Mixture Density Networks Improve Inter-Sensor Consistency of Optical Water Type Classification

Davide Lomeo, Stefan G. H. Simis, Mark A. Warren, David Moffat, Anne D. Jungblut, and Emma J. Tebbs

***Abstract*—In aquatic remote sensing, Optical Water Type (OWT) pre-classification has been used to map suitable algorithms to optically diverse targets and over optical gradients. However, current methods can yield inconsistent type assignments across sensors with different spectral capabilities. Probabilistic neural networks are increasingly used for constituent retrieval in optically complex inland water bodies because they can deliver higher accuracy than empirical or semi-analytical algorithms, while also quantifying estimation uncertainty. However, their interpretability within bio-optical theory remains elusive. To address the limitations brought by these two methodological streams, we train Mixture Density Networks (MDN) to predict OWT membership distributions. OWTs serve as an optical-biogeochemical interpretable scaffold, alongside associated classification uncertainty, while the classification misalignment between Sentinel-3 Ocean and Land Colour Instrument (OLCI) and Sentinel-2 Multispectral Instrument (MSI) is reduced. Using > 29,000 co-located OLCI and MSI observations from 59 lakes, MDN ensembles trained to reproduce OLCI-derived OWT membership scores achieved 99% overall accuracy for OLCI and improved MSI cross-sensor OWT agreement from 58% to 73%, compared to using spectral angle as OWT distance metric. The MDNs enabled identification of aleatoric (upstream bias) and epistemic (lack of representative data) components of the uncertainty envelope, with post-hoc uncertainty recalibration used to adjust uncertainty magnitudes to reliable confidence intervals, reducing miscalibration by 93% for OLCI and 80% for MSI, for both 68% and 95% intervals. Application of trained MDN ensembles across optically diverse systems confirmed that they learned generalisable optical relationships, improving OWT classification agreement across sensors, and providing diagnostics for bias identification. This approach preserves the interpretability of OWTs while enabling uncertainty aware, cross-mission processing, identifying likely sources of classification uncertainty, particularly useful for less capable sensors, which can guide appropriate post-processing and suitable algorithm selection and blending.**

***Index Terms*—Aquatic remote sensing, Optical Water Types (OWT), inland waters, Mixture Density Network (MDN), uncertainty, Sentinel-3, Sentinel-2, machine learning (ML).**

This work was supported by the Natural Environment Research Council (NERC) under Grant NE/S007229/1. *(Corresponding author: Davide Lomeo).*

Davide Lomeo is with King's College London, Department of Geography, London, United Kingdom (e-mail: davide.lomeo@kcl.ac.uk).

Stefan G. H. Simis is with Plymouth Marine Laboratory, Plymouth, United Kingdom (e-mail: stsi@pml.ac.uk).

Mark A. Warren is with Plymouth Marine Laboratory, Plymouth, United Kingdom (e-mail: mark1@pml.ac.uk).

David Moffat is with Plymouth Marine Laboratory, Plymouth, United Kingdom (e-mail: dmof@pml.ac.uk).

Anne D. Jungblut is with Natural History Museum, Department of Life Sciences, London, United Kingdom (e-mail: a.jungblut@nhm.ac.uk).

Emma J. Tebbs A. is with King's College London, Department of Geography, London, United Kingdom (e-mail: emma.tebbs@kcl.ac.uk).

This article has supplementary downloadable material available at [TO BE CONFIRMED], provided by the authors. The code is available at https://github.com/davidelomeo/mdn_owt.git

I. INTRODUCTION

MONITORING indicators of water quality in inland water bodies is challenging due to their optical complexity and variability. Optical Water Type (OWT) classifications provide an interpretable way to partition water bodies by predetermined optical characteristics. These are typically derived from clustering *in situ* or satellite-derived observations, creating bio-optical baselines for subsequent satellite remote sensing applications. Satellite capabilities have evolved both in the domain of ocean colour sensors, suitable to quantify specific water constituents in medium to large water bodies on the one hand, and sensors to derive land cover properties, featuring higher spatial but lower radiometric and spectral resolution. However, differences in waveband configurations have been shown to directly affect type assignment across multispectral missions [1], and OWT classification methods currently lack quantification of classification uncertainty to express these discrepancies. Machine learning approaches have demonstrated superior performance for water constituent retrieval [2], [3], [4], yet their learned representations operate outside established radiative transfer principles, limiting their interpretability. Addressing the cross-sensor consistency challenge while preserving the bio-optical interpretability of OWT classifications would support robust operational water quality monitoring across missions.

Decades of algorithm development have advanced the estimation of optically active water constituents, including chlorophyll-a (as phytoplankton biomass proxy) and phycocyanin (cyanobacteria) quantification [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] and suspended matter estimation across varying turbidity conditions [16], [17], [18]. However, algorithm performance remains constrained by the specific bio-optical relationships each approach attempts to exploit. The challenge of algorithm transferability across optical conditions has been approached using various switching techniques determined by localised spectral features wavebands or threshold amplitudes [12], [18], [19], and through optical pre-classification based on spectral shape. In the latter category, OWT classifications have been constructed to encompass and discretise the optical variability of natural waters in oceanic and coastal environments [20], [21] and inland water bodies [22], [23]. Building upon spectral end-member decomposition methodologies [24], [25], [26], OWTs facilitate dynamic algorithm selection through spectral similarity metrics, such as Mahalonobis distance [27] or the spectral angle [28], whereby satellite-derived reflectance is compared to reference OWT definitions. This enables

weighted ensemble predictions from constituent-specific retrieval algorithms validated within each type [27], [29]. One such classification for inland waters comprises 13 OWTs derived from a global-scale community repository of inland water bio-optical *in situ* observations including hyperspectral reflectance [22], providing the foundation for the European Space Agency (ESA) Lakes Climate Change Initiative (Lakes_cci) biogeochemical product generation [29], [30].

The operational use of OWT classifications for fuzzy algorithm blending encounters several limitations that affect their reliability across diverse optical conditions and sensor configurations. Firstly, OWTs may under-represent optical conditions that are inherently challenging to observe *in situ*, such as, for instance, optically shallow water, or near- or at-surface biomass accumulations which are easily disturbed by vessels. In addition, the OWT similarity scores that guide fuzzy algorithm blending lack explicit probabilistic quantification of classification confidence, which may impact subsequent algorithm selection and blending. Further, optically different water types under hyperspectral conditions can become indistinguishable when characterised by fewer sensor wavebands, thereby conflating ecologically distinct biogeochemical conditions and limiting cross-sensor classification consistency [1].

A fundamental limitation of optical remote sensing for water quality indicators remains the inverse problem, which involves deriving inherent optical properties, or optically active substances, from satellite-derived reflectance. The problem is inherently ill-conditioned, with non-uniqueness of solutions representing a fundamental mathematical challenge [31], [32]. This confronts the premises of deterministic linear, semi-analytical and Neural Network (NN)-type algorithms, whether applied in isolation or blended within OWT classifications, because they assume that unique solutions can be derived from spectral observations. Recent developments of probabilistic NNs have demonstrated resilience to some of these issues, including Bayesian or Recurrent NNs and Mixture Density Networks (MDN), because they explicitly model the non-uniqueness of solutions through distributional outputs rather than point estimates [4], [33], [34], [35], [36], [37], [38]. Among these, MDNs hold particular promise in ocean colour applications because they natively output full conditional probability distributions, addressing the multimodal nature of the inverse problem by representing multiple plausible solutions within a unified predictive framework [39]. Nevertheless, these approaches rely on learned representations within their NN backbones, meaning that when applied directly to biogeochemical parameter retrieval, the link between spectral observations and established optical principles remains unclear.

Beyond addressing the multimodal inverse problem, probabilistic NNs provide distributional outputs that enable uncertainty quantification [40], [41], a critical requirement for ocean colour remote sensing [42]. Uncertainty decomposition allows discerning uncertainty arising from sensor noise and upstream biases that cannot be reduced through additional data (aleatoric), from uncertainty reflecting model ignorance about unsampled optical conditions (epistemic) [43], [44]. Yet, while the relative magnitude of uncertainties offers meaningful insights, probabilistic NNs typically deliver over-confident predictive distributions [45], [46] that increase with predictive accuracy [47]. This likely occurs because training objectives do not account for uncertainty in model weights, making explicit regularisation or calibration necessary [43], [48], [49]. As such, post-hoc recalibration techniques help to align predicted confidence intervals with empirical error rates. These include Isotonic Regression [50], already used in recalibrating probabilistic NNs in remote sensing [51], and the single-parameter variant of Platt scaling [52], also known as temperature scaling [47]. Via recalibration, estimated uncertainties provide statistically reliable confidence intervals that can propagate to downstream products.

There is an opportunity to leverage MDNs in conjunction with OWT classification, whereby instead of predicting water constituents directly, MDNs predict how satellite-derived reflectance maps onto well-understood bio-optical conditions underpinned by operational OWT definitions. This approach has the potential to address an apparent gap in current OWT classifications by providing previously unavailable uncertainty quantification for type assignments. While advanced machine learning approaches have already demonstrated superior statistical performance for constituent retrieval (at least within their training scope), OWT classifications provide the optical-biogeochemical interpretability preserving decades of algorithm development linked to bio-optical principles. Thus, OWT classification uncertainties may be more meaningfully interpreted than uncertainties in direct constituent estimates, which can conflate multiple error sources including algorithm limitations, atmospheric correction biases, and fundamental ambiguities in the optical inverse problem.

Here, we develop and evaluate a probabilistic framework using MDNs to enable consistent OWT classification between Sentinel-3 Ocean and Land Colour Instrument (OLCI), one of the most capable operational multispectral ocean colour sensors with daily global coverage, and Sentinel-2 Multispectral Instrument (MSI), which provides complementary high spatial resolution observations but with fewer and broader wavebands in the visible domain. Using co-located observations from globally distributed lakes spanning diverse trophic states and optical conditions, we pursue three objectives. First, we train individually initialised MDN models (instances) for both OLCI and MSI to predict OLCI-derived OWT membership similarity scores, assessing whether multi-model averages (MDN ensembles – more below) provide more consistent OWT classifications than individual instances. Second, we decompose the resulting uncertainties into aleatoric and epistemic components and apply post-hoc recalibration to adjust their magnitudes to statistically reliable confidence intervals. Finally, we evaluate whether this framework reduces cross-sensor OWT classification misalignment between OLCI and MSI and assess the generalisability of trained ensembles to systems outside the training set. Ideally, this would enable missions to provide global water quality assessments across spatial scales,

consistent within estimated product uncertainty. Given the accelerating pace of environmental change in inland waters, establishing such probabilistic approaches may prove essential to maximise observational capacity across expanding satellite constellations and evolving spectral capabilities.

II. Datasets

A. Satellite data set

A globally representative dataset of satellite data was compiled to cover a range of lake size, trophic status, and optical water types from all continents except for Antarctica (Fig. 1). The size of the selected 59 lakes ranged from 16 km2 to 58,000 km2 (Table I). Concurrent images acquired by Sentinel-3A/B OLCI and Sentinel-2A/B MSI (within ±1.5 hours) were selected from the period 2022-2023 to include two full seasonal cycles. The number of days of concurrent observations per lake ranged from 2 to 78 days across the two years, commonly depending on the size of the lake. Images were processed using the candidate configuration for version 3.0 of the ESA Lakes_cci Climate Research Data Package, using Polymer v4.17b for atmospheric correction and obtain fully-normalised water-leaving reflectance (R_w) spectra [29], [30]. This configuration uses an extended range of initialisation conditions compared to previous versions, which is understood to improve retrieval of turbid water conditions including near or at-surface blooms [53], [54].

B. Sampling strategy and data quality assessment

Atmospherically corrected images were sampled using 3 × 3 pixel windows (macro-pixels) for OLCI, and 15 × 15 pixel windows for MSI, at approximately 300 m and 60 m equivalent spatial resolution at the equator, respectively. OLCI and MSI macro-pixels were co-located using the latitude and longitude coordinates of their central pixels. Macro-pixels for which the central pixel was flagged as non-valid during atmospheric correction due to clouds, shadows, or land adjacency, were removed. Macro-pixels that contained > 25% non-valid pixels, were also removed, following recommendations by Bailey & Verdell [55] and EUMETSAT [56]. MSI macro-pixels in which any of the pixels within the central 5 × 5 window, corresponding to the central pixel of the co-located OLCI macro-pixel, was flagged as non-valid, were also removed. The homogeneity of the spectra in the central window of the MSI macro-pixel was verified with the spectral angle metric (see section II.C for definition). Each central pixel spectrum was compared against all other pixels within the central 5 × 5 window, yielding a minimum spectral angle of 0.99 across all MSI macro-pixels. This confirmed that regions were sufficiently homogeneous for the central pixel to be used as a proxy for the whole area, allowing comparison with same-day OLCI observations. All macro-pixels in which the central pixel showed R_w > 0.4 in bands 412, 560 and 865 nm for OLCI and 443, 560, and 865 nm for MSI, as well as having a dominant OWT membership score < 0.8 were removed as likely affected by shadow or land-adjacency biases not flagged during atmospheric correction, or presenting optical conditions that were unlikely representative of optically deep water. Only co-located central pixels were used in the analysis, with a total of 29,639 observations.

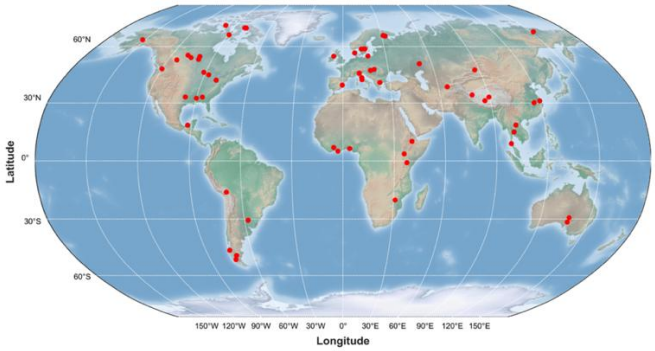

Fig. 1. Locations of lakes where co-located OLCI-MSI observations were used in the period 2022-2023.

C. Optical Water Type classifications

R_w spectra from all co-located observations were mapped onto the library of 13 OWTs for inland waters developed by Spyrakos et al. [22] and generally followed the same procedures as used in the Lakes_cci [29], [30]. In short, the hyperspectral OWT library of Spyrakos et al. [22] was formulated by clustering *in situ* hyperspectral R_w collated from the (no longer available) Lake Bio-optical Measurements and Matchup Data for Remote Sensing (LIMNADES) community repository onto a discrete number of classes. The resulting 13 OWTs are then used for algorithm selection and blending to generate the Lakes_cci R_w-derived products by convolving the hyperspectral OWT definitions to the spectral response of wavebands in common between MERIS and OLCI, or MODIS sensors, in the range 412-779 nm, excluding oxygen absorption bands centred at 761, 764 and 767 nm and sun-induced fluorescence bands centred at 674 and 681 nm.

All spectra are standardised before classification by dividing over their integrals to reduce the influence of varying reflectance amplitudes, focusing on the similarity of their shapes [29]. A further two type spectra derived from satellite observations were included to flag pixels that were likely affected by land adjacency [58].

The spectral angle metric [28] is then used to obtain OWT membership scores (S_{owt}) for each pixel as recommended by Liu et al. [29], with the equations:

$$a_j = \cos^{-1} \frac{\sum_{i=1}^{n} p_i r_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \sqrt{\sum_{i=1}^{n} r_i^2}} \quad (1)$$

$$S_{owt\,j} = 1 - a_j/\pi \quad (2)$$

where p_i and r_i are the standardised pixel and reference spectra in band i, respectively. The resultant $S_{owt\,j}$ is the membership score for OWT j, given as a number in range 0-1, where 1 indicates identical spectral shapes.

S_{owt} values for OLCI observations were calculated between the reference library of 13 OWTs convolved to OLCI bands as described above plus the 2 types associated to land adjacency. This set of 15 S_{owt} vectors is henceforth referred to as $OLCI_{SA}$.

TABLE I

NAME, LOCATION, AND SIZE OF THE 59 LAKES SELECTED IN THIS STUDY. THE TROPHIC STATUS WAS OBTAINED BY CALCULATING THE TROPHIC STATE INDEX [57] ON THE MEAN OLCI-DERIVED LAKES_CCI CHLOROPHYLL-A ESTIMATES FOR ALL AVAILABLE OLCI-MSI CO-LOCATED OBSERVATIONS BETWEEN JANUARY 2022 AND DECEMBER 2023. THE COLUMN 'OBS. DAYS' REFERS TO THE NUMBER OF UNIQUE OBSERVATION DAYS FOR EACH LAKE ACROSS THE TWO-YEAR PERIOD.

