Appendices

A SHIFTS DATASET GENERAL DATASHEET

Here we describe the motivation, uses, distribution as well as the maintenance and support plan for the Shifts 2.0 Dataset in the *datasheet for datasets* format (Gebru et al., 2018). The details of the composition, collection and pre-prossessing of each component dataset are provided in appendices C-D.

Motivation The primary goal for the creation of the Shifts 2.0 Dataset was the evaluation of uncertainty quantification models and robustness to distributional shift on industrial and medical tasks of large practical and societal importance. These datasets span multiple modalities and feature real examples of distributional shift. Making these datasets available allows models' robust generalisation and uncertainty quality to be assessed - something not possible with standard in-domain benchmarks. Furthermore, by construction a dataset using real medical or industrial tasks, the any insights reached can be directly applied without the need for adaptation. This is an important feature, as most novel ML methods fail at the stage of adaptation and scaling to actual applications.

Distribution It is our intention that the Shifts 2.0 dataset be freely available for research purposes. All the code will be makde available under an open-source Apache 2.0 licence.

The Shifts cargo vessel power estimation datasets is distributed under an open-source CC BY NC SA 4.0 license. The training and in-domain/shifted devevelopment sets, both with real and synthetic targets, will be freely distributed via the Zenodo platform. The evaluation sets will not be released, but will be hosted on permanent leaderboards on the Grand-Challenge platform. Should the leaderboards close for any reason, the evaluation sets will be similarly released via Zenodo. The reason for keeping the evaluation sets private is to ensure a truly clean 'out-of-domain generalisation scenario' and avoid any possible, even unintentional, data leakage.

The MS lesion segmentation dataset has a more complex structure. Part of the dataset (ISBI train set and PubMRI) is shared under a permissive CC BY NC SA 4.0 license. These components will be hosted on Zenodo. However, the MSSEG-1 component (Commowick et al., 2018) was only available via credentialized access via the Shanoir Platform under an OFSEP DUA. Getting this access took some time. However, we have reached an agreement with OFSEP to allow us to host our copy of the MSSEG-1 data on Zenodo under their DUA to facilitate faster and simpler credentialized access within a consistent, pre-processed data format. Thus, the in-domain training, dev and eval as well as the shifted dev set will be available for download from Zenodo. The data will be split into two archives - the MSSEG archive, which will require credentialized access which will be fast to achieve, and the remaining data, which will be freely hosted under a permissive CC BY NC SA 4.0 license. Researchers wishing to use the dataset will need to download both archives and then follow the included instructions to combine the two archives into the canonical splits we have defined.

Finally, the dataset sourced at Lausanne, which is used as the Shifted evaluation set, was collected in such a way that sharing the dataset itself is not possible, even via credentialized access. Specifically, patients have the right to withdraw their data from the dataset at any time - the only way to ensure this is for the data collectors to maintain both ownership and control over the dataset. However, the data owners are happy to freely allow researchers to evaluate their models on this data via dockers on a public leaderboard, which will be hosted in Grand-Challenge.

Societal Consequences Research on uncertainty estimation and robustness aims to make AI safer and more reliable, and therefore has limited negative societal consequences overall. As discussed in sections 3

and 4, both tasks considered for Shifts 2.0 have high societal importance. High-quality automatic segmentation of MS lesions can enable greater patient throughput, more regular checkups, and in the long term, a more personalised treatment plan. Similarly, accurate and reliable cargo vessel power consumption estimation can help optimize fuel usage, carry less surplus fuel and thereby decrease both the cost of marine cargo transport as well as it's climate impact.

Guidelines for Ethical Use Users of this dataset are encouraged to use it for the purpose of improving the reliability and safety of large-scale applications of machine learning. Furthermore, we encourage users of our dataset to develop compute and memory efficient methods for improving safety and reliability.

As part of this data features 3D MRI brain scans taken from MS patients, users of this dataset should not attempt to establish or retrieve the identity of the patients. Furthermore, users should not link this data to any other database in a way that could provide identifying information. Users similarly should not request the pseudonymisation key that would link this data to an individual's personal information. When sharing secondary or derivative data (e.g. group statistical maps, learnt models, etc...), users should only do so if they are on a group level, and information from individual participants cannot be deduced.

Responsibility The authors confirm that, to the best of our knowledge, the released dataset does not violate any prior licenses or rights. However, if such a violation were to exist, we are responsible for resolving this issue.

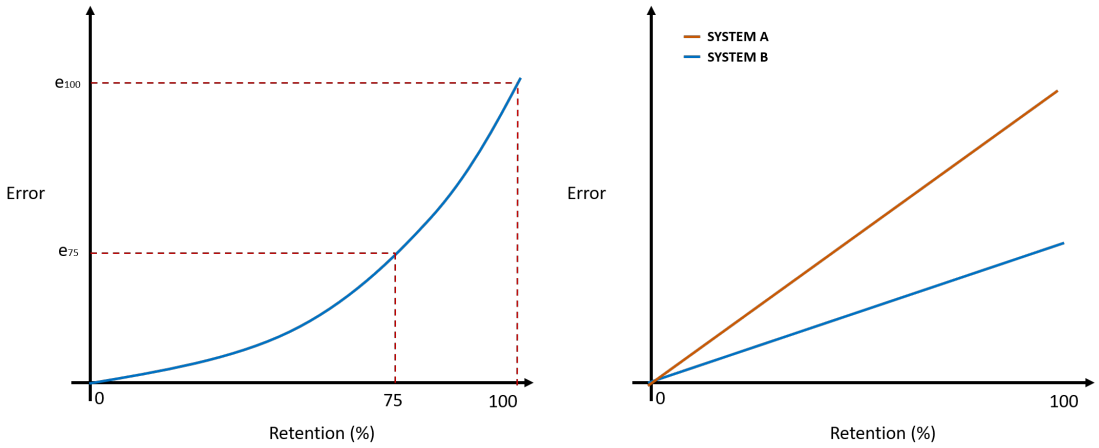
B Assessment Metrics

As discussed in Section 2, in this work we consider robustness and uncertainty estimation to be two equally important factors in assessing the reliability of a model. We assume that as the degree of distributional shift increases, so should a model's errors; in other words, a model's uncertainty estimates should be correlated with the degree of its error. This informs our choice of assessment metrics, which must *jointly* assess robustness and uncertainty estimation.

One standard approach to jointly assess robustness and uncertainty are *error-retention curves* Malinin (2019); Lakshminarayanan et al. (2017), which plot a model's mean error over a dataset, as measured using a metric such as error-rate, MSE, or nDSC, with respect to the fraction of the dataset for which the model's predictions are used. These retention curves are traced by replacing a model's predictions with ground-truth labels obtained from an oracle in order of *decreasing uncertainty*, thereby decreasing error. Ideally, a model's uncertainty is correlated with its error, and therefore the most errorful predictions would be replaced first, which would yield the greatest reduction in mean error as more predictions are replaced. This represents a hybrid human-AI scenario, where a model can consult an oracle (human) for assistance in difficult situations and obtain from the oracle a perfect prediction on those examples.

The area under the retention curve (R-AUC) is a metric for jointly assessing robustness to distributional shift and the quality of the uncertainty estimates. R-AUC can be reduced either by improving the predictions of the model, such that it has lower overall error at any given retention rate, or by providing estimates of uncertainty which better correlate with error, such that the most incorrect predictions are rejected first. It is important that the dataset in question contains both a subset "matched" to the training data, and a distributionally shifted subset.

Schematic explanations of error-retention curves are given in Figure 3, which demonstrates how these curves jointly assess robustness and uncertainty by measuring the area under such curves. Consider Figure 3a. Here we can see that replacing a certain percentage of the models predictions with ground truth labels will decrease the error rate. Specifically, e_{100} is the performance of the system using all the data while e_{75} is the error of the system using the top 75% of the data with the rejected data set to the ground-truth. Now

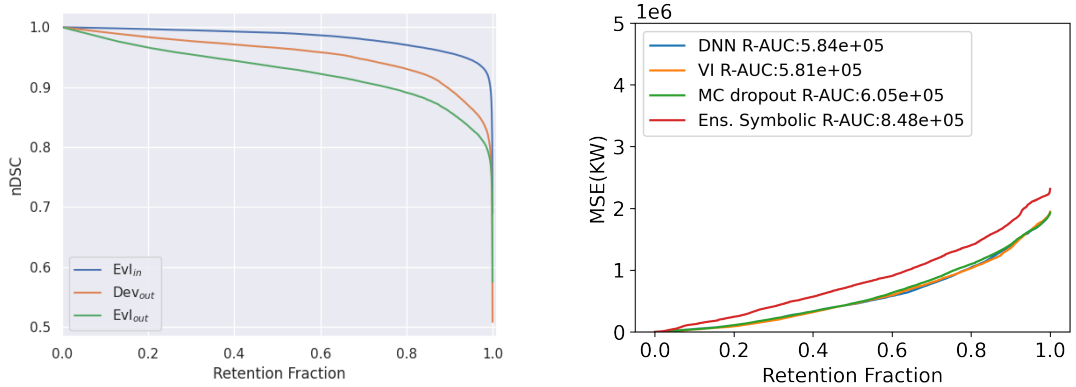


(a) Example (b) Robustness

(c) Uncertainty (d) Mixture

Figure 3: Schematic explanation of error retention curves.

