha, M.K., Chowdary, V.M., 2007. Challenges of using remote sensing and GIS in developing nations. Hydrogeol. J. 15, 197–200. http://dx.doi.org/10.1007/s10040-006-0117-1.

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., Kumar, V., 2019. Physics guided RNNs for modeling dynamical systems: a case study in simulating lake temperature profiles. In: Proceedings of the 2019 SIAM International Conference on Data Mining. SDM, In: Proceedings, Society for Industrial and Applied Mathematics, pp. 558–566. http://dx.doi.org/10.1137/1.9781611975673.63.

Jiang, D., Matsushita, B., Pahlevan, N., Gurlin, D., Fichot, C.G., Harringmeyer, J., Sent, G., Brito, A.C., Brotas, V., Werther, M., Mascarenhas, V., Blake, M., Hunter, P., Tyler, A., Spyrakos, E., 2023. Estimating the concentration of total suspended solids in inland and coastal waters from sentinel-2 MSI: A semi-analytical approach. ISPRS J. Photogramm. Remote Sens. 204, 362–377. http://dx.doi.org/10.1016/j.isprsjprs.2023.09.020.

Jorge, D.S.F., Loisel, H., Jamet, C., Dessailly, D., Demaria, J., Bricaud, A., Maritorena, S., Zhang, X., Antoine, D., Kutser, T., Bélanger, S., Brando, V.O., Werdell, J., Kwiatkowska, E., Mangin, A., Fanton d'Andon, O., 2021. A three-step semi analytical algorithm (3SAA) for estimating inherent optical properties over oceanic, coastal, and inland waters from remote sensing reflectance. Remote Sens. Environ. 263, 112537. http://dx.doi.org/10.1016/j.rse.2021.112537.

Keiner, L.E., Yan, X.-H., 1998. A neural network model for estimating sea surface Chlorophyll and Sediments from thematic mapper imagery. Remote Sens. Environ. 66 (2), 153–165. http://dx.doi.org/10.1016/S0034-4257(98)00054-6.

Köhler, J., Varga, E., Spahr, S., Gessner, J., Stelzer, K., Brandt, G., Mahecha, M.D., Kraemer, G., Pusch, M., Wolter, C., Monaghan, M.T., Stöck, M., Goldhammer, T., 2024. Unpredicted ecosystem response to compound human impacts in a European river. Sci. Rep. 14, 16445. http://dx.doi.org/10.1038/s41598-024-66943-9.

Krishna, S., Han, T., Gu, A., Jabbari, S., Wu, Z.S., Lakkaraju, H., 2024. The disagreement problem in explainable machine learning: a practitioner's perspective. Trans. Mach. Learn. Res. http://dx.doi.org/10.48550/arXiv.2202.01602.

Krogh, A., Hertz, J.A., 1991. A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems, vol. 4, Morgan-Kaufmann, Denver, Colorado, USA, pp. 950–957.

Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate uncertainties for deep learning using calibrated regression. In: Proceedings of the 35th International Conference on Machine Learning. PMLR, Stockholm, Sweden, pp. 2796–2804. http://dx.doi.org/10.48550/arXiv.1807.00263.

Kvålseth, T.O., 1985. Cautionary note about R^2. Amer. Statist. 39 (4), 279–285. http://dx.doi.org/10.1080/00031305.1985.10479448.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, Curran Associates, Inc, Long Beach, CA, USA, pp. 6405–6416. http://dx.doi.org/10.48550/arXiv.1612.01474.

Lee, Z., Carder, K.L., Arnone, R.A., 2002. Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters. Appl. Opt. 41 (27), 5755–5772. http://dx.doi.org/10.1364/AO.41.005755.