Continent	Lake Name	Area (Km^2)	Latitude	Longitude	Elevation	Trophic Status	Chla Range	Obs. days
Africa	Masinga	57.43	-0.889	37.49	1,046	Eutrophic	0.02 - 122.79	10
	Mutirikwi	84.43	-20.205	31.008	1,061	Mesotrophic	0.08 - 21.53	6
	Caddabassa	91.31	10.207	40.486	561	Eutrophic	0.03 - 66.51	12
	Lagos	420.28	6.532	3.573	0	Mesotrophic	0.01 - 85.84	21
	Aby	422.71	5.216	-3.161	78	Eutrophic	0.03 - 117.07	11
	Kossou	519.6	7.253	-5.59	239	Eutrophic	0.09 - 120.06	11
	Turkana	7,566.28	3.544	36.187	361	Mesotrophic	1.26 - 96.99	14
Asia	Talimardzhan	43.3	38.365	65.566	389	Mesotrophic	0.59 - 115.86	17
	Chem	110.04	34.16	79.778	4,961	Mesotrophic	1.65 - 41.71	6
	Yaggain	111.66	33.017	89.792	4,872	Mesotrophic	2.31 - 26.6	6
	Po	124.93	30.164	116.443	9	Eutrophic	2.31 - 73.96	10
	Pasak Chonlasit	153.68	14.982	101.057	39	Eutrophic	0.16 - 108.43	13
	Ratchaprapha	187.75	9.047	98.676	88	Mesotrophic	0.11 - 5.22	13
	Shalkar	233.22	50.559	51.685	13	Eutrophic	20.47 - 104.69	2
	Nam-Ngum	488.74	18.589	102.633	208	Mesotrophic	1.91 - 92.42	8
	Dang Reyongcuo	836.98	31.07	86.609	4,535	Oligotrophic	0.05 - 2.32	8
	Ulungar	864.74	47.259	87.289	478	Mesotrophic	0.12 - 44.13	4
	Taihu	2,416.41	31.199	120.193	0	Eutrophic	1.03 - 167.57	4
Europe	Tisza	16.61	47.493	20.561	83	Eutrophic	3.45 - 24.94	19
	L'Albufera	25.99	39.333	-0.354	-3	Hypereutrophic	3.25 - 113.9	6
	Bracciano	57.58	42.121	12.232	161	Oligotrophic	0.14 - 1.51	15
	Nissum Fjord	61.58	56.355	8.191	-2	Mesotrophic	0.02 - 46.41	8
	Volvi	68.3	40.669	23.482	30	Eutrophic	1.87 - 63.67	3
	Łebsko	70.08	54.717	17.411	-4	Eutrophic	0.28 - 159.16	18
	Chonduya	70.85	68.579	152.273	22	Eutrophic	0.04 - 162.04	11
	Glan	71.1	58.622	15.958	22	Eutrophic	0.14 - 83.16	16
	Trasimeno	120.98	43.137	12.103	253	Eutrophic	1.99 - 95.02	36
	Kukkureozero	195.87	65.919	32.888	88	Oligotrophic	0.07 - 98.72	7
	Iovskoye	245.55	66.538	30.936	70	Mesotrophic	0.06 - 48.54	10
	Garda	368.64	45.662	10.686	62	Oligotrophic	0.09 - 127.15	22
	Neagh	381.87	54.622	-6.405	10	Mesotrophic	0.07 - 137.91	8
	Balaton	583.59	46.883	17.846	100	Mesotrophic	1.38 - 57.83	30
	Vättern	1,887.88	58.33	14.537	88	Mesotrophic	0.03 - 66.04	19
	Vänern	5,500.90	58.907	13.298	44	Mesotrophic	0.05 - 186.44	15
North America	Katimik	50.33	52.884	-99.363	254	Eutrophic	0.21 - 122.06	15
	Minchumina	58.25	63.89	-152.235	195	Eutrophic	0.07 - 83.73	24
	Rice	61.25	45.927	-91.194	397	Mesotrophic	0.09 - 31.03	4
	Arsenault	67.52	55.099	-108.52	472	Mesotrophic	0.3 - 43.56	13
	Hargrave	80.06	54.477	-99.672	260	Mesotrophic	2.12 - 25.44	5
	Gull	80.82	52.537	-114	895	Mesotrophic	0.2 - 46.04	3
	Mitchell	85.77	32.932	-86.444	217	Eutrophic	0.08 - 47.2	10
	Ross R Barnett	87.72	32.453	-90.015	89	Oligotrophic	0.05 - 12.46	6
	Atasta	88.18	18.585	-92.148	5	Eutrophic	0.07 - 53.45	6
	Nina Bang	106.23	70.853	-79.41	52	Oligotrophic	0.05 - 20.33	6
	Lewisville	108.4	33.133	-96.982	158	Eutrophic	2.49 - 93.2	5
	Chelan	132.04	48.037	-120.355	331	Oligotrophic	0.07 - 26.91	12
	Candle	132.8	53.828	-105.305	491	Mesotrophic	0.08 - 127.58	15
	Quartz	137.12	70.968	-80.623	110	Mesotrophic	2.61 - 40.2	2
	Crooked	139.01	72.616	-98.503	14	Mesotrophic	2.22 - 39.51	9
	Curtis	149.01	66.719	-89.188	304	Mesotrophic	0.03 - 8.1	78
	Erie	25,937.72	42.144	-81.238	172	Mesotrophic	0.12 - 128.21	71
	Michigan	58,256.63	43.854	-87.081	175	Oligotrophic	0.03 - 104.42	9
Oceania	Torrens	4,264.39	-30.847	137.722	30	Eutrophic	0.02 - 25.35	8
	Eyre	9,400.67	-28.602	137.311	-15	Eutrophic	0.11 - 43.09	5
South America	El Toro	208.78	-51.206	-72.741	24	Oligotrophic	0.08 - 3.15	8
	Presidente Rios	310.01	-46.46	-74.405	16	Mesotrophic	0.03 - 3,120.73	16
	Cardiel	356.54	-48.917	-71.216	285	Mesotrophic	0.04 - 25.76	14
	Salto Grande	607.11	-30.934	-57.87	32	Eutrophic	0.14 - 138.18	4
	Titicaca	7,752.93	-15.916	-69.303	3,815	Oligotrophic	0.08 - 130.6	12

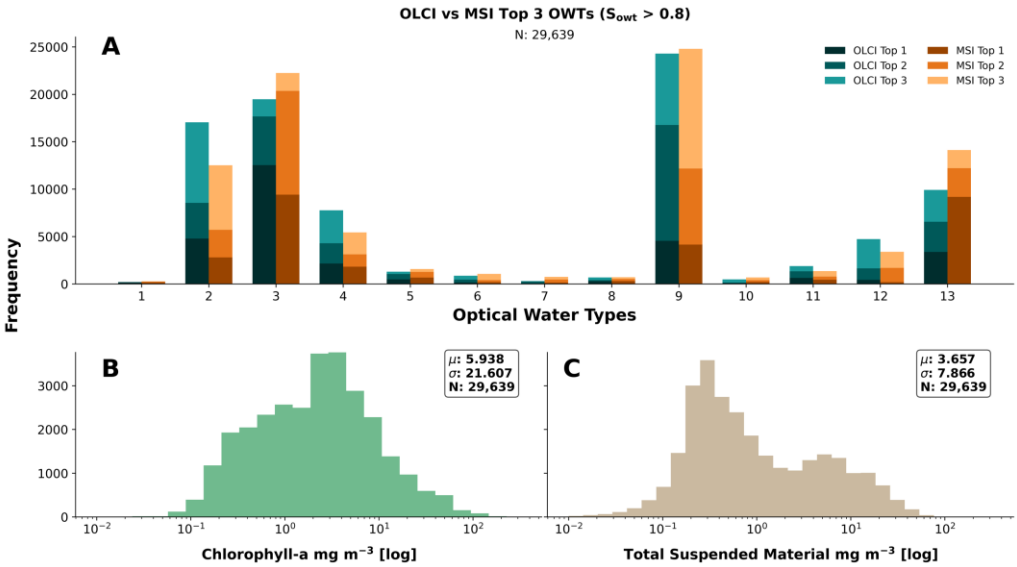

Fig. 2. **(A)** Frequency distribution of top-3 Optical Water Types (OWT) as determined by the spectral angle metric for OLCI and MSI after removal of observations with OWT membership scores (S_{owt}) < 0.8. **(B-C)** Histograms of chlorophyll-a and total suspended matter across the sampled 59 lakes using blended algorithms.

S_{owt} values for MSI were similarly calculated from the hyperspectral library of 13 OWTs, convolved to MSI bands in the range 443-783 nm. The two types associated to land adjacency have not been defined for MSI and were therefore not included. This set of 13 S_{owt} vectors is henceforth referred to as MSI_{SA}. However, when training MDN models, MSI spectra were mapped onto the $OLCI_{SA}$ vectors associated to the 15 types, hence obtaining 2 land adjacency types previously unavailable for MSI (see section III.A).

For fuzzy blending, including all S_{owt} scores risks introducing out-of-scope algorithms that can bias the blended estimates even at low similarity weights. On the other hand, relying on the single highest S_{owt} score, also referred to as dominant OWT [27], may not fully capture the optical conditions of a given observation and lead to algorithm edge effects. Combining the top-3 ranking OWTs represents a practical balance between including sufficient optical features for robust algorithm blending and excluding OWTs whose associated algorithms would operate entirely outside their valid scope. This approach is consistent with the Lakes_cci processing chain, where top-3 weighting was found to capture an appropriate range of likely optical conditions and avoid edge effects [29], [30]. Fig. 2 shows the distribution of the top-3 ranking OWTs and the resulting chlorophyll-a (Chla) and Total Suspended Matter (TSM) concentrations for the present data set as derived in the Lakes_cci processing chain for OLCI, excluding observations with S_{owt} < 0.8 because typically associated to upstream observational biases.

III. Methods

A. Mixture Density Network

MDNs are a class of neural networks (NN) with a dual architecture, comprising of a NN followed by a mixture layer head that provides the final outputs (Fig. 3). The NN part of the model consists of a series of hidden layers and associated activation functions (here we used the Rectified Linear Unit, ReLU), each composed of a number of nodes, that project the inputs into higher dimensions to understand the underlying structure of the data. The mixture layer head translates the mapping from the NN latent space into a conditional probability distribution as a mixture of multiple functions [39]. This probability density is expressed as follows:

$$p(y|R_w) = \sum_{i=1}^{M} \alpha_i(R_w)\phi_i(y|R_w) \tag{3}$$

where

$$\phi_i(y|R_w) = \frac{exp\left\{-\frac{1}{2}(y-\mu_i)^{\mathrm{T}}\sum_i^{-1}(y-\mu_i)\right\}}{\sqrt{(2\pi)^d|\Sigma_i|}}$$

$$s.t. \sum_{i=1}^{M} \alpha_i = 1;\ \alpha_i \geq 0$$

where M is the number of components in the mixture. $\alpha_i(R_w)$ represents the mixture coefficient, or prior probability, conditioned on R_w, of the target y vector (i.e., the Sowt values associated to the 15 types) generated from the i^{th} component of the mixture. All coefficients are always positive and sum to 1. $\phi_i(y|R_w)$ is the probability density function (PDF), here Gaussian, of the i^{th} component of the mixture of the target y vector, where μ_i is the centre (i.e., mean prediction) of the PDF. and Σ_i is the covariance matrix, which must be positive definite, and commonly regularised for a small ε to avoid numerical instability. The covariance matrix is parameterised using a Cholesky decomposition which computes the lower triangular matrix to save computational overhead and ensure positivity. The term $(2\pi)^d$ at the denominator is a normalisation constant that ensures that the PDF integrates to 1, where d represents the dimensionality of the data. The matrix Σ_i can be replaced by a diagonal covariance vector σ_i^2 if requiring a smaller parameter space.

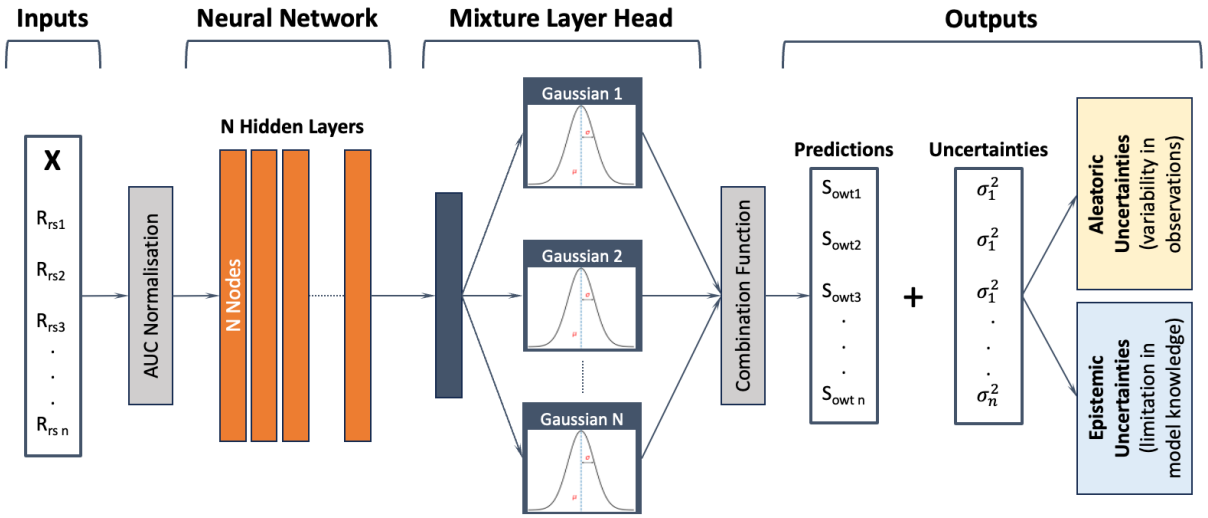

Fig. 3. The Mixture Density Network (MDN) architecture used to estimate Optical Water Types membership scores (S_{owt}) and associated predictive uncertainties. Uncertainties are provided as aleatoric (variability in observations) and epistemic (limitation in model knowledge). AUC normalisation refers to 'Area Under the Curve' normalisation, i.e., spectra are divided by the area between the curve and 0. The combination function refers to the method used for predictions. This can be a weighted mean across predictions from all mixture components (Gaussians) or selecting the component with the highest probability mass.

Thus, for each input R_w, the MDN model provides $M \times n$ Gaussians associated to $M \times n$ mean (predicted) y vectors, and $M \times n$ covariance matrices underpinning the probability distribution, where n is the number of types, here 15.

The maximum likelihood loss function $\mathcal{L}$ was used, computed as the negative log-likelihood over all mixture components:

$$\mathcal{L} = -\frac{1}{B}\sum_{n=1}^{B} \log[p(y|R_w)] \quad (4)$$

This representation aligns with common NN gradient descent optimisation algorithms (here we used the Adam optimisation algorithm) that aim at minimising the loss (hence the negation) between the values estimated by the MDN and the target y. The logarithm of the sum of batches B (then averaged) prevents common underflow problems, where very small numbers may be approximated to zero, causing numerical instability [59]. Also, a five-fold imputation process was implemented to fill missing values and prevent gradient collapse following Rubin [60], and similarly to other MDN applications in ocean colour [36], [61].

The MDN can output either the average prediction weighted by the set of $\alpha_i(R_w)$ priors, or the prediction associated with the highest probability mass, or maximum a-posteriori, i.e., the predictions associated to the highest $\alpha_i(R_w)$ component. These are derived as follows:

$$\hat{y} = \sum_{i=1}^{M} \alpha_i \mu_i \quad (5)$$

$$\hat{y} = \mu_i : i = \text{argmax}_i\, \alpha_i \quad (6)$$

The predictions from the highest probability posterior (6) were used in this study to underscore the need to promote the highest S_{owt} predictions within (expected) highly similar S_{owt} values, particularly considering the likely insufficiently distinct optically conditions when using MSI.

B. MDN initialisation

A known issue with MDNs, and of Gaussian mixture models (GMM) in general, is the tendency for a single mixture to absorb all probability mass irrespective of the number of mixtures used, effectively erasing their multimodality [62]. A common reason is the lack of an appropriate initialisation strategy, with models dwelling in locally optimal solutions and struggling to find global minima [63]. Therefore, the mixture head layer parameters of all MDNs trained in this work were first initialised with weights generated by training a small GMM on $OLCI_{SA}$ for a single pass. This process, also referred to as Expectation–Maximisation algorithm [64], consisted of clustering the 15 S_{owt} vectors in M clusters (i.e., the number of mixtures used) using the kmeans++ algorithm, known to improve cluster quality [65], and retaining the best of 20 random restarts, which is helpful to increase the probability of reaching the global-likelihood solution [63]. This approach provided 'soft initialisations' to inject meaningful reference mixture arrangements at epoch 0, reducing the risk of mode collapse (which was observed during preliminary training rounds) and accelerating early training.