consider Figure 3b, where we demonstrate robustness. Here we plot retention curves for two systems, where one system is broadly more robust than the other. Predictions are rejected in a random, uninformative order, yielding a straight line. Here we can see that the more robust system (System B) will have a lower area under the retention curve (R-AUC) than the less robust system A. Now consider Figure 3c, where we demonstrate uncertainty quality. For the same system, two different uncertainty approaches are consider where the uncertainty measure Y produces a better ranking which is more strongly correlated with the degree of error than uncertainty measure X. As a result, the largest errors are rejected first. Thus, the area under the retention curve constructed using the ranking defined by measure Y is smaller than under the retention curve defined using measure X. Finally, let's consider Figure 3d, where we show a *joint* assessment of robustness and uncertainty. Here, despite having worse predictive robustness, system A has a better uncertainty ranking measure, leading the a smaller R-AUC. Thus, this model is capable of achieving more operating points where it has lower error than system A, and is therefore better in terms of joint assessment of robustness and uncertainty. The converse scenario can also occur - we could have a more which is so robust, despite uninformative uncertainty, that is achieves superior performance at all retention percentages than a less robust model with informative uncertainty estimates.



(a) MS Segmentation (b) Power Estimation

Figure 4: Example error retention curves for the two tasks of the Shifts 2.0 Dataset.

In addition to area under an error-retention curve, we also consider an F1-retention curve, which is broadly similar, but uses the notion of 'acceptable' error to assess whether uncertainty estimates can be used to detect 'un-acceptable errors'. The metric is less susceptible to errors at the level of noise, but it is not always possible to define what is an 'acceptable error'. Thus, this metric is only used to assess the power estimation tasks, but not the segmentation task. For a detailed descriptions of the F1-retention curve, please see Malinin et al. (2021).

The area under the error-retention curve and F1-retention curve is a *summary statistic* which describes possible *operating points*. We can specify a particular operating point, such as 95% retention, and evaluate the error or F1 at that point for comparison. This is also an important figure, as all models work at a particular operating point which satisfies task-specific desiderata.

C LESION SEGMENTATION

The current appendix provides further details on the Medical data collected as well as more complete set of baseline results.

C.1 DATASET DESCRIPTION

Canonical Dataset Construction Here, we detail additional experiments using a UNET model that were run in order to select the canonical partitioning of the data. For both Tables 4 and 5, ensembles of size 5 were used. First, in Table 4, models are trained on data from each location and evaluated on all other location to identify the greatest shifts. Three different approaches to choosing the classification threshold were examined - in-domain on the corresponding dev set of the location, on the train and dev sets of the different location (not used for training), and finally on the actual test-set of each location. These different threshold tuning strategies allow us to examine the range of expected and upper bound performance on each location.

Thresholding	Train	Rennes	Bordeaux	Lyon	Ljubljana	Best	Lausanne
In-domain Dev	Rennes	50.51	**72.95**	54.81	35.78	47.05	40.63
	Bordeaux	49.46	68.18	55.12	34.70	50.13	46.71
	Lyon	58.73	69.75	**66.68**	42.51	54.84	52.00
	Ljubljana	**66.18**	70.29	65.98	**57.03**	**63.45**	62.12
	Best	59.03	71.28	63.93	46.95	63.27	55.74
Out-domain Train + Dev	Rennes	57.70	67.91	59.38	47.37	56.26	-
	Bordeaux	50.90	**71.80**	56.96	34.70	50.13	-
	Lyon	65.23	71.65	**69.00**	52.44	55.18	-
	Ljubljana	**66.91**	70.00	66.67	**59.03**	60.85	-
	Best	60.54	71.07	64.17	48.09	**61.97**	-
Out-domain Test	Rennes	65.11	**73.13**	60.19	47.40	57.71	54.26
	Bordeaux	50.90	71.87	56.96	34.70	50.13	46.71
	Lyon	65.79	72.50	**69.01**	52.44	57.29	60.13
	Ljubljana	**68.37**	70.25	66.73	**59.85**	64.17	66.30
	Best	61.19	71.35	64.17	48.58	**64.43**	58.34

Table 4: Cross-performance results using nDSC (↑) (%) for selected splits. Ensembles of 5 models are always used. Threshold is searched in increments of 0.01. The following threshold were used respectively for the in-domain dev threshold tuning: [0.8, 0.1, 0.47, 0.66, 0.50]. Here, R-AUC is calculated over all voxels in each image.

We performed N-fold cross-validation in Table 5 to determine which location should be considered as the shifted set as training on single locations may lead to unreliable conclusions due to the small size of the training sets. Train 5 systems (each one an ensemble) using all training data apart from one site at a time. Hyperparameters are tuned using all the dev sets apart from the site excluded. We evaluate this system on all the data (train + dev + test) from the excluded site (out) and in-domain test sets too. From the results, Ljubjana is an appropriate choice for the shifted development set as it faces the greatest degradation compared to in-domain performance.

Data distributions Here, a more detailed characterisation of the datasets (Trn, Dev_{in}, Evl_{in}, Dev_{out} and Evl_{out}) described in Section 3 is given. Distributions of total lesion volumes and number of lesions across patients are shown in Figure 5. General characteristics of the datasets are given in Table 6. It can be seen that

Excluded Location	Model	nDSC (%) (↑) In	Out	Lausanne	R-AUC (%) (↓) In	Out	Lausanne
Rennes	Single	$66.43_{\pm 0.50}$	$70.86_{\pm 0.42}$	$64.50_{\pm 0.83}$	$3.10_{\pm 0.21}$	$2.81_{\pm 0.55}$	$6.54_{\pm 0.96}$
	Ensemble	68.01	72.48	66.46	2.02	1.69	4.14
Bordeaux	Single	$65.66_{\pm 0.74}$	$72.14_{\pm 1.10}$	$63.21_{\pm 1.26}$	$3.09_{\pm 0.19}$	$2.48_{\pm 0.36}$	$6.61_{\pm 0.58}$
	Ensemble	66.33	72.73	63.25	1.87	1.33	4.05
Lyon	Single	$63.51_{\pm 0.18}$	$69.27_{\pm 0.69}$	$61.85_{\pm 1.69}$	$3.68_{\pm 0.76}$	$2.33_{\pm 0.62}$	$6.69_{\pm 1.54}$
	Ensemble	65.21	70.69	64.46	2.54	1.81	4.70
Ljubljana	Single	$67.59_{\pm 0.63}$	$49.33_{\pm 1.52}$	$55.70_{\pm 1.04}$	$2.77_{\pm 0.98}$	$7.84_{\pm 2.21}$	$9.87_{\pm 1.40}$
	Ensemble	68.89	50.85	57.53	1.76	4.66	7.40
Best	Single	$65.87_{\pm 1.62}$	$57.37_{\pm 0.79}$	$61.78_{\pm 2.21}$	$2.65_{\pm 0.48}$	$3.05_{\pm 0.69}$	$5.70_{\pm 1.22}$
	Ensemble	66.68	58.38	61.93	1.54	1.69	3.15

Table 5: N-fold cross-validation with nDSC (↑) (%) as the performance metric. The threshold is selected based on the (in-domain) development set. The following thresholds are used: [0.25, 0.55, 0.25, 0.35, 0.55]. Entropy is used as the uncertainty measure for single models and reverse mutual information for ensembles.

the difference in datasets comes not only from the location of the medical center or the scanner type, but also from the the sizes of lesions. Out-of-domain datasets have more subjects with smaller lesions. Per patient lesion counts, however, do not vary significantly across the datasets. Additionally, it was mentioned in the main paper that the component datasets of our benchmark are based on ISBI Carass et al. (2017b;a), MSSEG-1 Commowick et al. (2018) and PubMRI Lesjak et al. (2017). Table 7 offers additional meta-information on these source datasets with regard to age and gender ratio of the patient scans from each of these datasets.