Lehmann, M.K., Gurlin, D., Pahlevan, N., Alikas, K., Conroy, T., Anstee, J., Balasubramanian, S.V., Barbosa, C.C.F., Binding, C., Bracher, A., Bresciani, M., Burtner, A., Cao, Z., Dekker, A.G., Di Vittorio, C., Drayson, N., Errera, R.M., Fernandez, V., Ficek, D., Fichot, C.G., Gege, P., Giardino, C., Gitelson, A.A., Greb, S.R., Henderson, H., Higa, H., Rahaghi, A.I., Jamet, C., Jiang, D., Jordan, T., Kangro, K., Kravitz, J.A., Kristoffersen, A.S., Kudela, R., Li, L., Ligi, M., Loisel, H., Lohrenz, S., Ma, R., Maciel, D.A., Malthus, T.J., Matsushita, B., Matthews, M., Minaudo, C., Mishra, D.R., Mishra, S., Moore, T., Moses, W.J., Nguyễn, H., Novo, E.M.L.M., Novoa, S., Odermatt, D., O'Donnell, D.M., Olmanson, L.G., Ondrusek, M., Oppelt, N., Ouillon, S., Pereira Filho, W., Plattner, S., Verdú, A.R., Salem, S.I., Schalles, J.F., Simis, S.G.H., Siswanto, E., Smith, B., Somlai-Schweiger, I., Soppa, M.A., Spyrakos, E., Tessin, E., van der Woerd, H.J., Vander Woude, A., Vandermeulen, R.A., Vantrepotte, V., Wernand, M.R., Werther, M., Young, K., Yue, L., 2023. GLORIA - a globally representative hyperspectral in situ dataset for optical sensing of water quality. Sci. Data 10, 100. http://dx.doi.org/10.1038/s41597-023-01973-y.

Leymarie, E., Doxaran, D., Babin, M., 2010. Uncertainties associated to measurements of inherent optical properties in natural waters. Appl. Opt. 49 (28), 5415–5436. http://dx.doi.org/10.1364/AO.49.005415.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep learning for hyperspectral image classification: an overview. IEEE Trans. Geosci. Remote Sens. 57 (9), 6690–6709. http://dx.doi.org/10.1109/TGRS.2019.2907932.

Liu, X., Steele, C., Simis, S., Warren, M., Tyler, A., Spyrakos, E., Selmes, N., Hunter, P., 2021. Retrieval of Chlorophyll-a concentration and associated product uncertainty in optically diverse lakes and reservoirs. Remote Sens. Environ. 267, 112710. http://dx.doi.org/10.1016/j.rse.2021.112710.

Loisel, H., Jorge, D.S.F., Reynolds, R.A., Stramski, D., 2023. A synthetic optical database generated by radiative transfer simulations in support of studies in ocean optics and optical remote sensing of the global ocean. Earth Syst. Sci. Data 15 (8), 3711–3731. http://dx.doi.org/10.5194/essd-15-3711-2023.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc, Long Beach, California, USA, http://dx.doi.org/10.48550/arXiv.1705.07874.

McKerns, M.M., Strand, L., Sullivan, T., Fang, A., Aivazis, M.A.G., 2011. Building a framework for predictive science. In: Proceedings of the 10th Python in Science Conference. http://dx.doi.org/10.25080/Majora-ebaa42b7-00d.

McKinney, W., 2010. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. pp. 56–61. http://dx.doi.org/10.25080/Majora-92bf1922-00a.

Meyer, M.F., Topp, S.N., King, T.V., Ladwig, R., Pilla, R.M., Dugan, H.A., Eggleston, J.R., Hampton, S.E., Leech, D.M., Oleksy, I.A., Ross, J.C., Ross, M.R.V., Woolway, R.I., Yang, X., Brousil, M.R., Fickas, K.C., Padowski, J.C., Pollard, A.I., Ren, J., Zwart, J.A., 2024. National-scale remotely sensed lake trophic state from 1984 through 2020. Sci. Data 11 (1), 77. http://dx.doi.org/10.1038/s41597-024-02921-0.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M., 2021. Revisiting the calibration of modern neural networks. In: Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc, Online, pp. 15682–15694. http://dx.doi.org/10.48550/arXiv.2106.07998.

Moore, T.S., Campbell, J.W., Feng, H., 2001. A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. IEEE Trans. Geosci. Remote Sens. 39 (8), 1764–1776. http://dx.doi.org/10.1109/36.942555.

Moore, T.S., Dowell, M.D., Bradt, S., Ruiz Verdu, A., 2014. An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. Remote Sens. Environ. 143, 97–111. http://dx.doi.org/10.1016/j.rse.2013.11.021.

Morel, A., 1980. In-water and remote measurements of ocean color. Bound.-Layer Meteorol. 18 (2), 177–201. http://dx.doi.org/10.1007/BF00121323.

Morel, A., Prieur, L., 1977. Analysis of variations in ocean color. Limnol. Oceanogr. 22 (4), 709–722. http://dx.doi.org/10.4319/lo.1977.22.4.0709.