C. Modelling setup

OLCI- and MSI-derived R_w spectra were used as inputs for MDN model training, while $OLCI_{SA}$ was used as the target. A stratified random sampling approach was used due to the high OWT class distribution imbalance of the data set (Fig. 2A), whereby each data split respected the proportional distribution of the dominant OWT. This imbalance was expected and reflects the natural higher occurrence of certain OWTs in inland waters, particularly those associated to clear and low biomass conditions (e.g., OWTs 2, 3, 9, 13) compared to those with mixed and high biomass conditions or surface accumulations (e.g., OWTs 1, 6, 7), in line with a broadly log-normal distribution of phytoplankton biomass in nature and as reported in other studies [29], [66]. The data set was split into

72% reserved for training, 8% for validation, 2% for calibration, and 18% for testing. This resulted in 21,339 observations used for training, 2,371 for model validation, 592 for uncertainty calibration (discussed in section III.F), and 5,337 for model performance testing. The training set was designed to dynamically feed data to the models during learning [67], [68]. This 'undersampling' method consisted of randomly selecting 30 observations from each of the 15 types at the start of every training epoch, creating a balanced pool of 450 samples per epoch. Sampling was performed without replacement, ensuring no duplicate observations were present within the pool of samples at each epoch. This pool was then shuffled and divided into batches. The number of 30 observations per OWT was chosen because it was slightly under the count of the least frequent dominant OWT in the training set (OWT 7, 32 observations), finding a balance between having enough observations for each OWT while allowing for variability between epochs for less represented OWTs. Since the imbalance was severe (the most frequent dominant OWT was 250 times larger than the least frequent one), observations associated to the most frequent OWTs (Fig. 2) inherently had a higher probability of varying across epochs, as different subsets of 30 were selected, whereas less frequent ones with approximately 30 total samples were re-sampled nearly identically across epochs. A smaller number of observations to allow for greater variability between epochs was also tested, but the performance degraded significantly compared to the modelling reported on here.

Three sets of MDN models were trained, one for MSI and two for OLCI. The MSI models were named MSI_{MDN}. OLCI models were named OLCI-F_{MDN} and OLCI-R_{MDN}, with symbol F denoting use of the full range of 10 wavebands in the 412-779 nm region, versus the use of a reduced (R) set of seven wavebands with centres corresponding to those available in MSI, respectively. These seven wavebands were the bands centred at 443, 490, 560, 665, 709 (705), 754 (740), and 779 (783) nm, where numbers in parentheses refer to MSI waveband centres. The OLCI-R_{MDN} models helped to evaluate how OWT covariance changed when S_{owt} values were predicted by a capable ocean colour sensor with the same number of wavebands, positioned at a relatively comparable location in the spectrum, with MSI. While OLCI-F_{MDN} and MSI_{MDN} models were repeatedly trained to find the 'best' performing model configuration (see next section), OLCI-R_{MDN} was only trained on the same configuration of the best OLCI-F_{MDN} model, which was used as reference.

D. Hyperparameter search

While other studies have identified appropriate configurations for MDNs to model optically active substances and water quality indicators [4], [33], [34], [36], the present challenge to model 15 covarying S_{owt} values within a bounded region [0-1] was considered a sufficiently new case to require its own hyperparameter search. Therefore, 144 OLCI-F_{MDN} and MSI_{MDN} models were trained for each set of hyperparameter combinations (Table II) to determine the best performing model configuration.

TABLE II

HYPERPARAMETER COMBINATIONS USED TO TRAIN THE 144 OLCI-F_{MDN} AND MSI_{MDN}, ALONGSIDE THE BEST MODEL CONFIGURATIONS FOR BOTH SENSORS.

Hyperparameters		**Best OLCI Configuration**	**Best MSI Configuration**
Nodes	100, 500, 1000	1,000	1,000
Hidden Layers	3, 5, 7	5	3
Mixtures	3, 5, 7	3	5
Learning Rate	1e-3, 1e-4	1e-4	1e-4
Batch Size	64, 128	64	64
Regularisation (l2)	1e-4	1e-4	1e-4
Epsilon	1e-3	1e-3	1e-3
Covariance Matrix	Σ_i, σ_i^2	Σ_i	Σ_i

The number of iterations was fixed at 10,000, with an early stopping measure activated when the performance of the models did not improve after 500 consecutive epochs evaluated on the validation set loss. All hyperparameters were tested using both the full (Σ_i) and diagonal (σ_i^2) covariance matrices to evaluate the influence of parameter space size and network stability in PDF formulation. Each model was initialised with the same weights (i.e., using the same random seed). The OLCI-F_{MDN} and MSI_{MDN} model configurations showing the lowest mean MdSA across predictions were considered the 'best' configurations (Table II) and re-trained 15 times with different random seeds, though using the same set of seeds across sensors to ensure reproducibility. These 15 models are henceforth called 'instances'. OLCI-R_{MDN} was also trained using 15 different instances using the same configuration as OLCI-F_{MDN}. Given the highly similar results across instances (shown in section IV.C), training a larger number of instances was not considered necessary in this case.

Results are reported in two ways for all three MDNs. First, the single 'best instance' refers to the instance achieving the lowest MdSA on the test set. Second, 'ensemble' models refer to the average predictions and uncertainties across all 15 instances, where the instance with the lowest MdSA is included in this average. Although weighting schemes for Bayesian ensembles exist [69], [70], finding appropriate posterior distribution in multi-modal, high-dimensional settings remains challenging, and the solution can be approximated by simple non-weighted averages [71].

E. Uncertainty estimation

The MDN parameters α_i, μ_i, and their associated covariance matrices (Σ_i or σ_i^2) introduced in section III.A, allow computation of total predictive variance, also referred to as predictive uncertainty. By the law of total variance [72], this can be derived (as standard deviation) as follows:

$$\sigma_{TOT} = \sqrt{\sigma_{ALT}^2 + \sigma_{EPS}^2} \tag{7}$$

where σ_{ALT}^2 and σ_{EPS}^2 represent the *aleatoric* and *epistemic* components of the uncertainty, respectively. Aleatoric uncertainty relates to noise inherent in the observations that cannot be reduced by providing more data to the model [43], which in satellite remote sensing can be associated to sensor-specific biases (measurement errors or calibration issues) and/or atmospheric correction residuals. When derived individually, the aleatoric uncertainty (expressed as standard

deviation) is calculated as follows:

$$\sigma_{ALT} = \sqrt{\sum_{i=1}^{M} \alpha_i \Sigma_i} \quad (8)$$

representing the mean of the M variances in the MDN. On the other hand, epistemic uncertainty, also referred to as *model uncertainty*, relates to uncertainty in the model parameters that could be addressed by providing additional observation data to the model [43], and derived (as standard deviation) as follows:

$$\sigma_{EPS} = \sqrt{\sum_{i=1}^{M} \alpha_i (\mu_i - \hat{y})(\mu_i - \hat{y})^T} \quad (9)$$

representing the variance of the M means in the MDN.

Although predictions in this study were made using the top component of the mixture (6), it is useful to investigate uncertainties from the full set of mixtures available to all MDNs because they show the level of (dis)agreement between individual components towards a solution. Greater uncertainty typically indicates higher multimodality in the solution space, which can be associated to optical conditions that are either not well represented in the latent feature space of the model (epistemic), or that are affected by upstream biases such as atmospheric correction or adjacency effect that the model cannot resolve with additional data (aleatoric).

Two set of uncertainties are reported: the uncertainties from the MDN ensembles, computed as the mean σ_{TOT}, σ_{ALT}, σ_{EPS}, across the 15 instances, and the uncertainties from the single 'best instance' alone (as discussed in section III.D). In the case of ensemble models, the variance across instances predictions, or ensemble spread, was added as epistemic uncertainty [45], [72]. Uncertainties are converted into a percentage value relative to the final point estimate from the models, also known as relative uncertainty, using the following equation:

$$\sigma_*(\%) = \left(\frac{\sigma_*}{\hat{y}}\right) \times 100 \quad (10)$$

where σ_* is used as a placeholder for the uncertainty terms in (7-9). It is worth noting that $\sigma_{TOT}(\%)$ will not equal to the sum of $\sigma_{ALT}(\%)$ and $\sigma_{EPS}(\%)$ when these are derived individually due to the non-linearity of the square root function. Still, it is useful to report these separately to highlight sensor-noise-dominated areas (high aleatoric) versus model-ignorance areas (high epistemic).

F. Uncertainty recalibration

To track the origin of the estimated OWT classification uncertainty, it is useful to retain the out-of-the-box uncertainty components separability provided by MDNs and use a recalibration technique that preserves their additive relationship (7). Non-linear mapping methods such as the Isotonic Regression can be attractive because they help dealing with the boundaries of the prediction intervals, and work especially well if data availability is not a constraint [47], [48]. However, separating aleatoric and epistemic uncertainties after recalibration becomes a non-trivial task as it is not guaranteed that the ranking of the uncertainties is preserved. A parametric (linear) method such as temperature scaling, which finds an optimal τ that minimises the distance between nominal and empirical coverage, preserves such ranking, making it a preferable choice to ensure that the sources of uncertainty remain fully traceable [45], [47], [48].

For a Gaussian predictive distribution, coverage represents the fraction of true values that fall within a predicted confidence interval, defined as follows:

$$Coverage_k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{|y_i - \hat{y}_i| \leq k\sigma_i\} \quad (11)$$

where k is the nominal level of choice, σ_i the estimated uncertainty, $\hat{y}_i$ the predictions. $\mathbf{1}\{|y_i - \hat{y}_i| \leq k\sigma_i\}$ is an indicator that returns 1 when the k-σ band captures the true value and 0 otherwise, so the sum counts how many of the N test cases (observations) are successfully covered. With $k = 1$ (i.e., a nominal level of 68%, or 1-σ uncertainty), a coverage > 68% will denote under-confidence (intervals too wide), whereas a coverage < 68% will denote over-confidence (intervals too tight). This reflects the expectation, for instance, that a well-calibrated 90% confidence interval should contain the true value in 90% of cases [73].

Hence, an optimal recalibration scalar τ was determined by minimising the Miscalibration Area (MA) on the calibration set introduced in section III.C. Model calibration can be visualised by plotting nominal credibility levels p (the intended coverage) against empirical coverage probabilities $\hat{p}$ (the observed fraction of true values within the p-confidence interval). Perfect calibration yields a diagonal line where $p = \hat{p}$ across all confidence levels. The MA, defined as the integral of $|\hat{p} - p|$ over $p \in [0,1]$, ranges from 0 (perfect calibration) to 0.5 (maximally mis-calibrated) [74]. The recalibration scalar τ was obtained by minimising MA independently for each model instance and for ensembles across all sensors.

G. Evaluation of predictive performance and OWT classification accuracy

The performance of all trained MDN models was assessed using the median symmetric accuracy (MdSA) and the Signed Systematic Percentage Bias (SSPB) (Table III). MdSA correlated with the overall classification accuracy of the dominant and the top-3 scoring OWTs, for both sensors (Fig. S1). MdSA and SSPB can be interpreted as unsigned and a signed percentage error, respectively, with perfect accuracy achieved at 0%. These metrics are increasingly being used in aquatic remote sensing because they are symmetric, i.e., they penalise under- and over-estimations equally, and are robust in the presence of outliers [33], [75]. Moreover, since the target S_{owt} values are in range 0-1, and it is common for relatively different OWTs to present values in a narrow range within this interval, it is sensible to project S_{owt} values on a logarithmic scale to highlight small differences.

TABLE III

LIST OF MODEL PERFORMANCE METRICS USED.

Metric	Derivation	
Median Symmetric Accuracy (MdSA)	$\text{MdSA} = 100[exp(Md(\lvert Q \rvert)) - 1]$	
Signed Systematic Percentage Bias (SSPB)	$\text{SSPB} = 100\big[sgn\big(Md(Q)\big)\big][exp(\lvert Md(Q) \rvert) - 1]$	
Root Mean Squared Logarithmic Error (RMSLE)	$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} Q^2}$	Where $Q = \log\left(\frac{\hat{y}}{y}\right)$
Overall Accuracy (OA)	$OA = \frac{\sum_{i=1}^{T} C_{ii}}{\sum_{i=1}^{T}\sum_{j=1}^{T} C_{ij}}$	
Precision	$Precision_i = \frac{C_{ii}}{\sum_{j=1}^{T} C_{ji}}$	where $T = 13$
Recall	$Recall_i = \frac{C_{ii}}{\sum_{j=1}^{T} C_{ij}}$	
F1-Score (F1)	$F1_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i}$	

Q represents the *accuracy ratio* between the predicted S_{owt} vectors and the 'true' y (i.e., $OLCI_{SA}$) in log space.

C_{ij} represents the number of samples with true OWT i (as determined by $OLCI_{SA}$) predicted as OWT j, and C_{ii} the correctly classified samples (or diagonal elements of the confusion matrix).

Precision, recall, F1-score, and Overall Accuracy (OA) were used as metrics to evaluate the classification accuracy of dominant and top-3 OWT predictions (Table III). For dominant OWT predictions, confusion matrices were used to show how accurately the models assigned the highest S_{owt} values across the 13 OWTs, irrespective of the score residuals (or difference between true and predicted S_{owt} values). For the top-3 OWT predictions, model performance was assessed considering both the exact OWT ranking order as determined by $OLCI_{SA}$, and the unordered ranking, i.e., the rate at which OWTs were accurately predicted to be in the top-3 by spectral similarity, irrespective of the order. This approach evaluated the ability of the models to identify the correct group of OWTs that characterise each observation, which is important in downstream products where S_{owt} values are used as weights for algorithm blending, as, for example, in Moore et al. [27] or Liu et al. [29].

The classification performance into ranked OWTs was not extended to the two types associated to land adjacency because these are primarily used as quality control flags rather than OWT classification for algorithm selection. Given the likely small differences between S_{owt} values as mentioned above, it is useful to evaluate the relative deviation between true and predicted values using a log-base metric such as the Root Mean Squared Logarithmic Error (RMSLE; Table III). This metric is highly appropriate to highlight any 'tail-sensitive' multiplicative error that flags potential adjacency signal that affects pixels [76].

IV. RESULTS

A. OLCI and MSI Optical Water Type distribution and spectral alignment

The data set of 29,639 co-located OLCI-MSI observations from 59 lakes (Table I) showed a highly imbalanced OWT frequency distribution, in line with other global studies and expected natural variability in inland water bodies [29], [66]. OWTs commonly associated to high phytoplankton biomass, either mixed in the visible water column or accumulating at the surface (OWTs 1, 6, 7, and 8, as described by Spyrakos et al. [22], were infrequently observed, jointly representing < 0.1% of the total observations in the data set. The frequency distribution of OWT assignments to MSI-derived reflectance for the top-3 ranking OWTs tracked those from OLCI reasonably well, with some inconsistencies (e.g., OWT 2, 3, 9, 13 – Fig. 2). Comparison between OLCI and MSI standardised spectra grouped by dominant OWT (Fig. S2) confirmed that MSI can distinguish between the optically most contrasting water types such as OWT 1 (surface accumulations of cyanobacteria) and 13 (clear, blue appearance) while finding less agreement with OLCI for OWTs associated with more subtly varying contributions of optically active water constituents (e.g., OWT 5, 6, 12).

MSI and OLCI observations showed a band-to-band agreement (correlation) between r = 0.44 and r = 0.94 for 'matching' bands, namely bands centred at 443, 490, 560, 665, 709 (705), 754 (740), and 779 (783) nm, with numbers in brackets referring to MSI band centres (Fig. S3). The largest disagreement between OLCI and MSI was found for the bands centred at 443 nm (r = 0.44) and 490 nm (0.77), likely due to the different radiometric sensitivity in the blue region of the visible spectrum of the two sensors, affecting the atmospheric correction procedure differently, and contributing to misaligned upstream biases. The bands in the green-red region showed the highest agreement, with an average r = 0.92.