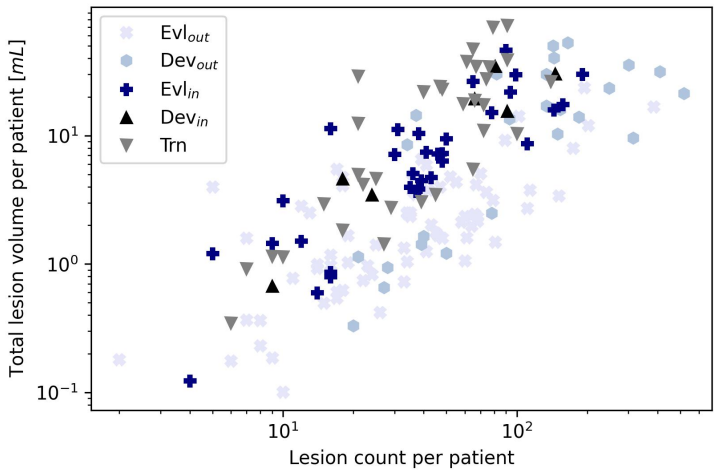


Figure 5: Loglog plot of white matter lesions characteristics in terms of per patient total lesion volume (TLV) and number of lesions for different datasets.

Parameters	Trn	Dev_{in}	Evl_{in}	Dev_{out}	Evl_{out}
Total lesion count	1628	435	1738	3544	3826
MS stages	RR, PP, SP*			RR, SP, PR, CIS	RR
Average across scans TLV, mL	$18.58_{\pm 18.75}$	$15.49_{\pm 12.42}$	$10.03_{\pm 10.28}$	$17.10_{\pm 15.56}$	$3.34_{\pm 4.13}$

Table 6: Additional characteristics of the datasets, such as total amount of lesions in a dataset, MS stages and average across scans total lesion volume (TLV) in milliliters. MS stages abbreviations: RR - relapsing remitting, PP - primary progressive, SP - secondary progressive, CIS - clinically isolated syndrome.
*Information about MS stages in MSSEG-1 was not found.

	ISBI	MSSEG-1	PubMRI
Age (years)	$40.4_{\pm 9.3}$	$45.3_{\pm 10.3}$	39 (median)
Gender ratio (M:F)	0.21	0.40	0.23
Inter-rater agreement (DSC)	0.63	0.71	0.78

Table 7: Age and gender meta-information for source datasets. Additionally, inter-rater agreement is reported as DSC.

Format The data will be shared as a series of compressed .nii files, all the data within will be pre-processed, interpolated to the 1mm iso-voxel space and skull-stripped for additional anonymisation. We will share both the T1 weighted and FLAIR modalities.

C.2 Performance metrics

We now detail performance metrics used to assess lesion segmentation models.

C.2.1 Normalized Dice Similarity Coefficient (nDSC)

Typically, the Dice Similarity Coefficient (DSC) is used as the performance metric between the ground-truth Y and its corresponding prediction $\hat{Y}$:

$$\text{DSC} = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} = \frac{2TP}{FP + 2TP + FN} = 2\frac{\texttt{precision * recall}}{\texttt{precision + recall}}$$

The reported score is usually the DSC averaged across all patient scans. However, DSC is biased to yield greater values for patients that have a greater lesion load i.e. a greater probability of the event occurring, where the event here is described as identifying a voxel as a lesion. To de-correlated DSC with lesion-load and obtain an unbiased metric of permormance, we consider a normalised DSC (nDSC). The following steps explain and justify how and why we calculate the proposed nDSC:

1. The probability of a successful event (identifying a lesion) influences the DSC score as the precision at 100% recall varies across the patients (the precision at 100% recall is simply the percentage of lesion voxels for the patient - i.e. the lesion load).
2. The DSC score is calculated as a geometric ratio of the precision, Pr_τ, and recall, Re_τ values at a selected threshold, τ (ML models typically have a probabilistic prediction for each voxel which must be compared against a threshold to classify as either a positive class or a negative class).

3. Here, the recall is held fixed and the precision for each patient is adjusted ($\text{Pr}_\tau \rightarrow \overline{\text{Pr}}_\tau$) by a different amount such that the cross-patient performance can be fairly evaluated.
4. The new value of the precision is determined by the scaling applied to the FP (false positives) which is scaled by a factor, k_p that is different for each patient, p.
5. k_p for each patient is determined by using the 100% recall rate point as this point is not influenced by model performance.
6. Hence, k_p for patient p is the factor the FP at 100% recall must be scaled by in order to ensure the precision achieved is a chosen reference value, r. Derivation of deducing k_p is given. The subscript 100% denotes operating at 100% recall.

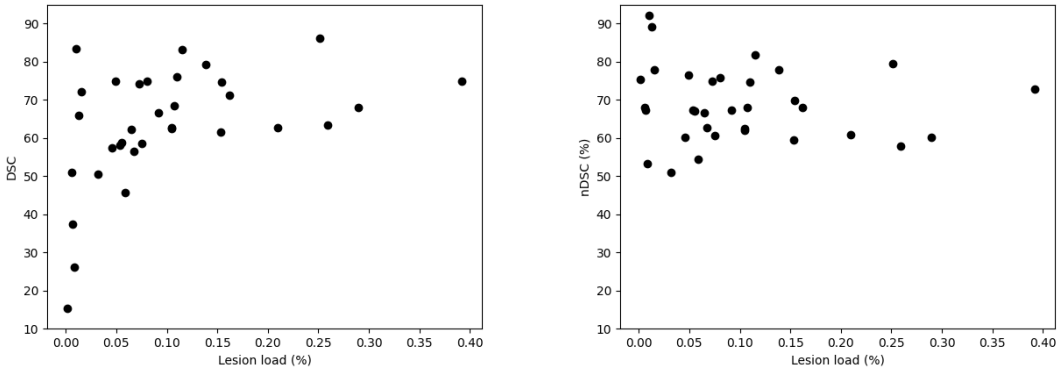
$$\text{Pr}_{100\%} = \frac{\text{TP}_{100\%}}{\text{TP}_{100\%} + \text{FP}_{100\%}}, \quad r = \overline{\text{Pr}}_{100\%} = \frac{\text{TP}_{100\%}}{\text{TP}_{100\%} + k_p\text{FP}_{100\%}}, \quad k_p = \frac{(1-r)\text{TP}_{100\%}}{r\text{FP}_{100\%}}$$

7. Here, r is selected as 0.1% because this is approximately the average precision across the patients at 100% recall (i.e. the average fraction of lesion voxels).
8. The recall is not influenced by scaling the FP by k_p.
9. The precision is directly affected as the new precision at our selected operating point (threshold to form the segmentation mask), τ^*, is given by:

$$\overline{\text{Pr}}_{\tau^*} = \frac{\text{TP}_{\tau^*}}{\text{TP}_{\tau^*} + k_p\text{FP}_{\tau^*}}$$

   Recall, k_p is given in step 6.
10. Thus, nDSC is calculated as the geometric mean of $\overline{\text{Pr}}_{\tau^*}$ and Re_{τ^*} for each patient.

The averaged nDSC is used as the predictive performance metric.



(a) DSC (b) nDSC

Figure 6: Empirical relationship of each metric with lesion load on Evl_{in} using UNET ensemble.

We empirically demonstrate that the nDSC metric is less dependent on the lesion load compared to DSC via Figure 6 and table8. Recall, lesion load is defined as the fraction of voxels that are lesion voxels for a given subject. Figure 6 plots the performance in terms of both DSC and nDSC against the lesion load for

each subject for Evl_{in}. It is clear that DSC is dependent on the lesion load while nDSC decorrelates this relationship by flat line average. Table 8 presents the transition table between DSC and nDSC as well as providing the Spearman's rank correlation coefficients between either DSC or nDSC with the lesion load. Notably, the nDSC metric is less correlated with the lesion load than DSC for each of the splits.

Split	Performance		Correlation	
	DSC	nDSC	DSC	nDSC
dev-in	71.71	68.54	0.63	-0.09
dev-out	49.85	49.33	0.57	0.46
eval-in	63.16	67.59	0.44	-0.10
eval-out	48.48	55.79	0.40	0.18

Table 8: Performance and Pearson's rank correlation coefficients between metric and the lesion load for the canonical white matter lesion segmentation splits using the baseline UNET model ensemble.

C.2.2 Lesion-scale F1 score

For MS lesion segmentation task it is important to assess not only the overall voxel-level segmentation quality, but also the lesion detection quality. Therefore, in addition to the nDSC we calculate the lesion-scale F1 score.

A general formula for computation of the F1-score:

$$F_1 = \frac{TP}{TP + 0.5(FP + FN)} \tag{1}$$

can be adapted for the assessment of lesions detection quality given a proper definition of true positive, false positive and false negative lesions.

We use the intersection over union (IoU) between lesions on a ground truth map and connected components on a corresponding prediction map to derive these definitions. In particular, the following condition were used:

- TP: If the maximum IoU between a connected component on the prediction map with lesions on the ground truth is greater than 0.5.
- FP: If the maximum IoU between a connected component on the prediction map with lesions on the ground truth is less than 0.5.
- FN: If the maximum IoU between a lesion on the ground truth map with connected components on the prediction map is less than 0.5.