Morley, S.K., Brito, T.V., Welling, D.T., 2018. Measures of model performance based on the log accuracy ratio. Space Weather. 16 (1), 69–88. http://dx.doi.org/10.1002/2017SW001669.

Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 55 (7), 3639–3655. http://dx.doi.org/10.1109/TGRS.2016.2636241.

Mouw, C.B., Chen, H., McKinley, G.A., Effler, S., O'Donnell, D., Perkins, M.G., Strait, C., 2013. Evaluation and optimization of bio-optical inversion algorithms for remote sensing of lake superior's optical properties. J. Geophys. Res. Ocean. 118 (4), 1696–1714. http://dx.doi.org/10.1002/jgrc.20139.

Muller-Karger, F., 2015. SFMBON. SeaBASS, http://dx.doi.org/10.5067/SEABASS/SFMBON/DATA001.

Neil, C., Spyrakos, E., Hunter, P.D., Tyler, A.N., 2019. A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sens. Environ. 229, 159–178. http://dx.doi.org/10.1016/j.rse.2019.04.027.

O'Shea, R.E., Pahlevan, N., Smith, B., Boss, E., Gurlin, D., Alikas, K., Kangro, K., Kudela, R.M., Vaičiūtė, D., 2023. A hyperspectral inversion framework for estimating absorbing inherent optical properties and biogeochemical parameters in inland and coastal waters. Remote Sens. Environ. 295, 113706. http://dx.doi.org/10.1016/j.rse.2023.113706.

Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M., Giardino, C., Gurlin, D., Fan, Y., Harmel, T., Hunter, P., Ishikaza, J., Kratzer, S., Lehmann, M.K., Ligi, M., Ma, R., Martin-Lauzer, F.-R., Olmanson, L., Oppelt, N., Pan, Y., Peters, S., Reynaud, N., Sander de Carvalho, L.A., Simis, S., Spyrakos, E., Steinmetz, F., Stelzer, K., Sterckx, S., Tormos, T., Tyler, A., Vanhellemont, Q., Warren, M., 2021. ACIX-aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. Remote Sens. Environ. 258, 112366. http://dx.doi.org/10.1016/j.rse.2021.112366.

Pahlevan, N., Smith, B., Alikas, K., Anstee, J., Barbosa, C., Binding, C., Bresciani, M., Cremella, B., Giardino, C., Gurlin, D., Fernandez, V., Jamet, C., Kangro, K., Lehmann, M.K., Loisel, H., Matsushita, B., Hà, N., Olmanson, L., Potvin, G., Simis, S.G.H., VanderWoude, A., Vantrepotte, V., Ruiz-Verdù, A., 2022. Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. Remote Sens. Environ. 270, 112860. http://dx.doi.org/10.1016/j.rse.2021.112860.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., Matsushita, B., Moses, W., Greb, S., Lehmann, M.K., Ondrusek, M., Oppelt, N., Stumpf, R., 2020. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. Remote Sens. Environ. 240, 111604. http://dx.doi.org/10.1016/j.rse.2019.111604.

Palmer, S.C.J., Kutser, T., Hunter, P.D., 2015. Remote sensing of inland waters: challenges, progress and future directions. Remote Sens. Environ. 157, 1–8. http://dx.doi.org/10.1016/j.rse.2014.09.021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (85), 2825–2830.

Pellegrino, A., Fabbretto, A., Bresciani, M., de Lima, T.M.A., Braga, F., Pahlevan, N., Brando, V.E., Kratzer, S., Gianinetto, M., Giardino, C., 2023. Assessing the accuracy of PRISMA standard reflectance products in globally distributed aquatic sites. Remote. Sens. 15 (8), 2163. http://dx.doi.org/10.3390/rs15082163.

Raissi, M., Perdikaris, P., Karniadakis, G.E., 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.

Rasmussen, M.H., Duan, C., Kulik, H.J., Jensen, J.H., 2023. Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. J. Cheminformatics 15, 121. http://dx.doi.org/10.1186/s13321-023-00790-0.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, Association for Computing Machinery, San Francisco, California, United States, pp. 1135–1144. http://dx.doi.org/10.1145/2939672.2939778.

Rollo, C., 2024. Cmcrameri.