B. MDN training

The initial phases of MDN model design and training showed a high dependence on the initialisation strategy used. When MDNs were trained from scratch and mixture weights updated with the first round of batched input data at epoch 1, they all showed a tendency to promote a single component,

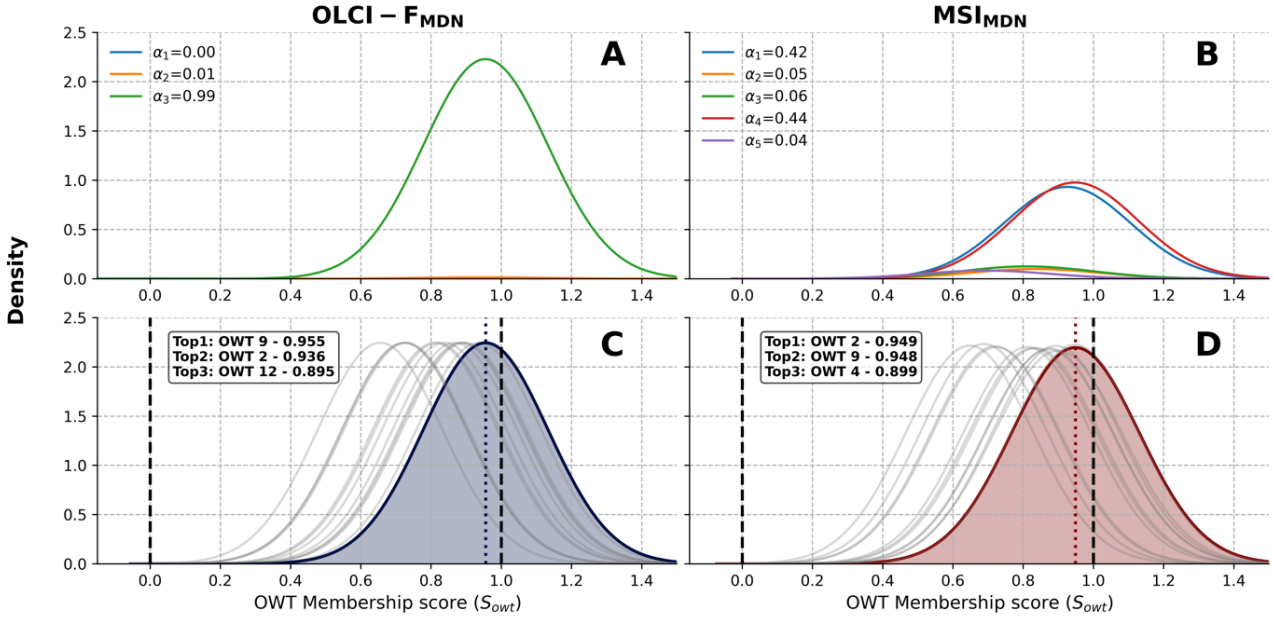

Fig. 4. Example of mixture layer components probabilities of the 'best' OLCI and MSI models (OLCI-F_{MDN} and MSI_{MDN}, respectively) for a co-located OLCI-MSI sample observation classified as Optical Water Type (OWT) 9 by the spectral angle metric (on the OLCI spectrum). **(A-B)** Probability mass of each component in the mixture in OLCI-F_{MDN} and MSI_{MDN}, respectively. **(C-D)** OWT membership score (S_{owt}) predictions for OLCI-F_{MDN} and MSI_{MDN}, respectively. The coloured curves show the (Gaussian) probability spread of the highest S_{owt} value, which is denoted by the coloured dotted lines. The black dotted lines show the valid [0, 1] S_{owt} value range. Panel A-B legends show the MDN components α_i, where i is the component number in the MDN mixture layer. Panel C-D legends show the S_{owt} values of the top-3 OWTs ranked by similarity.

irrespective of the number of components available. Instead, when updating weights with an initial solution space sweep, clustering the spectra of the reference OWT library convolved to OLCI into M components, the probability mass was no longer absorbed by a single component. This confirmed that without an appropriate initialisation strategy, MDNs may fail to fully utilise their multimodality, explore a wider solution space, and reach globally optimal solutions [63].

The hyperparameter search provided several insights. Firstly, MDNs collapsed, i.e., the gradient stopped updating due to null values populating the covariance matrices from the mixtures, when trained using a learning rate of 0.001 in conjunction with deeper networks (i.e., 5 or 7 layers), and irrespective of the type of covariance matrix used (full or diagonal). Second, the 'best' model configurations for both OLCI-F_{MDN} and MSI_{MDN} had 1,000 nodes per hidden layer and full covariances (Table II), showing that the mapping of sensor-derived spectra to S_{owt} values was best resolved at higher abstraction in the feature space, and gradients more successfully updated with multi-dimensional covariances. Third, the optimal number of mixtures for OLCI-F_{MDN} was generally lower than for MSI_{MDN}, demonstrating that the spectral characteristics of OLCI, being already highly correlated with the target $OLCI_{SA}$ vectors, assigned higher probabilities to individual components, making the use of more Gaussians redundant (Fig. 4A). On the other hand, MSI_{MDN} was prone to spread the probability mass to multiple Gaussians (Fig. 4B), likely due to the lower information carried by the input spectra and showing that the model required multiple mixture components to represent the uncertainty in mapping MSI spectra to OLCI-derived S_{owt} values. This characteristic led to multiple OWTs 'competing' for dominant class in the MSI_{MDN} models (Fig. 4D), which was something rarely observed in OLCI-F_{MDN} models (Fig. 4C). Fourth, all OLCI-F_{MDN} models reached their optimal solution by 9,000 iterations, whereas MSI_{MDN} did so around 5,500 iterations, confirming that information carried by MSI hit an earlier optimisation threshold.

Finally, the mean MdSA of the top 10 OLCI-F_{MDN} and MSI_{MDN} models ranged 0.06% – 0.09% and 1.57% – 1.68%, respectively, confirming the ability of several MDN architectures to reach a near-optimal solution. It is further worth noting that providing additional spectral features to the models, such as band ratios or further input spectra transformations like in other MDN studies [4], [33], [36] did not deliver improvements.

C. MDN prediction consistency between model instances

Training the best performing MDN configurations with different initialisations (weights) exposed the effect of initial conditions to prediction accuracy, while showing good consistency across instances, across both sensors. OLCI-F_{MDN} and MSI_{MDN} instances, as well as the OLCI-R_{MDN} instances trained on the reduced set of bands corresponding with MSI, showed similar performance patterns (Fig. 5). OLCI-F_{MDN} instances showed high accuracy with MdSA ranging 0.04% - 0.15% and near-zero bias compared to the target $OLCI_{SA}$. OLCI-R_{MDN} instances also showed high accuracy, with MdSA ranging 0.13% - 0.93% and an average bias of -0.04.

MSI_{MDN} instances showed larger absolute MdSA compared to the OLCI models, ranging 0.99% - 2.2% and with average bias of -0.05%. Across all sensors, the largest variability across instances occurred for OWTs typically associated to low biomass and clear water conditions (i.e., OWTs 3, 9, 13), which indicates that initial model weights can affect prediction

confidence for OWTs associated to optical conditions

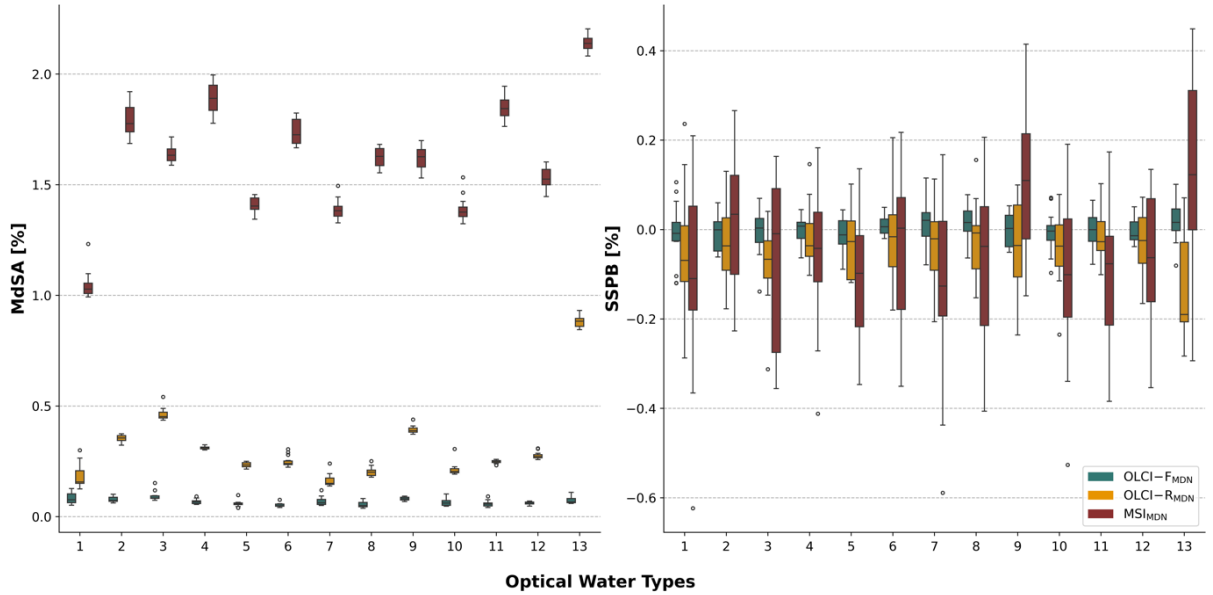

Fig. 5. Median Symmetric Accuracy (MdSA) (left panel) and Symmetric Signed Percentage Bias (SSPB) (right panel) variability across the 15 instances for OLCI-F_{MDN}, OLCI-R_{MDN}, and MSI_{MDN}, by Optical Water Types.

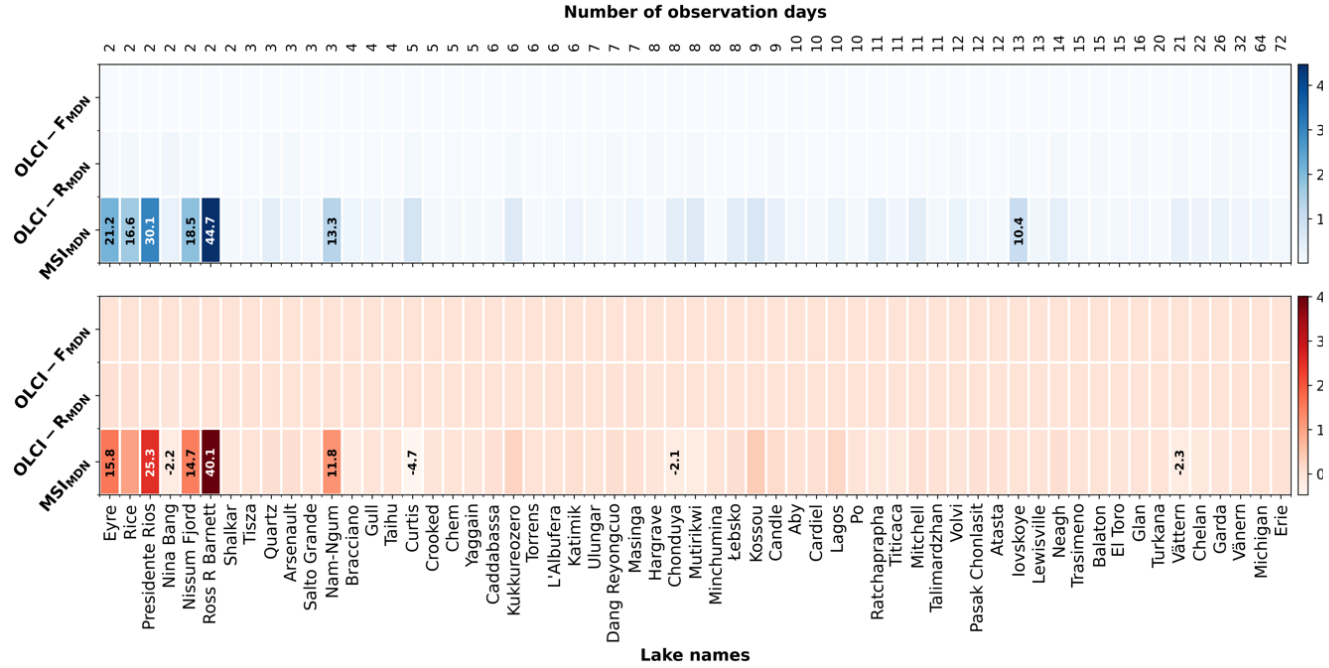

Fig. 6. MDN median performance by water body across the 15 instances for OLCI-F_{MDN}, OLCI-R_{MDN}, and MSI_{MDN}. (Top panel) Median Symmetric Accuracy (MdSA) (bottom panel) Symmetric Signed Percentage Bias (SSPB). Numbers at the top of the figure are the number of observation days per water body in the test set. Median MdSA > 10% and absolute SSPB exceeding 2% are quantified in their respective grid cells.

characterised by low NIR signal. For MSI_{MDN}, large variability also occurred for OWTs 4 and 11, both associated to high CDOM concentrations [22]. In this case, the broader bandwidths and lower radiometric sensitivity of the MSI wavebands in the blue region make these conditions inherently difficult to resolve, and likely explaining the higher variability across instances. Nevertheless, all MDNs showed strong alignment with the target $OLCI_{SA}$, and a relatively high consistency across instances.

All instances showed low errors across individual waterbodies, with low and widely consistent MdSA and SSPB for OLCI-F_{MDN} and OLCI-R_{MDN} (Fig. 6). MSI_{MDN} instances revealed relatively consistent MdSA and SSPB, with higher errors for lakes where fewer observations were available in the data set (Table I), suggesting that the errors were likely driven by epistemic uncertainty. The size of these lakes is between 61 km^2 and 9,400 km^2, and the distance from land (with respect to maximum water extent) for the available valid observations in the test set ranged from 0.5 km to 3.2 km, suggesting that land adjacency was unlikely the only driver of uncertainty. Some of these water bodies are narrow (e.g., Presidente Rios) or occasionally dry (e.g., Chonduya and Eyre), which may have impacted the quality of retrievals. For the three lakes that showed the largest MdSA and SSPB, i.e. lakes Presidente

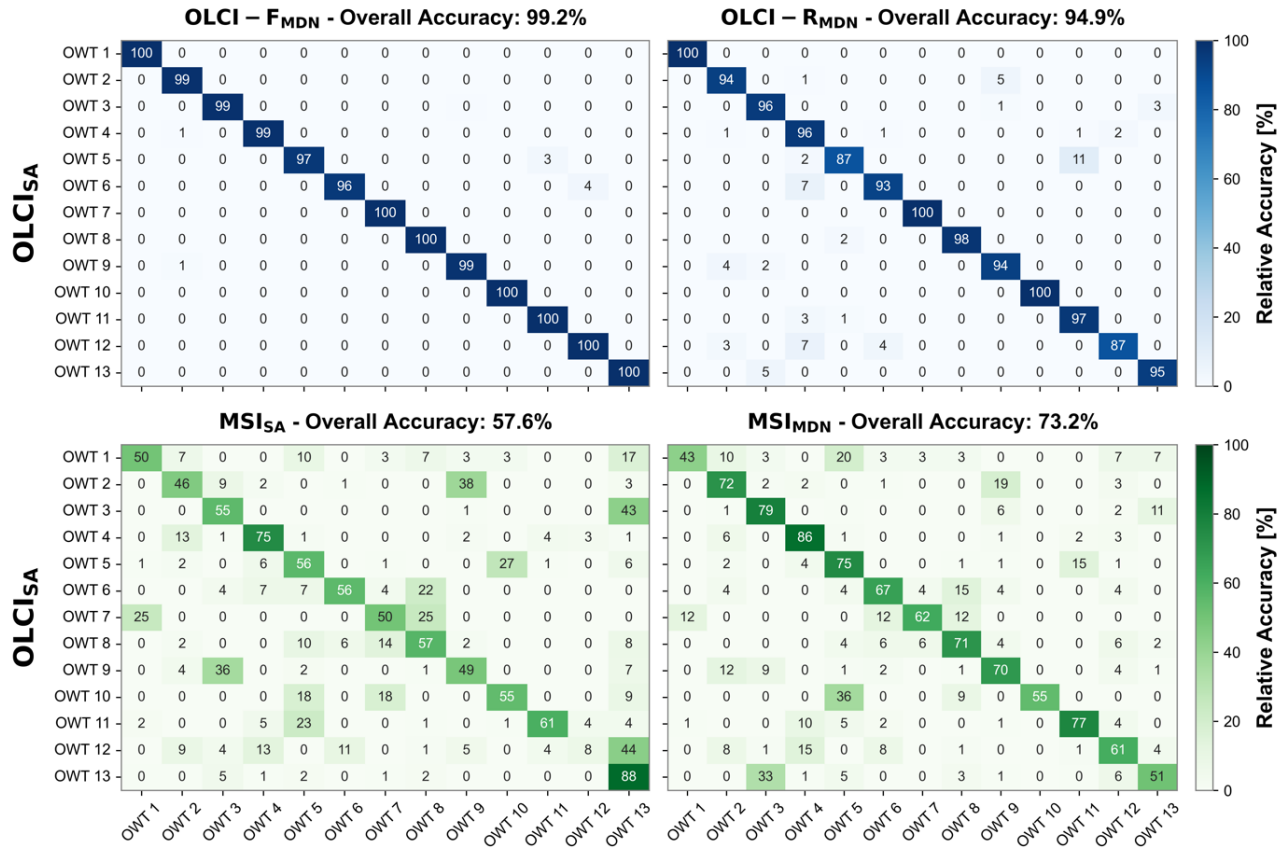

Fig. 7. Confusion matrices of the dominant Optical Water Type (OWT) classification as calculated with the spectral angle metric on standardised OLCI-derived spectra ($OLCI_{SA}$) and the ensembles (mean predictions) of the 15 instances for OLCI-F_{MDN}, OLCI-R_{MDN}, and MSI_{MDN}, as well as the spectral angle calculated on MSI (MSI_{SA}). The relative accuracy refers to the individual OWT 'hit rates', while the overall accuracy of the ensembles is shown at the top of each confusion matrix.