C.3 Additional Results

For completeness, Monte Carlo dropout Gal & Ghahramani (2016) based ensembles are considered here too using the UNET architecture. The baseline single models considered here have no dropout (as this gives best performance on Dev-in) and the deep ensembles are built using these single models. The deep ensemble is formed by averaging the output probabilities from 5 distinct single models. A separate set of 5 models are trained with 50% dropout in each model in order to be able to perform Monte Carlo Dropout (MCDP) as an additional comparison. The single models here have dropout usually turned off at inference time. For MCDP,

a single model is taken and dropout is turned on at inference time with an ensemble formed from 5 separate runs of the model (as the dropout introduces stochasticity). The process is repeated for each of the single models with dropout to get averaged results. As each single model yields a per-voxel probabilistic prediction, ensemble-based uncertainty measures Malinin (2019); Malinin & Gales (2021) are available for uncertainty quantification. Our ensembled models (Deep Ensemble and MCDP) use reverse mutual information Malinin & Gales (2021) as the choice of uncertainty measure. Single models use the entropy of the discrete binary probability distribution at each voxel to capture the uncertainties. All results reported for single models are the mean of the individual model performances with one standard deviation indicated.

Tables 9 and 10 present the performance ability of various baseline models. Table 9 focuses on the ability of the models to identify the exact delineations of lesions through nDSC (voxel-scale) while Table 10 compares the lesion detection ability of the models with F1 (lesion-scale). Comparing the in-domain performance against the out-of-domain performance, it is clear that the shift in the location naturally leads to severe degradation in performance at both the voxel-scale and the lesion-scale with drops exceeding 10% nDSC and F1. Comparing the deep ensembles against the single models, it is clear that ensembling such models boosts performance by about 1% nDSC and 1% F1 for each of the test sets. In particular, the transformer based architecture, UNETR, is able to outperform the fully convolutional architecture, UNET, for both the single and ensembled performance in terms of delineation and lesion detection of about 2% nDSC and 5% F1 respectively across the various splits. Introducing dropout in the models at training time costs the single model in performance at both voxel and lesion-scales with greater degradation observed in the in-domain splits. Consequently, the detrimental effect of dropout at training time seriously harms the performance of the MCDP systems that keep the dropout on at training time.

Arch	DP	Model	nDSC (%) (↑)			
			Dev_{in}	Dev_{out}	Evl_{in}	Evl_{out}
UNET	0.0	Single	$68.54_{\pm 0.68}$	$49.33_{\pm 1.52}$	$67.59_{\pm 0.63}$	$55.79_{\pm 1.04}$
		Deep Ensemble	69.70	50.85	68.89	57.53
	0.5	Single	$59.73_{\pm 1.17}$	$48.35_{\pm 1.73}$	$63.93_{\pm 0.45}$	$54.43_{\pm 1.41}$
		MCDP	$60.65_{\pm 0.91}$	$44.70_{\pm 1.35}$	$61.78_{\pm 0.90}$	$50.06_{\pm 1.67}$
UNETR	0.0	Single	$71.21_{\pm 0.96}$	$51.60_{\pm 1.66}$	$69.27_{\pm 0.94}$	$56.76_{\pm 2.63}$
		Deep Ensemble	72.51	53.46	71.41	59.49

Table 9: Lesion segmentation: Performance at voxel-level with nDSC with 1 standard deviation quoted for single results.

Arch	DP	Model	F1 (%) (↑)			
			Dev_{in}	Dev_{out}	Evl_{in}	Evl_{out}
UNET	0.0	Single	$25.02_{\pm 2.51}$	$8.17_{\pm 0.73}$	$25.46_{\pm 1.51}$	$14.79_{\pm 0.71}$
		Deep Ensemble	28.07	9.04	27.74	16.74
	0.5	Single	$14.42_{\pm 0.43}$	$6.75_{\pm 0.70}$	$18.66_{\pm 0.51}$	$11.85_{\pm 0.47}$
		MCDP	$12.61_{\pm 0.89}$	$4.59_{\pm 0.78}$	$17.31_{\pm 0.95}$	$10.70_{\pm 0.58}$
UNETR	0.0	Single	$33.60_{\pm 1.36}$	$15.03_{\pm 1.16}$	$33.85_{\pm 0.43}$	$17.19_{\pm 1.22}$
		Deep Ensemble	35.22	15.80	35.61	18.90

Table 10: Lesion segmentation: Performance at lesion-level with F1 with 1 standard deviation quoted for single results.

C.4 UNCERTAINTY ESTIMATION

Table 11 explores the joint robustness and uncertainty quantification performance using the R-AUC metric. Here, the deep ensemble of the UNETR outperforms all other systems, achieving R-AUC scores as low as 0.63 on Evl_{in} and 2.88 on Evl_{out}. It is interesting to note that despite performing worse at voxel-scale identification of lesions, the MCDP system does better than its equivalent single system when jointly assessing uncertainty and robustness. Therefore, it is clear that the quality of the uncertainty measures in the ensembled-based models (including both the deep ensemble and MCDP) allows the development of richer uncertainty quantification measures compared to single models. Figure 7 presents the corresponding retention curves (averaged across all the patients with one example model chosen for the single systems) using the deep ensembled UNET on the Evl_{in}, Dev_{out} and Evl_{out} splits. All systems substantially outperform a randomized ordering as a large volume of the input brain image is non white-matter tissue, for which the system is correctly certain that there are no white matter lesion voxels present in those regions. Particularly, the retention curve for the Evl_{in} appears to be very close to ideal which demonstrates the high quality of its voxel-scale uncertainties at identifying regions where the model is not confident in its prediction.

Arch	DP	Model	R-AUC (%) (↓)			
			Dev_{in}	Dev_{out}	Evl_{in}	Evl_{out}
UNET	0.0	Single	$2.51_{\pm 0.59}$	$7.84_{\pm 2.21}$	$2.77_{\pm 0.98}$	$9.87_{\pm 1.40}$
		Deep Ensemble	1.17	4.66	1.76	7.40
	0.5	Single	$2.62_{\pm 0.56}$	$8.76_{\pm 1.08}$	$2.66_{\pm 0.56}$	$9.71_{\pm 1.53}$
		MCDP	$1.92_{\pm 0.26}$	$6.77_{\pm 0.79}$	$2.52_{\pm 0.41}$	$7.89_{\pm 1.04}$
UNETR	0.0	Single	$1.89_{\pm 0.84}$	$6.17_{\pm 1.99}$	$1.95_{\pm 0.70}$	$6.47_{\pm 2.08}$
		Deep Ensemble	0.34	1.52	0.63	2.88

Table 11: Lesion segmentation: Joint robustness and uncertainty assessment (using reverse mutual information for ensembled models and entropy for single models) at voxel-level with R-AUC. 1 standard deviation is quoted for single results.

Figure 8 gives an idea about the spatial distribution of uncertainty. In particular, it can be seen that higher uncertainty regions are located around predicted lesions, therefore should be related to the quality of delineation. False negative lesions, however, can also have higher uncertainties in comparison to the background.

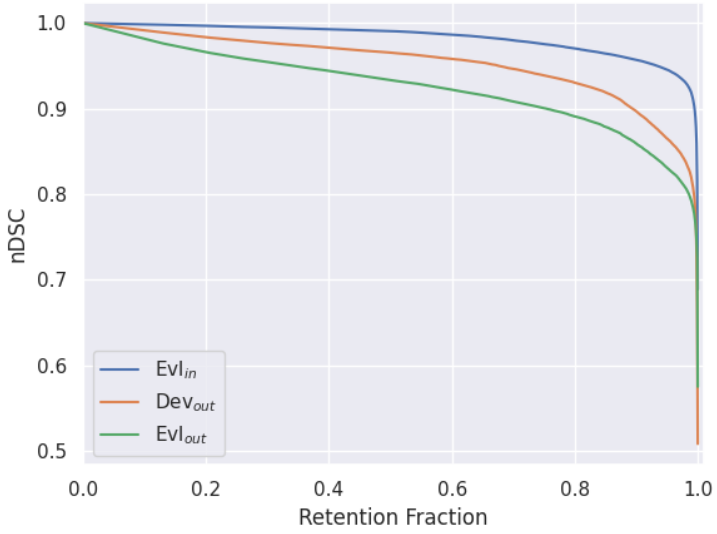


Figure 7: nDSC retention curves using the ensembled UNET on various canonincal splits.

C.5 Sub-Population Analysis

In this section we provide a sub-population analysis of the data. We examine the gender, age and lesion load distributions across the different splits, and how each of these properties interacts with model performance. As figure 9 shows, in all the datasets, there are roughly four times as many female patients as male patients. This overall is indicative of the higher incidence of MS in women than in men. Female patients, on average, are 2-6 years older than male patients in all datasets, except eval$_o ut$, where the female patients are on average younger. Patients in the shifted datasets are on average younger. Patients from Lausanne (eval$_o ut$) have a far lower lesion load, as they are in an earlier stage of MS. There are no clear gender and age related correlations with lesion load.