Saranathan, A.M., Smith, B., Pahlevan, N., 2023. Per-pixel uncertainty quantification and reporting for satellite-derived Chlorophyll-a estimates via mixture density networks. IEEE Trans. Geosci. Remote Sens. 61, 4200718. http://dx.doi.org/10.1109/TGRS.2023.3234465.

Saranathan, A.M., Werther, M., Balasubramanian, S.V., Odermatt, D., Pahlevan, N., 2024. Assessment of advanced neural networks for the dual estimation of water quality indicators and their uncertainties. Front. Remote. Sens. 5, 1383147. http://dx.doi.org/10.3389/frsen.2024.1383147.

Schaeffer, B., Salls, W., Coffer, M., Lebreton, C., Werther, M., Stelzer, K., Urquhart, E., Gurlin, D., 2022a. Merging of the case 2 Regional Coast colour and maximum-peak height chlorophyll-a algorithms: Validation and demonstration of satellite-derived retrievals across US lakes. Environ. Monit. Assess. 194 (3), 179. http://dx.doi.org/10.1007/s10661-021-09684-w.

Schaeffer, B.A., Urquhart, E., Coffer, M., Salls, W., Stumpf, R.P., Loftin, K.A., Werdell, P.J., 2022b. Satellites quantify the spatial extent of cyanobacterial blooms across the United States at multiple scales. Ecol. Indic. 140, 108990. http://dx.doi.org/10.1016/j.ecolind.2022.108990.

Schiller, H., Doerffer, R., 1999. Neural network for emulation of an inverse model operational derivation of case II water properties from MERIS data. Int. J. Remote Sens. 20 (9), 1735–1746. http://dx.doi.org/10.1080/014311699212443.

Schiller, H., Krasnopolsky, V., 2001. Domain check for input to NN emulating an inverse model. In: IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), vol. 3, IEEE, Washington, D.C., USA, pp. 2150–2152. http://dx.doi.org/10.1109/IJCNN.2001.938499.

Seegers, B.N., Stumpf, R.P., Schaeffer, B.A., Loftin, K.A., Werdell, P.J., 2018. Performance metrics for the assessment of satellite data products: An ocean color case study. Opt. Express 26 (6), 7404–7422. http://dx.doi.org/10.1364/OE.26.007404.

Silsbe, G.M., Behrenfeld, M.J., Halsey, K.H., Milligan, A.J., Westberry, T.K., 2016. The CAFE model: A net production model for global ocean Phytoplankton. Glob. Biogeochem. Cycles 30 (12), 1756–1777. http://dx.doi.org/10.1002/2016GB005521.

Smith, B., Pahlevan, N., Schalles, J., Ruberg, S., Errera, R., Ma, R., Giardino, C., Bresciani, M., Barbosa, C., Moore, T., Fernandez, V., Alikas, K., Kangro, K., 2021. A Chlorophyll-a algorithm for landsat-8 based on mixture density networks. Front. Remote. Sens. 1, 623678. http://dx.doi.org/10.3389/frsen.2020.623678.

Spyrakos, E., O'Donnell, R., Hunter, P.D., Miller, C., Scott, M., Simis, S.G.H., Neil, C., Barbosa, C.C.F., Binding, C.E., Bradt, S., Bresciani, M., Dall'Olmo, G., Giardino, C., Gitelson, A.A., Kutser, T., Li, L., Matsushita, B., Martinez-Vicente, V., Matthews, M.W., Ogashawara, I., Ruiz-Verdú, A., Schalles, J.F., Tebbs, E., Zhang, Y., Tyler, A.N., 2018. Optical types of inland and coastal waters. Limnol. Oceanogr. 63 (2), 846–870. http://dx.doi.org/10.1002/lno.10674.

Stock, A., Gregr, E.J., Chan, K.M.A., 2023. Data leakage jeopardizes ecological applications of machine learning. Nat. Ecol. Evol. 7, 1743–1745. http://dx.doi.org/10.1038/s41559-023-02162-1.

Stock, A., Subramaniam, A., 2022. Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing. GIScience Remote. Sens. 59 (1), 1281–1300. http://dx.doi.org/10.1080/15481603.2022.2107113.

Stoudt, S., Pintar, A., Possolo, A., 2021. Coverage intervals. J. Res. Natl. Inst. Stand. Technol. 126, 126004. http://dx.doi.org/10.6028/jres.126.004.