Rios, Ross R Barnett, and Eyre, the most frequent dominant OWTs were 13 and 5, with the former associated to clear blue waters, and the latter with a range of mixed optical conditions [22], both of which tend to be challenging to resolve for MSI (within SA space). The average S_{owt} values of the dominant OWTs across the available observations was 0.9, hinting at the possibility that while available observations were not flagged during atmospheric correction, they were likely affected by upstream biases that propagated into the model weights.

D. MDN classification performance

The OLCI-F_{MDN} ensemble demonstrated near-perfect alignment with $OLCI_{SA}$ with an overall accuracy of 99.2% when we only consider classification to the dominant OWT in terms of spectral similarity (Fig. 7). The OLCI-R_{MDN} ensemble showed an overall accuracy of 94.9%. Thus, reducing the number of OLCI bands did not limit performance, although some dominant OWT assignments marginally shifted. For example, OWT 5 was misclassified in 11% of cases as OWT 11, and OWT 12 was misclassified in 10% of cases as OWT 4 or 6. Comparing $OLCI_{SA}$ and MSI_{SA} (i.e., class assignments using the SA metric only) showed some significant divergence. Most notably, OWT 1 was correctly classified in 50% of cases, having 17% confusion with OWT 13 and 10% with OWT 5. OWT 12 was correctly classified in only 8% of cases, and instead misclassified 44% of cases as OWT 13, and 23% as OWTs 4 or 6. The MSI_{MDN} ensemble outperformed the MSI_{SA}, increasing overall accuracy from 57.6% to 73.2%. While the hit rate for OWT 1 lowered from 50% to 43%, the ensemble reduced the (unlikely) assignment to OWT 13 to 7%, increasing in the hit rate of OWT 5 to 20%. The assignment to OWT 12 increased dramatically from 8% to 61% of cases, showing that abstracting input spectral features in latent space helped to capture the optical characteristics of this OWT.

The classification accuracy of the best instances for OLCI-F_{MDN} and OLCI-R_{MDN} did not differ noticeably from the ensembles, with only up to 0.2% overall accuracy reduction (Fig. S4). The overall accuracy of the best MSI_{MDN} instance was 1.2% lower than the MSI_{MDN} ensemble, but with some non-negligible discrepancies. For example, the best MSI_{MDN} instance correctly classified OWT 13 only 34% of times, with the remainder classified in 46% of cases as OWT 3. OWT 10 was correctly classified only 27% of times and misclassified 55% of cases as OWT 5. OWTs 3 and 8, on the other hand, showed better hit rates compared to the ensemble (84% and 76%, respectively). The distribution of F1-scores assigned to each dominant OWT by the ensembles (Table S1) and best instances (Table S2) was statistically different for OLCI-F_{MDN} and MSI_{MDN} (Wilcoxon signed-rank test; $p = 0.018$ and $p = 0.006$, respectively), but statistically similar for OLCI-R_{MDN} ($p = 0.474$), revealing that OLCI-R_{MDN} likely reached its classification accuracy capacity. This showed that when looking at dominant OWTs alone, ensembles offered a better

solution than individual instances, though with marginal gains for OLCI-F_{MDN}.

Classification accuracy for the per-case top-3 ranking OWTs in terms of spectral similarity showed that OLCI-F_{MDN} and OLCI-R_{MDN} maintained a high accuracy in respecting the exact OWT classification ranking by $OLCI_{SA}$. OLCI-F_{MDN} and OLCI-R_{MDN} ensembles had an average F1-score across OWTs of 0.98 and 0.89, respectively, and 0.97 and 0.87 for the best instances, respectively (Table S1-S2). The distribution of F1 score between ensembles and best instances was found significantly different ($p = 0.005$ and $p = 0.013$, for OLCI-F_{MDN} and OLCI-R_{MDN}, respectively), confirming the ability of ensembles to provide more consistent classifications with the target $OLCI_{SA}$. MSI_{MDN}, and especially MSI_{SA}, had some challenges in producing classifications consistent with OLCI. MSI_{SA} had an average F1-score across OWTs of 0.36 when considering the exact top-3 OWT ranking order, showing a poor overall performance, but an average of 0.68 when the order of the top-3 OWTs was ignored and only inclusion in the top-3 was considered (Table S3). This showed that, while MSI_{SA} less consistently aligns with the S_{owt} values distribution of $OLCI_{SA}$, it converges relatively well on the overall optical characteristics captured by OLCI. MSI_{MDN} improved the classification accuracy both in terms of exact top-3 OWT ranking, with an average F1-score across OWTs of 0.49 and 0.47 for the ensemble and best instance, respectively, and when disregarding the order, with an average of 0.72 and 0.71 for the ensemble and best instance, respectively. In this case, the ensemble and best instance showed statistically similar F1-score distribution across OWTs ($p = 0.068$ and $p = 0.091$, for ordered and unordered top-3 ranking, respectively, showing marginal gains if using the ensemble.

S_{owt} values assigned by the OLCI-F_{MDN} and OLCI-R_{MDN} ensembles to the land adjacency types 14 and 15 closely aligned with the target $OLCI_{SA}$. The average RMSLE was 0.002 for the former and 0.012 for the latter, with an average SSPB of -0.01% and -0.08%, respectively (Fig. S5). MSI_{MDN} showed and an average RMSLE of 0.068 and an average SSPB of -0.068% for both types. Equivalent trends were observed for the best instances. This shows that the spectral resolution and radiometric sensitivity of MSI poorly adapts to the OLCI-derived land adjacency, particularly type 15.

Given that MDN ensembles offered a better solution than individual best instances for the top-3 ranking OWTs as well as dominant OWT membership, subsequent results focus only on ensembles while referring to $OLCI_{SA}$ and MSI_{SA} for comparison with SA-derived memberships for both sensors.

E. Uncertainty recalibration and decomposition

The classification uncertainty generated by the MDNs was found to be highly mis-calibrated, in line with other studies [45], [46], [47], [51]. The OLCI-F_{MDN} ensemble showed the highest Miscalibration Area (MA) with 0.488 (out of a maximum of 0.5), followed by OLCI-R_{MDN} (MA = 0.407) and MSI_{MDN} (MA = 0.24) (Table IV), with all instances largely aligning with the ensembles (Fig. S6). Once recalibrated, the MAs dropped by 93%, 87% and 80% for the OLCI-F_{MDN}, OLCI-R_{MDN}, and MSI_{MDN} ensembles, respectively, demonstrating the effectiveness of a simple linear transformation with τ to flatten the reliability curve. Before recalibration, all ensembles were extremely under-confident, i.e., their intervals were too wide, and the expected true values often fell within them (Fig. S6). In addition, the OLCI-R_{MDN} and MSI_{MDN} models showed s-shape coverage curves, which denote a skewed or heavy-tailed error distribution, resulting in the ensembles being slightly under-confident for smaller errors, and over-confident for larger errors.

TABLE IV

POST-CALIBRATION RELIABILITY OF THE OLCI-F_{MDN} AND OLCI-R_{MDN} AND MSI_{MDN} ENSEMBLES. τ IS THE TEMPERATURE-SCALING FACTOR APPLIED TO EVERY PREDICTED STANDARD DEVIATION. MA_{UNC} AND MA_{CAL} ARE THE MISCALIBRATION AREAS BEFORE AND AFTER RECALIBRATION BY APPLYING τ. COV68, COV95 GIVE THE EMPIRICAL FRACTIONS OF TRUE OBSERVATIONS THAT FALL INSIDE THE $\pm 1\,\Sigma$ ($\approx 68\%$) AND $\pm 1.96\,\Sigma$ ($\approx 95\%$) PREDICTION BANDS, RESPECTIVELY, REPORTED FOR EACH OPTICAL WATER TYPE (OWT). VALUES CLOSE TO THE NOMINAL LEVEL INDICATE WELL-CALIBRATED UNCERTAINTIES, LARGER VALUES SIGNAL UNDER-CONFIDENCE, SMALLER VALUES SIGNAL OVER-CONFIDENCE.

	OLCI-F_{MDN} (τ = 0.014)		**OLCI-R_{MDN}** (τ = 0.105)		**MSI_{MDN}** (τ = 0.339)	
	MA_{unc} 0.488	**MA_{cal}** 0.03	**MA_{unc}** 0.407	**MA_{cal}** 0.051	**MA_{unc}** 0.24	**MA_{cal}** 0.049
OWT	**Cov68**	**Cov95**	**Cov68**	**Cov95**	**Cov68**	**Cov95**
1	0.73	0.93	0.95	0.99	0.87	0.94
2	0.60	0.83	0.57	0.82	0.68	0.87
3	0.48	0.73	0.46	0.71	0.63	0.84
4	0.65	0.87	0.63	0.86	0.58	0.80
5	0.79	0.95	0.78	0.93	0.62	0.82
6	0.80	0.94	0.73	0.91	0.57	0.79
7	0.75	0.94	0.92	0.97	0.69	0.85
8	0.82	0.95	0.83	0.95	0.60	0.80
9	0.55	0.78	0.52	0.78	0.59	0.78
10	0.78	0.95	0.84	0.95	0.68	0.84
11	0.80	0.95	0.74	0.92	0.57	0.78
12	0.71	0.92	0.69	0.88	0.58	0.78
13	0.59	0.85	0.31	0.52	0.59	0.82

The coverage test for two nominal values at 68% and 95% (or $\pm$ 1 σ and 1.96 σ), revealed that the OLCI-F_{MDN} ensemble maintained a reasonable central coverage for both intervals, while OLCI-R_{MDN} and MSI_{MDN} ensembles were systematically under- or over-confident across OWTs (Table IV). OLCI-F_{MDN} had an average deviation of 2% and -6% for the 68% and 95% intervals, respectively. The largest over-confidence was for OWTs 2, 3, 9 and 13, for both intervals, while the largest under-confidence was for OWTs 6, 8 and 11 at 68%, although showing good calibration at 95%. OLCI-R_{MDN} had an average deviation of 1% and -9% for the 68% and 95% intervals, respectively, with a similar overall pattern to OLCI-F_{MDN}. The largest over-confidence was for OWTs 3 and 13 for both intervals, and the largest under-confidence for OWTs 1 and 7, for both intervals. MSI_{MDN} had an average deviation of -5% and -13% for the 68% and 95% intervals, respectively, showing a tendency to be over-confident in the estimated uncertainties. The MSI_{MDN} ensemble showed under-confidence for OWT 1 in the 68% interval, but with good calibration at 95%. Overall, the coverage test showed that after recalibration, all models produced realistic prediction

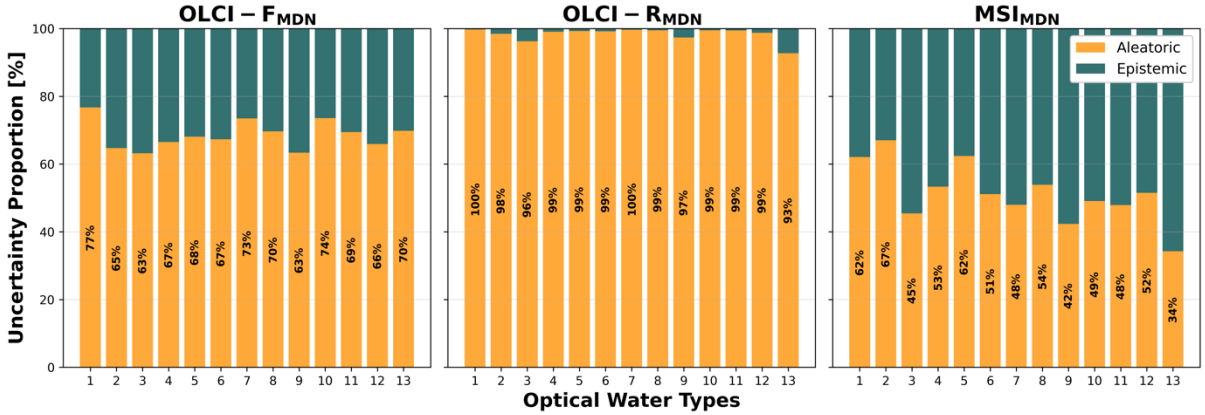

Fig. 8. Proportion of estimated recalibrated mean aleatoric and epistemic uncertainties by Optical Water Type (OWT) for the OLCI-F_{MDN}, OLCI-R_{MDN}, and MSI_{MDN} ensembles.

intervals, with OLCI-R_{MDN} and MSI_{MDN} showing moderately stronger skew, in line with the calibration curves (Fig. S6).

The distribution of decomposed uncertainties after recalibration revealed three distinct patterns across the ensembles, with OLCI-F_{MDN} dominated by aleatoric uncertainty, OLCI-R_{MDN} showing almost no epistemic uncertainty, and MSI_{MDN} revealing an overall balanced uncertainty budget (Fig. 8). The average aleatoric uncertainty for OLCI-F_{MDN} across OWTs was 68.6% (± 3.9%), with the highest epistemic uncertainty associated to OWT 3 and 9 (37% for both) and the lowest to OWT 1 and 10 (23% and 26%, respectively). OLCI-R_{MDN} had almost absent epistemic uncertainty, likely because the model maxed its predictive power, and unlikely to improve with new data. The average aleatoric uncertainty for the MSI_{MDN} ensemble was 51.4% (± 8.5%), with the highest epistemic uncertainty associated to OWT 13 (66%) and lowest to OWT 2 (33%).

The contrast between OLCI-R_{MDN} (< 1% average epistemic) and MSI_{MDN} (> 48% average epistemic), despite both models using seven wavebands centred at similar wavelengths, demonstrates that the elevated epistemic uncertainty in MSI_{MDN} is also likely associated to incomplete learning within the available feature space. This suggests that while additional training data could improve the performance of MSI_{MDN}, the lower radiometric sensitivity and broader bandwidths of MSI impose an information-theoretic ceiling below that achievable by OLCI, which the MDN expresses as epistemic uncertainty.

The uncertainty proportions across individual lakes for the OLCI-F_{MDN} and MSI_{MDN} ensembles showed a relatively consistent split between aleatoric and epistemic, with some exceptions (Fig. S7). The aleatoric uncertainty for the OLCI-F_{MDN} ensemble ranged 61-79%, confirming that the model generalised relatively well across systems and that additional data is not expected to widely shift the internal representation of the covariance between learned optical conditions. On the other hand, the aleatoric uncertainty for the MSI_{MDN} ensemble ranged 21-61%, with epistemic uncertainty averaging around 52%, and confirming both lack of representative training data and insufficient feature space in MSI.

Finally, the uncertainty of the OLCI-F_{MDN} and MSI_{MDN} ensembles relative to predictions was low for both, with an average of 0.07% ± 0.004% for the former, and 2.57% ± 0.78% for the latter. The OLCI-F_{MDN} ensemble yielded the largest mean relative uncertainty for OWT 1 with 0.08% ± 0.004%, and lowest for OWTs 3 and 9, with 0.06% ± 0.004% for both. The MSI_{MDN} ensemble showed the largest mean relative uncertainty for OWT 1, with 4.22% ± 1%, and lowest for OWTs 12 with 2% ± 0.59%. Overall, this showed that both ensembles were able to close the gap with the target $OLCI_{SA}$ by producing relatively well calibrated uncertainty ranges.

F. OWTs distribution and associated uncertainties in other water bodies

OLCI-F_{MDN} and MSI_{MDN} ensembles deployed over the hypersaline, cyanobacteria-dominated Lake Bogoria (Kenya) and the freshwater, oligotrophic to mesotrophic, Yellowstone Lake (United States) aligned with the global uncertainty patterns previously discussed, exposing sensor-specific classification biases, and revealing cross-sensor classification alignment improvements (Fig. 9 and 10). These systems were chosen to evaluate the performance of MDNs in different bio-optical conditions, climate, latitude, and typical phytoplankton assemblages.

For Lake Bogoria, OLCI-F_{MDN} provided near identical classification to the target $OLCI_{SA}$, with a hit rate, or overlap between class-proportion, of 98% for both dominant and top-3 OWTs, with the largest divergence for OWT 8, where OLCI-F_{MDN} assigned 1.2% fewer observations to this type in favour of OWT 7 and 12 (Table S4). Surprisingly, both OLCI-F_{MDN} and $OLCI_{SA}$ classified 25% of pixels as OWT 13 despite visible biomass throughout the lake (Fig. 9A). Areas classified as OWTs 13 exhibited the largest aleatoric uncertainty proportion (77%), suggesting that upstream artefacts such as land adjacency effect or surface scums interfering with atmospheric correction might be affecting the reflectance product (Fig. 9C). However, it is also possible that the confusion with clear water is due to the use of normalised R_w in the absence of clear spectral features in the NIR, causing the spectra to look relatively flat.