	TRN		DEV'$_{in}$ + EVL'$_{in}$		DEV$_{out}$		EVL$_{out}$	
	M	F	M	F	M	F	M	F
mean	43.12	45.13	41.14	47.59	33.0	39.21	37.55	33.96
median	40.5	44.0	36.0	47.0	31.0	39.0	35.0	35.0
max	54.0	66.0	59.0	61.0	43.0	60.0	62.0	48.0
min	35.0	24.0	33.0	29.0	25.0	26.0	20.0	21.0

Table 12: Age Breakdown by Gender across datasets. Trn, Dev$_i n$ and Eval$_i n$ statistics do not feature data from ISBI (Best), as per-patient metadata is not available.

We have provided a performance break-down by gender for each datasets, as well as performance breakdown by age and lesion load, for both UNet and UNEtr models in figures 10 and 11, respectively. The results show that there are no clear gender-based differences in model performance. Similarly, there is no clear correlation between model performance and age. There is a minor correlation (0.46-0.47) between model performance

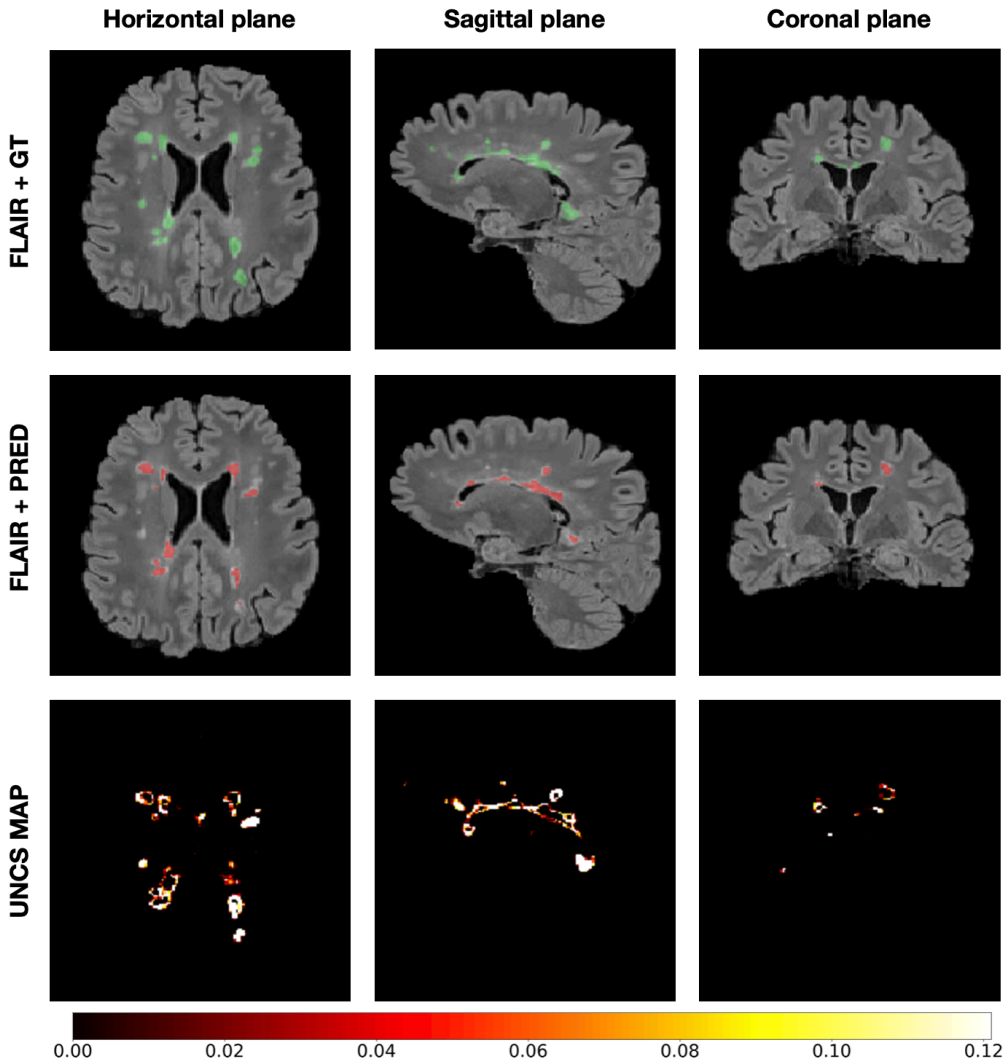


Figure 8: Examples on one subject of a FLAIR image with a ground truth (**FLAIR + GT**) and predicted (**FLAIR + PRED**) WML maps overlays and an uncertainty map (**UNCS MAP**). For each of the 3D maps single horizontal, sagittal and coronal slices are displayed. Predictions were obtained using an ensemble of 5 UNET models. Uncertainty map was computed as reversed mutual information from the probabilistic voxel-wise predictions of models in ensemble. Color bar corresponds to the uncertainty map, where outlying values above 0.121 are displayed in white. All images were displayed using ITK-SNAP software Yushkevich et al. (2006).

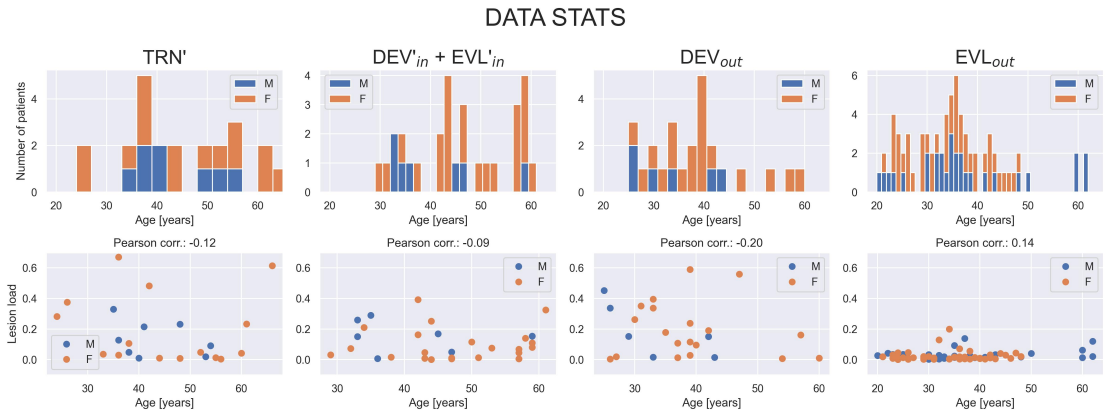


Figure 9: Sub-Population Statistics. Trn, Dev$_i n$ and Eval$_i n$ statistics do not feature data from ISBI (Best), as per-patient metadata is not available.

	TRN		DEV'$_{in}$ + EVL'$_{in}$		DEV$_{out}$		EVL$_{out}$	
	M	F	M	F	M	F	M	F
mean	0.134	0.197	0.154	0.099	0.187	0.192	0.031	0.024
median	0.109	0.047	0.154	0.070	0.151	0.161	0.021	0.013
max	0.330	0.670	0.290	0.392	0.452	0.589	0.138	0.200
min	0.010	0.003	0.007	0.001	0.014	0.004	0.002	0.001

Table 13: Lesion Load by Gender across datasets. Trn, Dev$_i n$ and Eval$_i n$ statistics do not feature data from ISBI (Best), as per-patient metadata is not available.

and lesion load on datasets featuring low lesion load. This is, however, expected, as it shows the intrinsic difficulty of the task. Very small, hard-to-detect lesions are harder to segment accurately. Tables 14 and 15 show the mean, median, minimum, and maximum per-patient nDSC for male and female patients across all datasets. The results on in-domain data show that the performance on male and female patients is similar. Out of domain, performance on female patients is a little lower on average. However, it is important to highlight that these results and statistics are collected based on a small sample size and are therefore noisy. Furthermore, there are 4 times as few male patients than female patients, so statistics on male patients are noisier. Thus, we would be hesitant to make any strong statements regarding sub-population performance bias on our models.