Tang, W., Llort, J., Weis, J., Perron, M.M.G., Basart, S., Li, Z., Sathyendranath, S., Jackson, T., Sanz Rodriguez, E., Proemse, B.C., Bowie, A.R., Schallenberg, C., Strutton, P.G., Matear, R., Cassar, N., 2021. Widespread Phytoplankton blooms triggered by 2019–2020 Australian wildfires. Nature 597 (7876), 370–375. http://dx.doi.org/10.1038/s41586-021-03805-8.

Tsallis, C., Stariolo, D.A., 1996. Generalized simulated annealing. Phys. A 233 (1–2), 395–406. http://dx.doi.org/10.1016/S0378-4371(96)00271-3.

Valdenegro-Toro, M., Mori, D.S., 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, IEEE, New Orleans, Louisiana, USA, pp. 1508–1516. http://dx.doi.org/10.1109/CVPRW56347.2022.00157, arXiv:2204.09308.

Vanhellemont, Q., Ruddick, K., 2018. Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. Remote Sens. Environ. 216, 586–597. http://dx.doi.org/10.1016/j.rse.2018.07.015.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17 (3), 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2, arXiv:1907.10121.

Wager, S., Wang, S., Liang, P.S., 2013. Dropout training as adaptive regularization. In: Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc, Lake Tahoe, Nevada, USA, pp. 351–359. http://dx.doi.org/10.48550/arXiv.1307.1493.

Wan, L., Zeiler, M., Zhang, S., LeCun, Y., Fergus, R., 2013. Regularization of neural networks using DropConnect. In: Proceedings of the 30th International Conference on Machine Learning, vol. 28, PMLR, Atlanta, Georgia, USA, pp. 1058–1066.

Wang, P., Boss, E.S., Roesler, C., 2005. Uncertainties of inherent optical properties obtained from semianalytical inversions of ocean color. Appl. Opt. 44 (19), 4074–4085. http://dx.doi.org/10.1364/AO.44.004074.

Warren, M.A., Simis, S.G.H., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C., Ansper, A., 2019. Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral imager over coastal and inland waters. Remote Sens. Environ. 225, 267–289. http://dx.doi.org/10.1016/j.rse.2019.03.018.

Werdell, P.J., Bailey, S., Fargion, G., Pietras, C., Knobelspiesse, K., Feldman, G., McClain, C., 2003. Unique data repository facilitates ocean color satellite validation. Eos Trans. Am. Geophys. Union 84 (38), 377–387. http://dx.doi.org/10.1029/2003EO380001.

Werdell, P.J., Franz, B.A., Bailey, S.W., Feldman, G.C., Boss, E., Brando, V.E., Dowell, M., Hirata, T., Lavender, S.J., Lee, Z., Loisel, H., Maritorena, S., Mélin, F., Moore, T.S., Smyth, T.J., Antoine, D., Devred, E., Fanton d'Andon, O.H., Mangin, A., 2013. Generalized ocean color inversion model for retrieving marine inherent optical properties. Appl. Opt. 52 (10), 2019–2037. http://dx.doi.org/10.1364/AO.52.002019.

Werther, M., Burggraaff, O., 2023. Dive into the unknown: embracing uncertainty to advance aquatic remote sensing. J. Remote. Sens. 3, 0070. http://dx.doi.org/10.34133/remotesensing.0070.

Werther, M., Odermatt, D., Simis, S.G.H., Gurlin, D., Lehmann, M.K., Kutser, T., Gupana, R., Varley, A., Hunter, P.D., Tyler, A.N., Spyrakos, E., 2022. A Bayesian approach for remote sensing of chlorophyll-a and associated retrieval uncertainty in oligotrophic and mesotrophic lakes. Remote Sens. Environ. 283, 113295. http://dx.doi.org/10.1016/j.rse.2022.113295.

Werther, M., Spyrakos, E., Simis, S.G.H., Odermatt, D., Stelzer, K., Krawczyk, H., Berlage, O., Hunter, P., Tyler, A., 2021. Meta-classification of remote sensing reflectance to estimate trophic status of inland and nearshore waters. ISPRS J. Photogramm. Remote Sens. 176, 109–126. http://dx.doi.org/10.1016/j.isprsjprs.2021.04.003.

Zaneveld, J.R.V., 1994. Optical closure: From theory to measurement. In: Ocean Optics. Oxford University Press, New York City, New York, USA, pp. 59–72. http://dx.doi.org/10.1093/oso/9780195068436.003.0007.