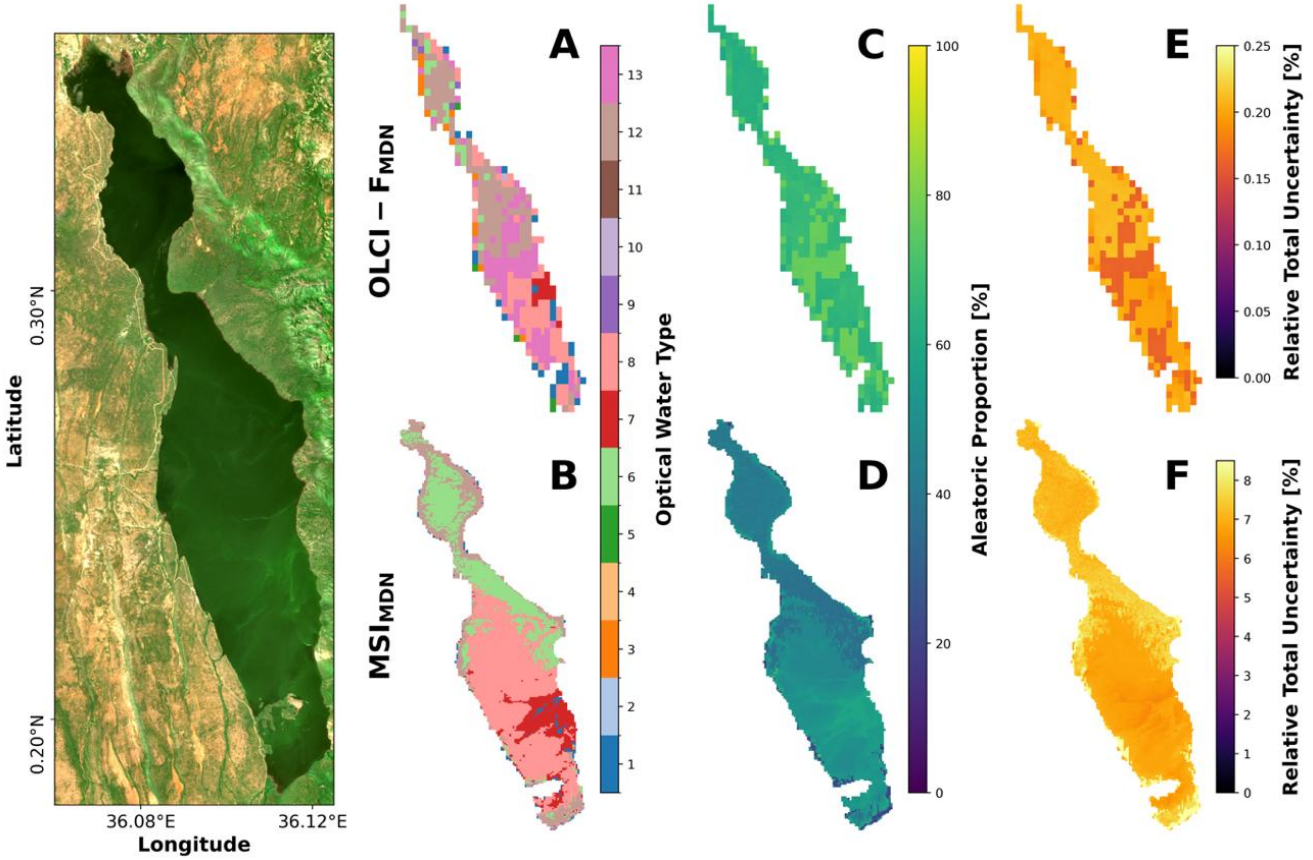

Fig. 9. Optical water type maps for Lake Bogoria (Kenya) acquired on 31 May 2021 by Sentinel-2 MSI (which was used to produce the natural colour image at far left) and Sentinel-3 OLCI. **(A-B)** Class assignments from the OLCI-F_{MDN} and MSI_{MDN} ensembles, respectively. **(C-D)** Aleatoric proportion (%) derived from the OLCI and MSI models after weighting the variance of each OWT variance by membership similarity score and summing across all 13 OWTs. **(E-F)** Corresponding relative total OWT assignment uncertainty (%) expressed as the calibrated 1-σ spread divided by the dominant membership similarity score.

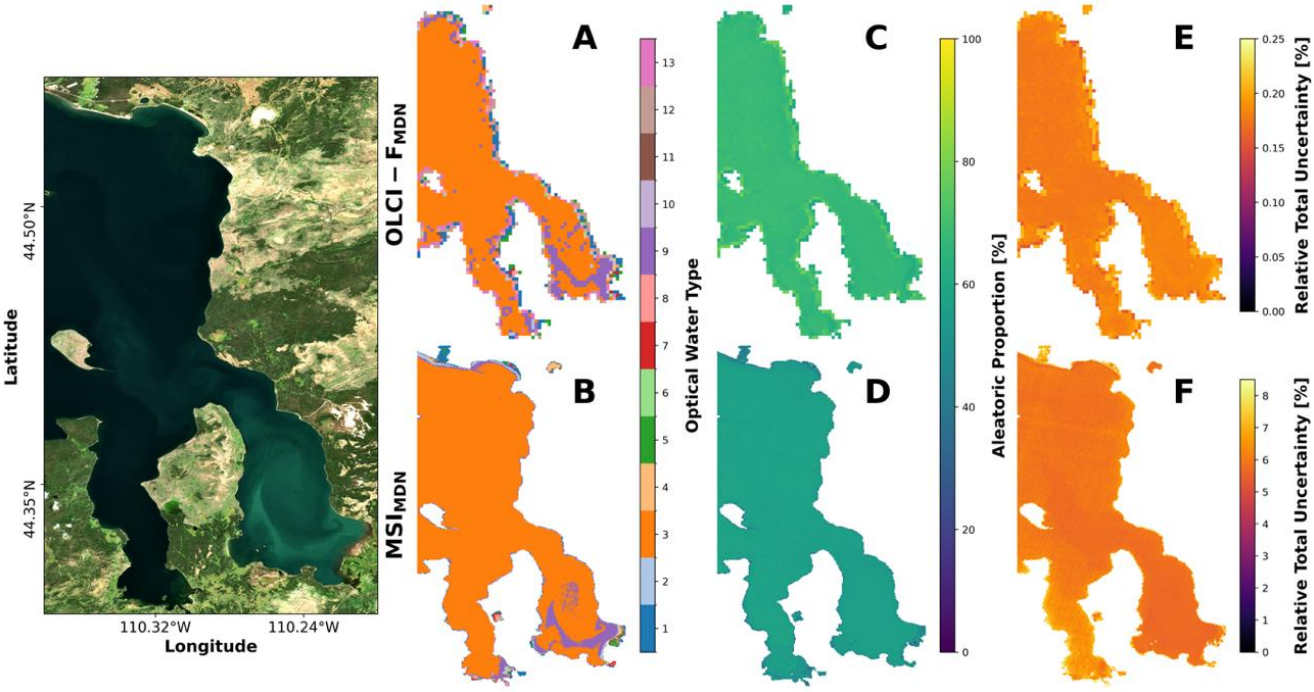

Fig. 10. Optical water type (OWT) maps for Lake Yellowstone (United States) acquired on 30 Aug 2020 by Sentinel-2 MSI (which was used to produce the natural colour image at far left) and Sentinel-3 OLCI. **(A-B)** Class assignments from the OLCI-F_{MDN} and MSI_{MDN} ensembles, respectively. **(C-D)** Aleatoric proportion (%) derived from the OLCI and MSI models after weighting the variance of each OWT variance by membership similarity score (S_{owt}) and summing across all 13 OWTs. **(E-F)** Corresponding relative total uncertainty (%) expressed as the calibrated 1-σ spread divided by the dominant S_{owt}.

MSI_{MDN} diverged markedly from MSI_{SA}, reallocating 8% and 17% more pixels to OWTs 6 and 12, respectively, and 9% and 10% fewer pixels to OWTs 7 and 8, respectively (Table S4). MSI_{MDN} also assigned 20% and 23% more pixels to OWTs 6 and 8, respectively, and 13% fewer pixels to OWT 12 compared to OLCI-F_{MDN} (Fig. 9B). While the hit rate of dominant and top-3 OWTs between MSI_{SA} and $OLCI_{SA}$ was only 37% and 44%, respectively, representing what the current classification for the two sensors looks like in the Lakes_cci, this improved to 50% and 60% between MSI_{MDN} and OLCI-F_{MDN}, respectively, showcasing the ability of the MSI_{MDN} ensemble to close the gap with OLCI. The average aleatoric fraction for the MSI_{MDN} ensemble was 45%, ranging between 71% for OWT 10 and 11% for OWT 3, confirming that the MSI_{MDN} ensemble cannot reliably resolve some low biomass optical conditions (Fig. 9D). Similarly to OLCI-F_{MDN}, areas of higher epistemic uncertainty were those associated to the highest total uncertainty relative to predicted S_{owt} values (Fig. 9F), particularly in areas where the dominant OWT classification from MSI_{MDN} diverged from MSI_{SA} (Fig. S8).

For Lake Yellowstone, the OWT classification was more consistent between the OLCI-F_{MDN} and MSI_{MDN} ensembles (Fig. 10). OLCI-F_{MDN} matched $OLCI_{SA}$ almost identically, with a hit rate of 99% for both dominant and top-3 OWTs, while MSI_{MDN} improved the hit rate compared to MSI_{SA} from 82% to 86% for the dominant, and from 68% to 84% for the top-3 OWTs. This classification improvement reduced the gap with OLCI-F_{MDN} to 14% difference for OWT 3 (from almost 20%), and 2% difference (from 6%) for OWT 9 (Table S4) and shown as a better overall spatial pattern alignment (Fig. S8). The uncertainty from the OLCI-F_{MDN} ensemble confirmed to be primarily aleatoric (average of 67%), while that from the MSI_{MDN} ensemble epistemic (average of 53%) (Fig. 10C-D). The aleatoric uncertainty in OLCI-F_{MDN} was mostly found along shorelines, hinting at (expected) biases likely linked to land adjacency, particularly where the land was brightest (Fig. 10C) and averaging 0.2%, confirming that the biases affected all OWTs equally (Fig. 10E). The uncertainty in MSI_{MDN} was homogenously spread across the lake (Fig. 10D), with the largest total uncertainty relative to predicted S_{owt} values in areas associated to OWT 13 (average 11.5%) (Fig. 10F). This aligns with what shown in Fig. 8, suggesting once again that the higher epistemic uncertainty associated to low-biomass OWTs may be due to missing diagnostic wavebands in MSI.

V. Discussion

Monitoring inland waters ideally requires producing consistent optical-biogeochemical products leveraging the complementary capabilities of multiple satellite missions, especially given that no sensor specifically designed to monitor inland water bodies currently exists. However, there are several limitations. On one hand, the high spatial resolution needed to monitor small or dendritic water bodies is primarily offered by sensors designed for land observations like MSI, that, however, do not offer appropriate diagnostic capabilities. On the other, highly capable ocean colour sensors like OLCI are limited to observing medium to large water bodies. While the ongoing transition to hyperspectral missions such as the Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) and the Copernicus Hyperspectral Imaging Mission for the Environment (CHIME) will enable a better characterisation of the global optical variability of natural waters, the most widely validated and currently operational, near-real time ocean colour sensors remain multispectral, creating a capability gap that must be bridged to maintain observational continuity. Thus, preserving the scientific value of historical multispectral observations requires developing approaches that maintain consistent water quality assessment across sensors with different capabilities, ensuring that decades of monitoring data remain useful as we transition to more capable sensors.

OWT pre-classification already underpins operational multispectral satellite water quality monitoring, partitioning the optical complexity of natural waters into discrete classes that enable appropriate algorithm selection and blending for downstream product generation. Since widely validated algorithms, such as those underpinning the Lakes_cci processing chain [29], [30] are developed and optimised for specific optical conditions defined within each type, preserving consistent OWT classification across sensors is essential for maintaining the integrity of long-term monitoring records. The core challenge is that different sensors provide varying capabilities to identify OWTs, and current approaches using metrics like spectral angle can yield inconsistent type assignments across multispectral sensors with different spectral configurations [1], likely affecting the consistency of downstream products. While probabilistic NNs have emerged in water quality monitoring as powerful alternatives to empirical and semi-analytical algorithms [4], [33], [37], [38], their accuracy typically degrades in out-of-scope applications [51], and they offer little interpretability within established bio-optical principles. Even where uncertainty decomposition is available in constituent retrieval MDNs, the resulting aleatoric and epistemic components cannot be attributed to specific optical conditions or sensor limitations, since the learned representations do not expose the bio-optical relationships governing predictions, limiting their diagnostic value for operational monitoring.

The development of probabilistic OWT classifications via MDNs addresses these limitations by transitioning to full probability distributions over possible water types while maintaining bio-optical interpretability. By grounding uncertainty quantification within OWT classifications, the decomposed uncertainty components carry direct optical meaning, whereby aleatoric uncertainty reflects upstream biases associated with specific optical conditions (e.g., atmospheric correction that fails to resolve highly scattering conditions such as surface accumulations), and epistemic uncertainty reflects sensor-specific spectral limitations in resolving certain optical conditions (more below). Reduced spectral information naturally translates to broader probability distributions rather than implausible type assignments, creating a framework where algorithm selection can adapt to

varying classification confidence levels. By augmenting established spectral angle based OWT classifications with rigorous uncertainty quantification, this approach preserves the physical interpretability and decades of algorithm development essential for water quality assessment, with the probability layer propagating directly through the inversion chain and providing an uncertainty baseline for algorithm selection and blending rather than being bolted on after constituent estimation. Spectra that fall outside the reference library are likely to show wider probability distributions, thereby exposing optical conditions that are rare or challenging to sample, but that are nevertheless observed by satellite sensors.

OWT classifications of co-located OLCI and MSI observations illustrate the primary challenge of using the spectral angle metric across different sensors, as MSI can track the same overall optical characteristics observed by OLCI, but with several type assignment inconsistencies (Fig. 2). The overall accuracy of MSI_{SA} for dominant OWTs was 57.6 % (Fig. 7), due to the inability to discern between certain classes with a reduced set of wavebands. For example, OWT 12 was almost completely absent in the MSI_{SA} classification (only 8% hit rate), which may be explained by the sensitivity of this class to the covariance between bands that are absent in MSI (wavebands centred at 620 and 681 nm, see also Lomeo *et al.* [1]), and without which certain optical conditions cannot be accurately characterised. Another example is the recurring confusion between OWT 1 and OWT 13, representing contrasting optical conditions, from at-surface accumulations to clear blue water [22]. Here, the misclassification is likely driven by atmospheric correction biases or the lower radiometric sensitivity of MSI in both the NIR and blue regions, where the most distinctive signals for biomass accumulation and clear water conditions are respectively expressed. After normalisation, the shapes of MSI-derived spectra can appear highly similar, causing the two types to present equivalent membership scores, with direct consequences for the accuracy and reliability of downstream product blending. The misalignment of the OWT classification by MSI remains when considering the top-3 ranking OWTs by similarity, where class order assigned by MSI rarely aligned with that by OLCI, with marginal improvements when the exact ranking is disregarded (Table S2-S3). Overall, these results confirm the limited ability of MSI_{SA} to discern between OWTs compared to the classification obtain with a capable ocean colour sensor like OLCI.

MDN OWT models achieve higher classification consistency between OLCI and MSI. When moving the training field to an abstraction such as S_{owt} values, MSI spectra are not required to replicate OLCI. Instead, MDNs trained on MSI spectra to map $OLCI_{SA}$ in latent space, learn to classify OWTs like OLCI while preserving the covariance structure of MSI bands, which is an aspect typically disregarded when using spectral angle (as opposed to, for example, using Mahalonobis distance used in several ocean colour applications). This abstraction fundamentally shifts the cross-sensor harmonisation challenge from attempting to reconcile differences in spectral shape to learning the underlying optical relationships that define water type membership, providing a pathway for consistent OWT classification across sensors with different spectral capabilities. $OLCI_{SA}$ is the most appropriate training target for this purpose because OLCI represents the most capable operational ocean colour sensor with global daily coverage and with nearly a decade of continue operation. The resulting OWT membership distribution derived with the spectral angle method represents the most reliable baseline against which other multispectral sensors should be evaluated, particularly in lakes where classification on normalised spectral shape provides better separation of bio-geochemically distinct water types, with subtle differences in shape at key wavebands, over Mahalonobis-type approaches where reflectance amplitude is more likely to dominate the classification.