	TRN		DEV'$_{in}$ + EVL'$_{in}$		DEV$_{out}$		EVL$_{out}$	
	M	F	M	F	M	F	M	F
mean	73.92	75.85	68.51	70.72	56.59	46.35	58.98	55.45
median	75.85	75.89	69.79	73.84	54.36	47.05	62.37	58.76
max	83.89	93.57	76.55	92.21	74.44	68.5	76.22	87.21
min	61.55	64.64	57.91	49.11	41.24	0.0	19.26	0.0

Table 14: UNet Performance (% nDSC) by gender across Datasets

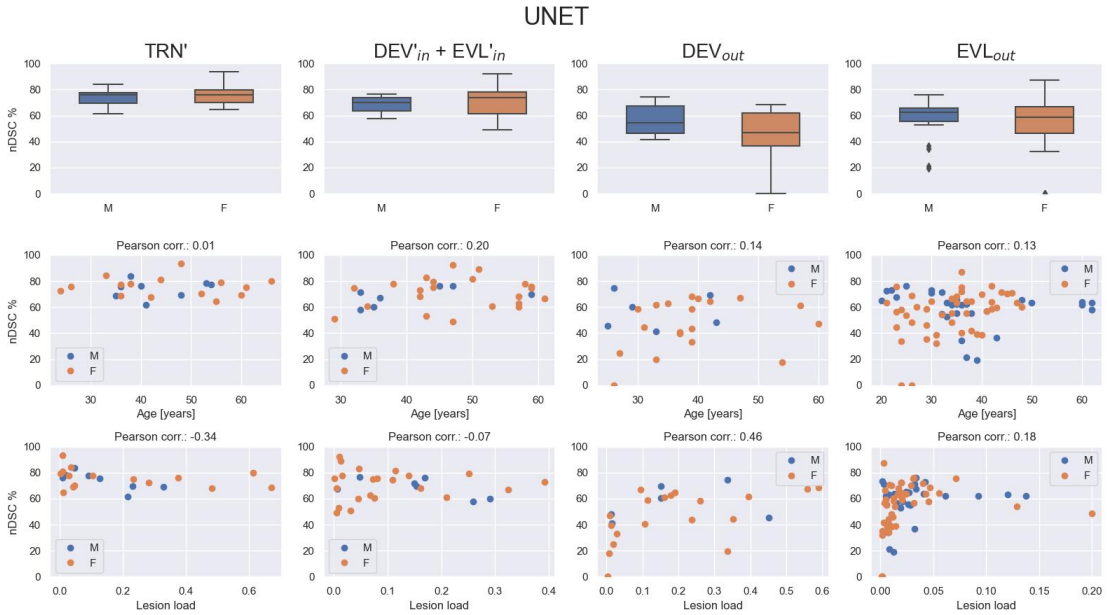


Figure 10: Sub-Population Statistics.

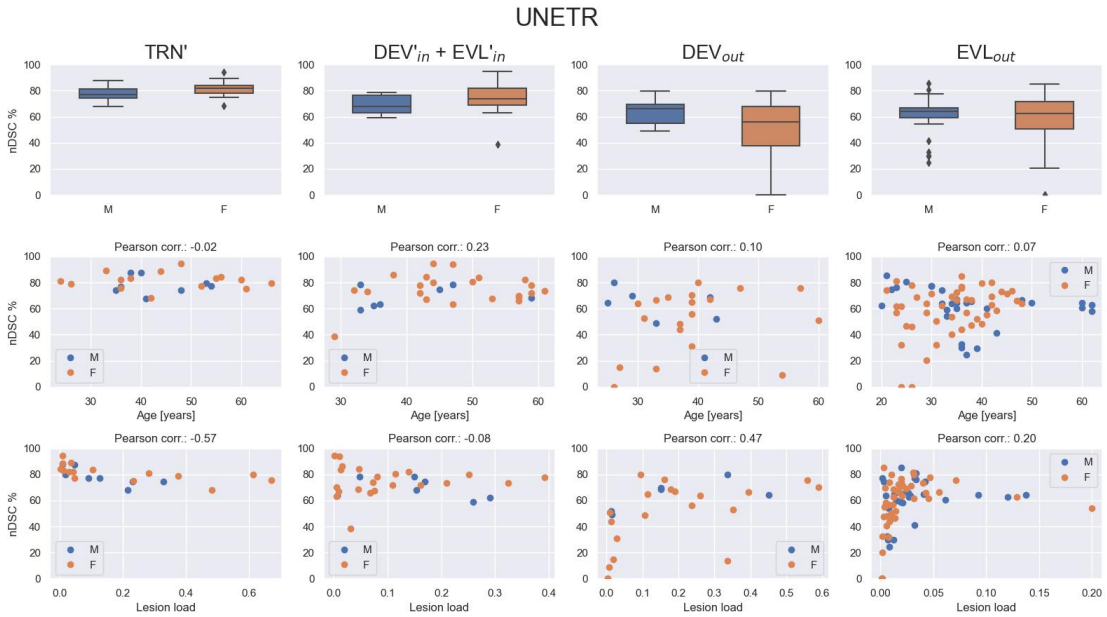


Figure 11: Sub-Population Statistics.

	TRN		DEV'$_{in}$ + EVL'$_{in}$		DEV$_{out}$		EVL$_{out}$	
	M	F	M	F	M	F	M	F
mean	78.21	81.55	68.51	70.72	63.92	50.16	60.89	58.67
median	77.14	82.02	69.79	73.84	66.45	55.97	63.88	62.72
max	87.75	94.39	76.55	92.21	79.95	79.93	85.44	85.08
min	67.81	68.2	57.91	49.11	49.03	0.0	24.66	0.0

Table 15: UNetR Performance (% nDSC) by gender across Datasets

D SHIP POWER CONSUMPTION

D.1 DATASET DESCRIPTION

Collection Process Data is collected from a real vessel as it experiences various weather, loading and operational conditions. The sampling frequency is 1 min, and achieved by either interfacing directly with the vessels sensors or through already existing onboard signal aggregation systems like the Electronic Chart Display and Information System (ECDIS) or the vessels Alarm Monitoring System (AMS). The interfacing in either case is performed by dedicated IoT edge devices that gather all relevant data and transmit them via satellite link.

Preprocessing, Cleaning and Labeling The available features are recorded by on-board sensors and the global positioning system (GPS) is being used to complement the acquired data with weather data from a global weather provider. The data is preprocessed to remove extreme outliers and stationary states, for example when a vessel is at port, by applying feature filters. Furthermore, we create a second dataset, the synthetic dataset, by combining the real samples with synthetic power labels generated by our synthetic model (detailed below).

Partitioning into train, development, and evaluation sets We create a canonical partitioning of power estimation dataset so that it contains both in-domain and shifted components. In order to define the distributional shifts, the data split along two dimensions: time and true wind speed, as shown in Figure 1, using the wind speed intervals from Table 16.

The time dimension is intended to capture the non-stationary effects of fouling (no cleaning events occur during the time period under study), whereas the wind speed dimension is intended to capture weather effects (by acting as a proxy since wind is correlated with wind-waves) and to better expose the model's performance in bad or uncertain weather. Partitioning the datasets in more dimensions would have added complexity without adding any practical benefits because the most important uncertainty factors (weather and fouling) are already represented.

Given these shifts, three main subsets are created:

- **Train set**: It covers the time range of 39.4 months starting after a dry docking cleaning event and includes data with true wind speed up to 19 kn.
- **Development set**: It consists of an in-domain partition dev_in and an out-of-domain partition dev_out, with equal representatives. Dev_in is sampled from the same partitions as the train set while dev_out includes more recent records (time period of 6.6 months) that correspond to wind speeds in the range [19, 26) kn.
- **Evaluation set**: Evaluation set, like development set, have an in-domain eval_in and an out-of-domain split eval_out with equal populations.Eval_in is sampled from the same subsets as the train

set. Eval_out is the most shifted partition from the in-domain distribution, containing the most recent records spanning an 18 months period and the most severe wind conditions seen in the whole dataset, corresponding to wind speeds ranging between [19, 40] kn.

The number of records of the proposed partitions (rows) along with the respective populations in each 2D segmentation (columns with prefix group) of the synthetic and real datasets are reported in Tables 17 and 18 respectively.

Wind interval	Range (kn)	Range in Beaufort
1	[0, 9)	Up to 3
2	[9, 14)	3-4
3	[14, 19)	4-5
4	≥ 19	≥ 5

Table 16: Wind intervals considered for data partitioning. Beaufort ranges are defined approximately.

Data	pct (%)	total	Group 1	Group 2	Group 3	Group 4
train	80.3	523190	231626	118698	172866	0
dev_in	-	18108	8017	4108	5983	0
dev_out	-	18108	0	0	0	18108
dev	5.6	36216	8017	4108	5983	18108
eval_in	-	46021	20355	10448	15218	0
eval_out	-	46021	0	0	0	46021
eval	14.1	92042	20355	10448	15218	46021

Table 17: Number of records in the canonical partitioning of the synthetic dataset. The color notation is the same as in Figure 1 and indicates the data segments from which the partitions are sampled.