The hyperparameter search revealed that several MDN architectures can achieve high accuracy beyond the proposed best configurations (Table II), likely because the bounded [0, 1] range of the target S_{owt} values prevented the models from moving beyond reasonable predictions, resulting in several models converging to similar solutions. For some configurations, training collapsed due to null values appearing in the covariance matrices despite the safeguard of the imputations, typically caused by 'exploding gradients' in sufficiently deep networks at large learning rates [77]. The GMM-based initialisation strategy proved crucial to avoid mode collapse [63], as without it models could produce acceptable predictions but at the cost of absent intra-component variability, and therefore, appropriate uncertainty characterisation. When properly initialised, the 15 trained instances of each model showed high consistency across OWTs and sensors (Fig. 5), confirming that starting model weights mattered less than the initialisation strategy itself. The MSI_{MDN} instances showed a higher average MdSA compared to OLCI-F_{MDN} instances, although within an acceptable range (0.99-2.2%). OLCI-R_{MDN} instances only marginally diverged from the average error and biases of OLCI-F_{MDN} despite the absence of three wavebands (i.e., 412, 510, 620 nm. This is likely because the dropped bands do not substantially affect the covariance structure of the S_{owt} vectors within the 13-OWT library, though further investigation beyond the scope of this work is warranted.

Classification performance analysis confirmed that MDN OWTs successfully learn the mapping between reflectance spectra and OWT membership scores, with ensembles outperforming individual instances across sensors and OWTs (Fig. 7, Fig. S4), consistent with what commonly described in the literature [78], [79]. The near-perfect accuracy of OLCI-F_{MDN} was expected, since the natively high correlation between OLCI-derived spectra and $OLCI_{SA}$ means the mapping task is well-conditioned, with MDNs effectively learning a smooth function over a highly structured input space. For MSI, the MSI_{MDN} ensemble improved cross-sensor OWT agreement from 58% to 73% compared to using only spectral angle alone as the OWT distance metric. This result demonstrates that learning optical relationships in latent space

is an effective harmonisation strategy across sensors with different capabilities and overcomes some of the limitations in classification on normalised spectral shape. This is primarily because MDNs were trained within the boundary of a pre-determined S_{owt} covariance matrix, inherently absorbing information on the reflectance amplitude as seen by OLCI. Where residual misalignment persisted, particularly within the top-3 OWT ranking, this was attributable to cases where small differences in probability mass between competing mixture components produced overlapping probability distributions (Fig. 4B), causing rank swaps (Fig. 4D). The OLCI-F_{MDN} and OLCI-R_{MDN} ensembles offered similar performances, showing that the bulk of spectral information, at least in relation to the reference OWT library, is carried primarily by the seven OLCI bands also provided by MSI, albeit with different band width and sensitivity.

All MDN OWT models showed severe initial miscalibration, in line with the literature [45], [46], [47], [48], and recently observed in remote sensing MDN applications [51], with post-hoc temperature scaling reducing miscalibration areas by 80-93% across MDNs. MSI_{MDN} exhibited the lowest initial miscalibration not as a reflection of better uncertainty characterisation, but likely because the wider intervals produced by reduced spectral information happened to sit closer to well-calibrated predictions, cautioning against interpreting lower miscalibration as a quality indicator in isolation. Post-calibration coverage analysis exposed systematic sensor-specific limitations, with OLCI-F_{MDN} and OLCI-R_{MDN} ensembles showing a tendency towards over-confidence for OWTs associated to low biomass conditions (e.g., OWTs 3, 13) and under-confidence for OWTs associated to high-biomass to surface accumulation conditions (e.g., OWTs 1, 7). Over-confidence for low biomass OWTs is consistent with the well-conditioned nature of the mapping task for well-represented types, where tight probability distributions persist after recalibration. Under-confidence for high biomass OWTs likely reflects the wider mixture components arising from their sparser representation in the training distribution, producing over-dispersed prediction intervals. The MSI_{MDN} ensemble showed a tendency towards over-confidence for most OWTs, consistent with the compression of the detectable spectral gradient imposed by the reduced waveband configuration of MSI. This is analogous to what Lomeo *et al.* [1] demonstrated for MODIS relative to MERIS/OLCI, whereby optically distinct conditions are collapsed into a narrower detectable range, which in this case the MDN translates into narrow, falsely confident distributions. In contrast, OWT 1 remained under-confident. Despite the high NIR reflectance amplitude associated with surface biomass accumulations being typically detectable by MSI, the MDNs likely encountered high variability in the spectral signatures associated to this type during training, possibly due to atmospheric correction failures in resolving these highly scattering conditions, causing mixture components to compete, and producing wider prediction intervals than warranted.

Collectively, the patterns described above highlight that uncertainty recalibration is fundamentally influenced by the match between spectral information content and the optical complexity of the classification task, with sensor capability determining the reliability of the resulting uncertainty estimates as much as model architecture. In practical terms, OWT-specific coverage values in Table IV provide the most direct indicator of classification reliability. Where coverage approaches the nominal 68% or 95% levels, confidence intervals can be deemed reliable for downstream algorithm selection and blending, while substantial deviations flag OWTs where uncertainty estimates should be treated with greater caution.

The use of a single linear rescaling factor (applied after uncertainty decomposition) preserved the separability between aleatoric and epistemic uncertainty while maintaining the ranking of predictions, enabling meaningful interpretation of their relative contributions across sensors and OWTs. Aleatoric uncertainty dominated the total uncertainty envelope in the OLCI-F_{MDN} ensemble, with epistemic largest for low phytoplankton biomass OWTs (e.g., OWTs 3, 9) (Fig. 8). This pattern was somewhat surprising given that these water types are widely represented in the dataset (Fig. 2) but suggests that the original OWT definitions may have captured these as single types, while multispectral sensors, each with their own upstream biases, encounter a wider range of optical conditions that lack representation in the original *in situ* observation library, effectively creating 'catch-all' categories that encompass conditions not originally sampled. This interpretation aligns with the broader challenge of defining OWTs from limited *in situ* sampling, where satellite sensors inevitably observe optical conditions beyond the scope of field measurements, like shown in Lomeo *et al.* [80] for spectra associated to cyanobacteria bloom development through varying optical conditions. Another explanation may be that when dealing with normalised R_w spectra associated to low biomass conditions, the lack of distinctive spectral features, particularly in the NIR region, causes these spectra to appear highly similar and relatively flat, with the models encountering multiple plausible OWT mappings, and inherently resulting in elevated epistemic uncertainty. For OWTs associated with high biomass conditions, either mixed in the water column or accumulating at the surface (OWTs 1, 7), uncertainty was primarily aleatoric despite their lower representation in the dataset, indicating that these optical conditions are subject to higher upstream biases from atmospheric correction challenges in dense biomass conditions [81], which appropriately propagate as aleatoric uncertainty in the MDN OWT models. The OLCI-R_{MDN} ensemble showed virtually no epistemic uncertainty across all OWTs (Fig. 8), suggesting that the model reached its predictive capacity because its waveband count and radiometric sensitivity allowed it to extract the available information fully, leaving negligible residual epistemic uncertainty. In contrast, the MSI_{MDN} ensemble exhibited the largest proportion of epistemic uncertainty, particularly for OWTs associated to low biomass conditions, indicating the inability of the model to fully resolve the OWT membership covariance structure when

these water types are dominant. While this result aligns with OLCI-F_{MDN}, though at different magnitude, it likely occurs for different reasons, including the lower radiometric sensitivity of the available wavebands in the blue and NIR region, as well as the absence of wavebands centred at 412 nm and 510 nm, preventing MSI to fully resolve these optical conditions.

The deployment of the OLCI-F_{MDN} and MSI_{MDN} ensembles in lakes outside the training set demonstrates that the proposed MDN OWT models provide spatially appropriate uncertainty characterisation and showcasing improved cross-sensor classification alignment (Fig. 9 and 10). The OWT classification of the OLCI-F_{MDN} ensemble in Lake Bogoria exposed systematic upstream biases, likely linked to atmospheric correction, manifesting as unlikely assignments of OWT 13 to areas visibly dominated by biomass (Fig. 9). As a model trained to reproduce $OLCI_{SA}$, OLCI-F_{MDN} inherited these classification biases, with their origin identified through the uncertainty decomposition, whereby areas assigned to OWT 13 also exhibited the largest aleatoric fraction (Fig. 9C), delivering a diagnostic capability previously absent from spectral angle-based classifications. The OWT classification of the MSI_{MDN} ensemble highlighted the improved alignment with OLCI, both for dominant and top-3 OWTs (Table S4), while not showing the same atmospheric correction issues as OLCI (Fig. 9C-D) and exhibiting predominantly epistemic uncertainty across the lake, in line with previous results (Fig. 8; Fig. S7), In the less optically dynamic conditions of Lake Yellowstone, MSI_{MDN} showed an improved overall agreement with OLCI for the top-3 OWTs (Table S4) but a similar dominant OWT distribution, showing that in less productive waters MSI_{MDN} is likely to provide more appropriate OWT classifications compared to MSI_{SA} (Fig. S8). Together, these results confirm the suitability of MDN OWT models for integration into operational processing chains, given their ability to deliver detailed uncertainty characterisation from atmospherically corrected satellite-derived spectra alone, without recourse to ancillary data.

Overall, two important clarifications are warranted regarding the interpretation of the estimated uncertainties. First, while epistemic uncertainty is conventionally associated with insufficient training data, MDNs applied to reflectance spectra introduce an additional source. When the available spectral information is insufficient to unambiguously resolve certain optical conditions, multiple mixture components naturally compete for probability mass, elevating the variance of the component means (9) regardless of training set size. This behaviour is an inherent and expected property of MDNs, reflecting irreducible spectral ambiguity imposed by sensor characteristics, meaning that a portion of the epistemic uncertainty estimated by the MDNs, and especially MSI_{MDN}, originates from information-theoretic constraints of the sensor itself rather than from gaps in the training distribution, and is therefore unlikely to diminish with additional training data alone. Second, reduced spectral information naturally translate to wider Gaussian components when optical conditions are more challenging to resolve (Fig. 4B), producing substantially larger total relative uncertainty envelopes in MSI_{MDN} compared to OLCI-F_{MDN}. Interpreting total relative uncertainty across sensors therefore requires accounting for this sensor-driven baseline difference, whereby uncertainty magnitudes are most meaningfully evaluated within each sensor rather than across sensors, while cross-sensor comparisons are better served by the spatial structure of uncertainties and the consistency of dominant and top-3 ranking OWT assignments. Where MSI_{MDN} produces spatially coherent classifications alongside wide prediction intervals, the appropriate interpretation is that OWT assignments are plausible but carry sensor-driven classification uncertainty that can be used to modulate the confidence assigned to downstream algorithm selection and blending. This is illustrated in Fig. 9A-B, where the classification maps for Lake Bogoria show broad spatial coherence between MSI_{MDN} and OLCI-F_{MDN}, yet with systematic divergence for OWTs associated with various mixed water column cyanobacteria conditions (Table S4 - OWTs 6, 8, 12), for which diagnostic spectral features fall in wavebands absent from MSI, and that cannot be fully resolved regardless of classification confidence.

In our view, future work should prioritise six development streams. First, it may be beneficial to explore alternative ways to model the conditional probabilities in the mixture layer by using other objectives that do not represent probabilities solely as normal distributions and preventing values to move beyond the expected bounded [0, 1] range. Second, it would be useful to train MDN OWT models with a wider set of definitions beyond the current *in situ* libraries, especially considering that certain optical conditions may not be successfully captured during a single sampling campaign. Third, it would be worth exploring ways to integrate physics informed NNs within MDN OWTs, combining the interpretability of radiative transfer models with the flexibility of machine learning, potentially resolving the persistent challenge of separating atmospheric and aquatic contributions to observed spectra. Fourth, it seems appropriate to expose these models to other atmospheric correction algorithms, perhaps leveraging their multimodal nature and create atmospheric correction-specific branches of the MDN OWT models to deal with the range of possible satellite-derived R_w spectra as obtained with different atmospheric corrections. Fifth, it will be beneficial to extend the investigation to other land sensors like the Landsat series to allow expanding the available temporal window, and small, high spatial resolution satellite constellations like PlanetScope to enhance monitoring of small to very small water bodies. Finally, it will be useful to explore ways to determine the portion of epistemic uncertainty deriving from the lack of (sensor-specific) spectral information, which is unlikely to be resolved with additional data, and highlight what samples, instead, should be expected to improve model performance if fed to the models.

VI. Conclusion

This study demonstrates that OWT classification via MDNs fundamentally reframes cross-sensor harmonisation from a

spectral matching problem to one of learning transferable optical relationships, providing previously unavailable quantified OWT classification uncertainties essential for operational water quality monitoring across multispectral satellite sensors. The MDNs ensembles achieved near-perfect reproduction of OLCI-based classifications using the spectral angle metric (> 99% accuracy) while improving MSI classification alignment with OLCI from 58% to 73%, showing that a sensor with reduced spectral capabilities can contribute meaningfully to monitoring programs when inherent spectral limitations are statistically characterised. The decomposition of uncertainties into aleatoric and epistemic components, and subsequent recalibration through temperature scaling, exposed distinct sensor-specific constraints, with OLCI limited primarily by upstream biases (69% aleatoric), and MSI by insufficient training representation and limited spectral information (52% epistemic), providing actionable diagnostics for targeting improvements. The spatial coherence of uncertainty patterns across contrasting systems, including examples outside the training set, validates that MDN OWTs learn generalisable optical relationships, essential for global-scale deployment where training data cannot encompass all possible conditions.

Critically, this approach addresses fundamental limitations of current MDN applications in ocean colour that directly estimate biogeochemical parameters without preserving interpretable connections to established bio-optical principles. While MDNs have demonstrated superior statistical performance for constituent retrieval [4], [34], [36], their learned representations do not expose the bio-optical relationships governing predictions, limiting interpretation. By constraining MDNs to predict OWT membership scores rather than constituent concentrations directly, our framework maintains the mechanistic understanding essential for operational monitoring while leveraging probabilistic NNs to handle the multimodal nature of the inverse problem. This directly links to the ability to trace prediction failures to specific optical conditions or sensor limitations, rather than untraceable network weights, enabling targeted improvements that would be impossible with end-to-end learning approaches.

Ultimately, this work establishes that acknowledging and quantifying OWT classification uncertainty provides the foundation for robust multi-mission water quality monitoring, enabling uncertainty-aware processing chains where algorithm selection and blending weights can adapt dynamically to classification confidence, while preserving decades of algorithm development and validation. As satellite constellations expand and sensor capabilities evolve, probabilistic frameworks that explicitly model the relationship between spectral information and OWT classification confidence may become essential for maintaining consistent long-term records while adapting to technological advances. The transition to MDN OWT classifications represents a fundamental shift in how the integration of diverse Earth observation data may be approached for understanding and managing aquatic ecosystems under accelerating environmental change.

ACKNOWLEDGMENT

The authors acknowledge funding to Davide Lomeo from the Natural Environmental Research Council (NERC) through the London NERC DTP (NE/S007229/1) and the Natural Environment Research Council Earth Observation Data Analysis and Artificial-Intelligence Service (NEODAAS) for providing satellite data.

REFERENCES

[1] D. Lomeo, S. G. H. Simis, N. Selmes, A. D. Jungblut, and E. J. Tebbs, ‘Colour-informed ecoregion analysis highlights a satellite capability gap for spatially and temporally consistent freshwater cyanobacteria monitoring’, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 228, pp. 323–339, Oct. 2025, doi: 10.1016/j.isprsjprs.2025.07.030.

[2] Y. Zhang, J. Pulliainen, S. Koponen, and M. Hallikainen, ‘Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data’, *Remote Sensing of Environment*, vol. 81, no. 2–3, pp. 327–336, Aug. 2002, doi: 10.1016/S0034-4257(02)00009-3.

[3] M. W. Matthews, ‘A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters’, *International Journal of Remote Sensing*, vol. 32, no. 21, pp. 6855–6899, Nov. 2011, doi: 10.1080/01431161.2010.512947.

[4] N. Pahlevan *et al.*, ‘Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach’, *Remote Sensing of Environment*, vol. 240, p. 111604, Apr. 2020, doi: 10.1016/j.rse.2019.111604.

[5] A. G. Dekker, *Detection of optical water quality parameters for entrophic waters by high resolution remote sensing*. Amsterdam: Free University, 1993.

[6] D. Jupp, J. Kirk, and G. Harris, ‘Detection, identification and mapping of cyanobacteria — Using remote sensing to measure the optical quality of turbid inland waters’, *Mar. Freshwater Res.*, vol. 45, no. 5, p. 801, 1994, doi: 10.1071/MF9940801.

[7] H. J. Gons, ‘Effect of a waveband shift on chlorophyll retrieval from MERIS imagery of inland and coastal waters’, *Journal of Plankton Research*, vol. 27, no. 1, pp. 125–127, Sep. 2004, doi: 10.1093/plankt/fbh151.

[8] J. Gower, S. King, G. Borstad, and L. Brown, ‘Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer’, *International Journal of Remote Sensing*, vol. 26, no. 9, pp. 2005–2012, May 2005, doi: 10.1080/01431160500075857.

[9] G. Dall’Olmo, A. A. Gitelson, D. C. Rundquist, B. Leavitt, T. Barrow, and J. C. Holz, ‘Assessing the potential of SeaWiFS and MODIS for estimating chlorophyll concentration in turbid productive waters using red and near-infrared bands’, *Remote Sensing of Environment*, vol. 96, no. 2, pp. 176–187, May 2005, doi: 10.1016/j.rse.2005.02.007.

[10] S. G. H. Simis, A. Ruiz-Verdú, J. A. Domínguez-Gómez, R. Peña-Martinez, S. W. M. Peters, and H. J. Gons, ‘Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass’, *Remote*

Sensing of Environment, vol. 106, no. 4, pp. 414–427, Feb. 2007, doi: 10.1016/j.rse.2006.09.008.

[11] A. A. Gilerson *et al.*, 'Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands', 2010.

[12] M. W. Matthews, S. Bernard, and L. Robertson, 'An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters', *Remote Sensing of Environment*, vol. 124, pp. 637–652, Sep. 2012, doi: 10.1016/j.rse.2012.05.032.

[13] S. Mishra, D. R. Mishra, Z. Lee, and C. S. Tucker, 'Quantifying cyanobacterial phycocyanin concentration in turbid productive waters: A quasi-analytical approach', *Remote Sensing of Environment*, vol. 133, pp. 141–151, Jun. 2013, doi: 10.1016/j.rse.2013.02.004.

[14] M. W. Matthews and D. Odermatt, 'Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters', *Remote Sensing of Environment*, vol. 156, pp. 374–382, Jan. 2015, doi: 10.1016/j.rse.2014.10.010.

[15] G. Liu *et al.*, 'A Four-Band Semi-Analytical Model for Estimating Phycocyanin in Inland Waters From Simulated MERIS and OLCI Data', *IEEE Trans. Geosci. Remote Sensing*, vol. 56, no. 3, pp. 1374–1385, Mar. 2018, doi: 10.1109/TGRS.2017.2761996.

[16] C. E. Binding, D. G. Bowers, and E. G. Mitchelson-Jacob, 'Estimating suspended sediment concentrations from ocean colour measurements in moderately turbid waters; the impact of variable particle scattering properties', *Remote Sensing of Environment*, vol. 94, no. 3, pp. 373–383, Feb. 2005, doi: 10.1016/j.rse.2004.11.002.

[17] B. Nechad, K. G. Ruddick, and Y. Park, 'Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters', *Remote Sensing of Environment*, vol. 114, no. 4, pp. 854–866, Apr. 2010, doi: 10.1016/j.rse.2009.11.022.

[18] A. I. Dogliotti, K. G. Ruddick, B. Nechad, D. Doxaran, and E. Knaeps, 'A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters', *Remote Sensing of Environment*, vol. 156, pp. 157–168, Jan. 2015, doi: 10.1016/j.rse.2014.09.020.

[19] M. E. Smith, L. Robertson Lain, and S. Bernard, 'An optimized Chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters', *Remote Sensing of Environment*, vol. 215, pp. 217–227, Sep. 2018, doi: 10.1016/j.rse.2018.06.002.

[20] A. Morel and L. Prieur, 'Analysis of variations in ocean color1', *Limnology & Oceanography*, vol. 22, no. 4, pp. 709–722, Jul. 1977, doi: 10.4319/lo.1977.22.4.0709.

[21] T. S. Moore, J. W. Campbell, and Hui Feng, 'A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms', *IEEE Trans. Geosci. Remote Sensing*, vol. 39, no. 8, pp. 1764–1776, Aug. 2001, doi: 10.1109/36.942555.

[22] E. Spyrakos *et al.*, 'Optical types of inland and coastal waters: Optical types of inland and coastal waters', *Limnol. Oceanogr.*, vol. 63, no. 2, pp. 846–870, Mar. 2018, doi: 10.1002/lno.10674.

[23] K. Uudeberg, I. Ansko, G. Põru, A. Ansper, and A. Reinart, 'Using Optical Water Types to Monitor Changes in Optically Complex Inland and Coastal Waters', *Remote Sensing*, vol. 11, no. 19, p. 2297, Oct. 2019, doi: 10.3390/rs11192297.

[24] J. T. Kent and K. V. Mardia, 'Spatial classification using fuzzy membership models', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 5, pp. 659–671, Sep. 1988, doi: 10.1109/34.6774.

[25] F. Wang, 'Fuzzy classification of remote sensing images', *IEEE Trans. Geosci. Remote Sensing*, vol. 28, no. 2, pp. 194–201, Mar. 1990, doi: 10.1109/36.46698.

[26] T. S. Moore, J. W. Campbell, and M. D. Dowell, 'A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product', *Remote Sensing of Environment*, vol. 113, no. 11, pp. 2424–2430, Nov. 2009, doi: 10.1016/j.rse.2009.07.016.

[27] T. S. Moore, M. D. Dowell, S. Bradt, and A. Ruiz Verdu, 'An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters', *Remote Sensing of Environment*, vol. 143, pp. 97–111, Mar. 2014, doi: 10.1016/j.rse.2013.11.021.

[28] F. A. Kruse *et al.*, 'The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data', in *AIP Conference Proceedings*, Pasadena, California (USA): AIP, 1993, pp. 192–201. doi: 10.1063/1.44433.

[29] X. Liu *et al.*, 'Retrieval of Chlorophyll-a concentration and associated product uncertainty in optically diverse lakes and reservoirs', *Remote Sensing of Environment*, vol. 267, p. 112710, Dec. 2021, doi: 10.1016/j.rse.2021.112710.

[30] S. Simis *et al.*, 'D2.2: Algorithm Theoretical Basis Document (ATBD)', 2022.

[31] IOCCG, *Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters*, vol. 3. in Reports of the International Ocean-Colour Coordinating Group, vol. 3. Dartmound, Canada, 2000.

[32] IOCCG, *Observation of Harmful Algal Blooms with Ocean Colour Radiometry.* Dartmouth, Canada: International Ocean Colour Coordinating Group (IOCCG), 2021. doi: 10.25607/OBP-1042.

[33] B. Smith *et al.*, 'A Chlorophyll-a Algorithm for Landsat-8 Based on Mixture Density Networks', *Front. Remote Sens.*, vol. 1, p. 623678, Feb. 2021, doi: 10.3389/frsen.2020.623678.

[34] R. E. O'Shea *et al.*, 'Advancing cyanobacteria biomass estimation from hyperspectral observations: Demonstrations with HICO and PRISMA imagery', *Remote Sensing of Environment*, vol. 266, p. 112693, Dec. 2021, doi: 10.1016/j.rse.2021.112693.

[35] M. Werther *et al.*, 'A Bayesian approach for remote sensing of chlorophyll-a and associated retrieval uncertainty in oligotrophic and mesotrophic lakes', *Remote Sensing of Environment*, vol. 283, p. 113295, Dec. 2022, doi: 10.1016/j.rse.2022.113295.

[36] R. E. O'Shea *et al.*, 'A hyperspectral inversion framework for estimating absorbing inherent optical properties and biogeochemical parameters in inland and coastal waters', *Remote Sensing of Environment*, vol. 295, p. 113706, Sep. 2023, doi: 10.1016/j.rse.2023.113706.

[37] A. M. Saranathan, M. Werther, S. V. Balasubramanian, D. Odermatt, and N. Pahlevan, 'Assessment of advanced neural networks for the dual estimation of water quality indicators and their uncertainties', *Front. Remote Sens.*, vol. 5, Jul. 2024, doi: 10.3389/frsen.2024.1383147.

[38] S. V. Balasubramanian *et al.*, ‘Mixture density networks for re-constructing historical ocean-color products over inland and coastal waters: demonstration and validation’, *Front. Remote Sens.*, vol. 6, p. 1488565, Feb. 2025, doi: 10.3389/frsen.2025.1488565.

[39] C. M. Bishop, ‘Mixture Density Networks’, 1994. [Online]. Available: https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf

[40] Y. Gal and Z. Ghahramani, ‘Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning’, in *Proceedings of the 33 rd International Conference on Machine Learning*, New York, NY, USA, 2016.

[41] A. Kristiadi, S. Däubener, and A. Fischer, ‘Predictive Uncertainty Quantification with Compound Density Networks’, Dec. 29, 2019, *arXiv*: arXiv:1902.01080. doi: 10.48550/arXiv.1902.01080.

[42] P. C. Gray, E. Boss, J. X. Prochaska, H. Kerner, C. B. Demeaux, and Y. Lehahn, ‘The Promise and Pitfalls of Machine Learning in Ocean Remote Sensing’, *Oceanography*, vol. 37, no. 3, pp. 52–63, 2024, doi: https://doi.org/10.5670/oceanog.2024.511.

[43] A. Kendall and Y. Gal, ‘What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?’, 2017.

[44] E. Hüllermeier and W. Waegeman, ‘Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods’, *Mach Learn*, vol. 110, no. 3, pp. 457–506, Mar. 2021, doi: 10.1007/s10994-021-05946-3.

[45] B. Lakshminarayanan, A. Pritzel, and C. Blundell, ‘Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles’, Nov. 04, 2017, *arXiv*: arXiv:1612.01474. doi: 10.48550/arXiv.1612.01474.

[46] V. Kuleshov and S. Deshpande, ‘Calibrated and Sharp Uncertainties in Deep Learning via Density Estimation’, in *Proceedings of the 39 th International Conference on Machine Learning*, Baltimore, Maryland, 2022.

[47] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, ‘On Calibration of Modern Neural Networks’, Aug. 03, 2017, *arXiv*: arXiv:1706.04599. doi: 10.48550/arXiv.1706.04599.

[48] V. Kuleshov, N. Fenner, and S. Ermon, ‘Accurate Uncertainties for Deep Learning Using Calibrated Regression’, Jul. 01, 2018, *arXiv*: arXiv:1807.00263. doi: 10.48550/arXiv.1807.00263.

[49] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier, ‘Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference’, Nov. 18, 2019, *arXiv*: arXiv:1909.13550. doi: 10.48550/arXiv.1909.13550.

[50] A. Niculescu-Mizil and R. Caruana, ‘Predicting good probabilities with supervised learning’, in *Proceedings of the 22nd international conference on Machine learning - ICML ’05*, Bonn, Germany: ACM Press, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.

[51] M. Werther *et al.*, ‘On the generalization ability of probabilistic neural networks for hyperspectral remote sensing of absorption properties across optically complex waters’, *Remote Sensing of Environment*, vol. 328, p. 114820, Oct. 2025, doi: 10.1016/j.rse.2025.114820.

[52] J. C. Platt, ‘Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.’, in *Advances in Large Margin Classifiers*, vol. 3, in 10, vol. 3. , MIT Press, 1999, pp. 61–74.

[53] S. G. H. Simis *et al.*, ‘Lakes Climate Change Inititiave Algorithm Theoretical Basis Document (ATBD), version 3.0.0’, 2025, doi: 10.5281/zenodo.18234692.

[54] S. G. H. Simis *et al.*, ‘Lakes Climate Change Initiative Product Validation and Intercomparison Report (PVIR), version 3.0.0’, 2025, doi: 10.5281/zenodo.18234692.

[55] S. W. Bailey and P. J. Werdell, ‘A multi-sensor approach for the on-orbit validation of ocean color satellite data products’, *Remote Sensing of Environment*, vol. 102, no. 1–2, pp. 12–23, May 2006, doi: 10.1016/j.rse.2006.01.015.

[56] EUMETSAT, ‘Recommendations for Sentinel-3 OLCI Ocean Colour product validations in comparison with in situ measurements – Matchup Protocols’, vol. EUM/SEN3/DOC/19/1092968, no. v8B, p. 12, Jan. 2022.

[57] R. E. Carlson, ‘A trophic state index for lakes’, *Limnology & Oceanography*, vol. 22, no. 2, pp. 361–369, Mar. 1977, doi: 10.4319/lo.1977.22.2.0361.

[58] D. Jiang *et al.*, ‘A data-driven approach to flag land-affected signals in satellite derived water quality from small lakes’, *International Journal of Applied Earth Observation and Geoinformation*, vol. 117, p. 103188, Mar. 2023, doi: 10.1016/j.jag.2023.103188.

[59] P. Blanchard, D. J. Higham, and N. J. Higham, ‘Accurately computing the log-sum-exp and softmax functions’, *IMA Journal of Numerical Analysis*, vol. 41, no. 4, pp. 2311–2330, Oct. 2021, doi: 10.1093/imanum/draa038.

[60] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. in Wiley series in probability and mathematical statistics. New York: Wiley, 1987.

[61] N. Pahlevan *et al.*, ‘Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3’, *Remote Sensing of Environment*, vol. 270, p. 112860, Mar. 2022, doi: 10.1016/j.rse.2021.112860.

[62] O. Makansi, E. Ilg, Ö. Cicek, and T. Brox, ‘Overcoming Limitations of Mixture Density Networks: A Sampling and Fitting Framework for Multimodal Future Prediction’, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7137–7146. doi: 10.1109/CVPR.2019.00731.

[63] E. Shireman, D. Steinley, and M. J. Brusco, ‘Examining the effect of initialization strategies on the performance of Gaussian mixture modeling’, *Behav Res*, vol. 49, no. 1, pp. 282–293, Feb. 2017, doi: 10.3758/s13428-015-0697-6.

[64] A. P. Dempster, N. M. Laird, and D. B. Rubin, ‘Maximum Likelihood from Incomplete Data Via the *EM* Algorithm’, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 39, no. 1, pp. 1–22, Sep. 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[65] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, ‘The effectiveness of lloyd-type methods for the k-means problem’, *Journal of the ACM*, vol. 59, no. 6, 2012.

[66] X. Liu, M. Warren, N. Selmes, and S. G. H. Simis, ‘Quantifying decadal stability of lake reflectance and chlorophyll-a from medium-resolution ocean color sensors’, *Remote Sensing of Environment*, vol. 306, p. 114120, May 2024, doi: 10.1016/j.rse.2024.114120.

[67] A. Amidi and S. Amidi, *A detailed example of how to use data generators with Keras*. (2017). [Online]. Available: https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly

[68] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, and E. I. Zacharaki, ‘EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation’, Jul. 19, 2017, *arXiv*: arXiv:1707.06017. doi: 10.48550/arXiv.1707.06017.

[69] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, ‘Bayesian Model Averaging: A Tutorial’, vol. 14, no. 4, pp. 82–417, 1999.

[70] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, ‘Using Stacking to Average Bayesian Predictive Distributions (with Discussion)’, *Bayesian Anal.*, vol. 13, no. 3, Sep. 2018, doi: 10.1214/17-ba1091.

[71] R. Rahaman and A. H. Thiery, ‘Uncertainty Quantification and Deep Ensembles’, Nov. 02, 2021, *arXiv*: arXiv:2007.08792. doi: 10.48550/arXiv.2007.08792.

[72] M. Valdenegro-Toro and D. S. Mori, ‘A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement’, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 1508–1516. doi: 10.1109/CVPRW56347.2022.00157.

[73] S. Stoudt, A. Pintar, and A. Possolo, ‘Coverage Intervals’, *J. RES. NATL. INST. STAN.*, vol. 126, Mar. 2021, doi: 10.6028/jres.126.004.

[74] M. H. Rasmussen, C. Duan, H. J. Kulik, and J. H. Jensen, ‘Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets’, *J Cheminform*, vol. 15, no. 1, Dec. 2023, doi: 10.1186/s13321-023-00790-0.

[75] S. K. Morley, T. V. Brito, and D. T. Welling, ‘Measures of Model Performance Based On the Log Accuracy Ratio’, *Space Weather*, vol. 16, no. 1, pp. 69–88, Jan. 2018, doi: 10.1002/2017SW001669.

[76] K. Zolfaghari, N. Pahlevan, S. G. H. Simis, R. E. O’Shea, and C. R. Duguay, ‘Sensitivity of remotely sensed pigment concentration via Mixture Density Networks (MDNs) to uncertainties from atmospheric correction’, *Journal of Great Lakes Research*, vol. 49, no. 2, pp. 341–356, Apr. 2023, doi: 10.1016/j.jglr.2022.12.010.

[77] G. Philipp, D. Song, and J. G. Carbonell, ‘The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions’, Apr. 06, 2018, *arXiv*: arXiv:1712.05577. doi: 10.48550/arXiv.1712.05577.

[78] T. G. Dietterich, ‘Ensemble Methods in Machine Learning’, in *Multiple Classifier Systems*, vol. 1857, in Lecture Notes in Computer Science, vol. 1857. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[79] IOCCG, *Uncertainties in Ocean Colour Remote Sensing*, vol. 18. in Reports of the International Ocean-Colour Coordinating Group, vol. 18. Dartmouth, Canada, 2019.

[80] D. Lomeo *et al.*, ‘A novel cyanobacteria occurrence index derived from optical water types in a tropical lake’, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 223, pp. 58–77, May 2025, doi: 10.1016/j.isprsjprs.2025.03.006.

[81] IOCCG, *Atmospheric Correction for Remotely-Sensed Ocean-Colour Products.*, vol. 10. in Reports of the International Ocean-Colour Coordinating Group, vol. 10. Dartmouth, Canada, 2010.