Data	pct (%)	total	Group 1	Group 2	Group 3	Group 4
train	80.2	530706	236401	119084	175221	0
dev_in	-	18368	8182	4122	6064	0
dev_out	-	18368	0	0	0	18368
dev	5.6	36736	8182	4122	6064	18368
eval_in	-	47227	21037	10597	15593	0
eval_out	-	47227	0	0	0	47227
eval	14.3	94454	21037	10597	15593	47227

Table 18: Number of records in the canonical partitioning of the real dataset. The color notation is the same as in Figure 1 and indicates the data segments from which the partitions are sampled.

Data analysis The violin plots of the features for the canonical partitions for the synthetic dataset (Figure 12) and the real dataset (Figure 13), demonstrate the comparability of the in-domain subsets and the distributional changes that are seen in the out-of-domain partitions, particularly for the target and wind related features.

Synthetic Data Generation For the synthetic dataset, real and sampled features are combined with power labels predicted a synthetic, physics-based model. The synthetic model Tsompopoulou et al. (2022) is a generative function ($f_{synthetic}$) which takes as input a time-series of features (i.e. signals), as recorded from

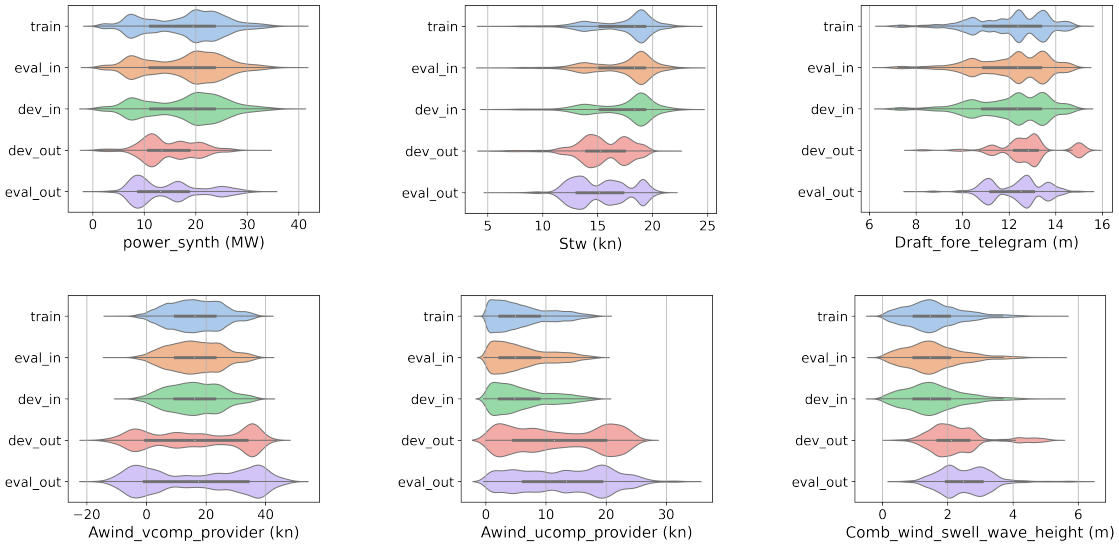


Figure 12: Violin plots for the canonical partitions of the synthetic dataset after the noise injection (scaled to have the same width for better visualization).

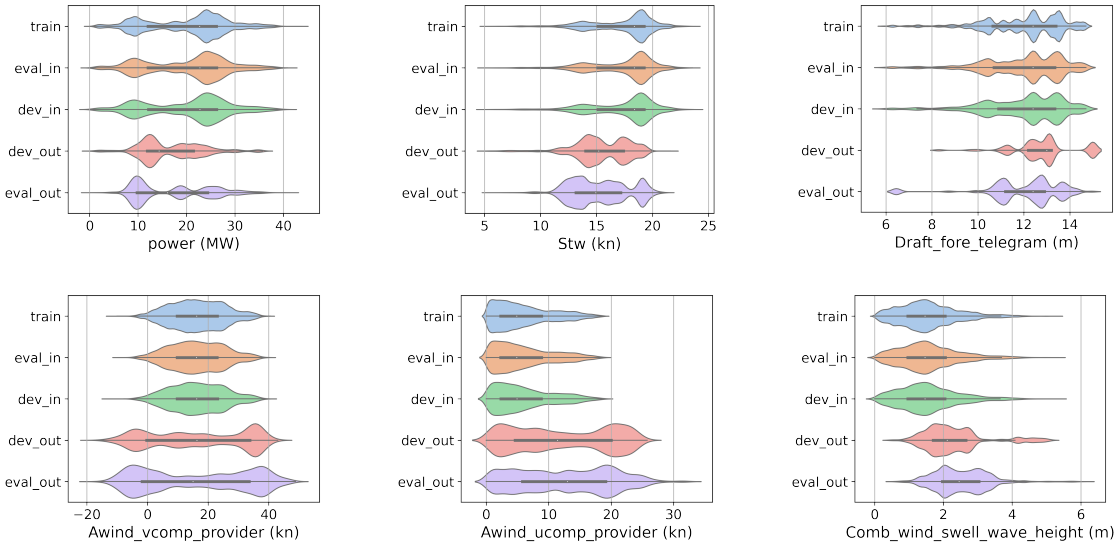


Figure 13: Violin plots for the canonical partitions of the real dataset (scaled to have the same width for better visualization).

a real vessel, and calculates the power consumed by the vessels hull. This function finds the propeller cooperation point after calculating all the components of resistance (bare hull, appendages, wind, waves, fouling drag) for given speed, draft and trim. More specifically, for the generation of synthetic data, a non-linear solver script was created to find the operating point of a given propeller and hull resistance for each desired condition, as described by Bose Bose (2008). The propeller curves (KT, KQ) can either be user defined or use the B-Series Van Lammeren et al. (1969). For the resistance part, the calculation of each component can be described as follows: having the full hydrostatics table of the vessel for the whole range of drafts and trims, along with a series of geometric characteristics (bulb shape and size, transom, appendages etc), calm water resistance is calculated by employing the Holtrop method Holtrop & Mennen (1982) for slender ships (i.e. containers, RoRo, gas carriers) and Modified Holtrop Nikolopoulos & Boulougouris (2019) is used for bulkier ships like large Tankers and Bulk carriers. Following the ISO 15016 ISO (2015), the weather added resistance is found by calculating the wind effect by using the regressions of Fujiwara Fujiwara et al. (2006), while the wave effects are modelled according to STAwave1 and STAwave2 as also introduced by Tsujimoto Tsujimoto et al. (2008). Hull interaction factors are calculated depending on ship type, using empirical formulas, a summary of which can be found in Carlton Carlton (2018). Scale effect corrections, cavitation criteria and corrections were also taken from Carlton Carlton (2018) and Bertram Bertram (2012) . The effect of wake affecting energy saving devices can be modelled by adjusting the interaction factors. Fine-tuning of the method to fit a specific vessel (when there is not enough hydrostatic data, or discrepancies are observed), can be done by using sea trial data and/or detailed factors when available from a towing tank report, or actual measurements of well known conditions. Last but not least, the effect of fouling is modelled as the result of its manifestations (drag, propeller and interaction). The change in drag coefficient is modelled after Townsin Townsin et al. (1981), the effect of fouling on the propeller performance is modelled as in Seo Seo et al. (2016) (increase in torque coefficient), as also described in Carlton Carlton (2018) and the change of interaction factors are modelled after Farkas Farkas et al. (2020). All the aforementioned models produce the effect of fouling on each component over time, which is measured from each drydock / cleaning event. While this allows for a sophisticated modelling of the interaction of features and power used, it still nevertheless a model which is simpler than reality and has fewer factors of variation.

One of the key goals of this research is to look into the quality of uncertainty estimation both within and outside of domain areas. Working with a synthetic dataset allows for the insertion of well-controlled noise patterns, which should be reflected in the model's heteroscedastic predictive uncertainty Malinin (2019). To make the synthetic set realistic for this task, we apply two types of Gaussian noise with non-constant variance (heteroscedasticity) to the synthetic target y_i:

- heteroscedastic Gaussian noise correlated with power, $\varepsilon_{power,i} = N(0, a \cdot y_i)$. This type of noise simulates the scenario of linear deterioration of the torque meter accuracy as power increases,
- heteroscedastic Gaussian noise correlated with true wind speed, $\varepsilon_{wind,i} = N(0, b \cdot w_i)$. Synthetic data is partitioned based on true wind speed, therefore adding the noise wind with variance linearly increasing with wind speed, simulates an increasing data uncertainty as we move from the in-domain partitions to out-of-domain ones. The goal of this approach is to capture the empirical observation that the most severe wind conditions encountered in the dataset are the most uncertain.

Here, $i = 1, \cdots, M$ stands for the i-th record, w is the true wind speed, $a = 0.025$ (at power 40 MW the standard deviation of heteroscedastic power noise is 1MW) and $b = 25$ (at wind speed 40 kn the standard deviation of heteroscedastic wind noise is 1MW). The synthetic power with noise is defined as:

$$y_i' = y_i + \varepsilon_{power,i} + \varepsilon_{wind,i}$$

Furthermore, to emulate the effect of signal intrinsic noise coming from the data gathering process, we add Gaussian white noise $N(0, \sigma)$ to the training features (sensor noise, weather hindcast errors, transmission

Feature	σ
draft_aft_telegram	0.15 m
draft_fore_telegram	0.15 m
stw	0.25 kn
diff_speed_overground	0.25 kn/3min
awind_speed_provider	0.5 kn
rcurrent_vcomp	0.05 kn
rcurrent_ucomp	0.05 kn
comb_wind_swell_wave_height	0.1 m

Table 19: Standard deviation of the added Gaussian noise per input feature.

errors to name a few sources of inherent data variability). The standard deviation per feature (Table 19) is determined using the average expected noise magnitude of these signals. The effect of the injected noise on the data variance is illustrated in Figure 14 via the correlation plots of the noisy data with the respective original signal per feature.

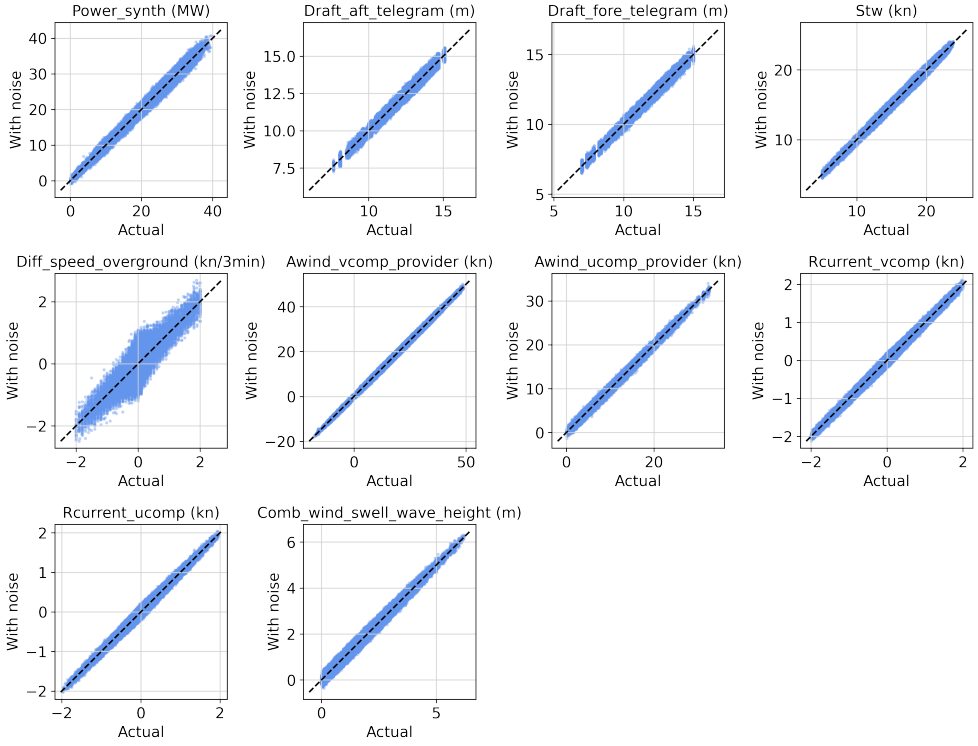


Figure 14: Correlation plots illustrating the effect of injected noise per feature.

Generalization set In order to further evaluate the generalization capability of models under research in out-of-domain regions, we introduce for the first time, the notion of a generalization set as an augmented

synthetic dataset (about 2.5 million records) by applying independently uniform sampling on input features from a predefined range of values shown in Table 20. The idea of sampling from the convex hull of the full range of possible conditions (operational and weather) is a necessary condition to assure that the measure of performance is robust. For each sample point, power labels are generated by the synthetic model which by design can cover the complete feature space. The idea is depicted in Figures 1a and 1b. Independently uniform sampling of features enables the creation of new records with feature vectors not regularly or even not physically possible to be met. For instance, in regular vessel operation, there is a low probability for a vessel to have high speed in extreme weather conditions. Beyond navigational preferences, confounding features such as apparent winds and waves is another source of spurious correlations in the dataset that could lead to a biased model.

In cases where the downstream task is implemented by a combinatorial optimization algorithm (e.g weather routing) while the model produces the cost function, it is critical that the model is unbiased across the board, even under unlikely conditions. Because the optimization algorithm actively searches the space of all feasible states, it may select a poorly modeled one as optimal and drive the entire solution in an entirely wrong direction.

Having a biased dataset apart from training unreliable models also prevents from detecting such model as performance metrics also affected from the same biases. As the test set is practically following the same (biased) distribution with the rest of the dataset, a model could show good performance while in reality would fail to generalize or even worse to properly disentagle all the causal factors received as input. Uniform sampling assures that all possible conditions are equally represented practically eliminating spurious correlation between input features. Also allows for measuring performance consistently even with the classic metrics (e.g mse, mae etc.) by providing an unbiased estimate of the performance of a model across the board.

The generalization set has two important characteristics making it suitable for model evaluation:

1. There are no correlations between the input features (both causal and spurious correlations).

2. The generalization set is suitable for evaluating model performance both in and out of domain, covering a wide range of operational conditions.

Feature	Range
speed over ground	[5, 23] kn
draft aft	[8, 15] m
draft fore	[8, 15] m
true wind speed	[0, 40] kn
relative wind angle	[0, 360] degrees
current speed	[0, 2] kn
relative current angle	[0, 360] degrees
waves	[0, 6] m

Table 20: Range of values of the input features used to create the generalization set by uniform sampling.

Format The data will be shared as several comma-separate value (CSV) files.

D.2 DESCRIPTION OF FEATURES AND TARGETS

Feature name	Units	Description	Source
draft_aft_telegram	m	Draft at stern as reported by crew in daily reports	Telegrams
draft_fore_telegram	m	Draft at bow as reported by crew in daily reports	Telegrams
stw	kn	Speed through water (i.e. relative to any currents) of the vessel as measured by speed log	Onboard sensor
diff_speed_overground	kn/3min	Acceleration of the vessel relative to ground	GPS
awind_vcomp_provider	kn	Apparent wind speed component relative to the vessel along its direction of motion	Weather provider
awind_ucomp_provider	kn	Apparent wind speed component relative to vessel perpendicular to its direction	Weather provider
rcurrent_vcomp	kn	Component of currents relative to the vessel along its direction of motion	Weather provider
rcurrent_ucomp	kn	Component of currents relative to vessel perpendicular to its direction	Weather provider
comb_wind_swell_wave_height	m	Combined wave height due to wind and sea swell	Weather provider
timeSinceDryDock	minutes	Time since the last dry dock cleaning of the vessel	Calculated
time_id	–	Run number representing time. It may be used as an index of the records	Calculated

Table 21: Description of the input features.

Feature name	Units	Description	Source
power	kW	Propeller shaft power as measured by torquemeter	Onboard sensor
power_synth	kW	Synthetic power generated by the synthetic model	Estimated

Table 22: Description of the targets.

D.3 TRAINING DETAILS

To evaluate the proposed dataset partitioning through the prism of uncertainty, we use the following methods in the form of an ensemble, that are able to capture both epistemic and aleatoric uncertainty:

- Deep ensemble of 10 variational inference neural networks (Deep Ensemble VI)
- Deep ensemble of 10 Monte-Carlo (MC) dropout (Gal & Ghahramani, 2016) neural networks (Deep Ensemble MC dropout)

- Ensemble of 10 deep neural networks (Ensemble DNN)
- A proprietary domain-constrained model is also introduced in the form of an ensemble of 10 dense neural-symbolic networks Tsamoura et al. (2021) incorporating specific domain knowledge priors derived from the physics of the problem (Ensemble Symbolic). Domain specific knowledge is encoded via known relationships between input and output features, for example the cubic relationship between speed and power. Such physics priors are integrated with the rest of the network in a neural-symbolic fashion and as a result the model is still trained end-to-end like a normal deep regression model.

Each model outputs two parameters, the predicted mean and the predicted standard deviation of the conditional Normal distribution of the target (power) given the input. The variance of the predicted means across the members of the ensemble corresponds to the epistemic uncertainty and the mean of the predicted variances across the members is the measure of aleatoric uncertainty of the ensemble Malinin et al. (2021)).

For all the methods except for the proprietary model (i.e Ensemble Symbolic) we use the same architecture: 2 hidden layers with 50 and 20 nodes and softplus activation function. The output layer has 2 nodes and a linear activation function. To satisfy the constraint of positive standard deviation the second output is fed through a softplus function and a constant 10^{-6} is added for numerical stability as proposed by Lakshminarayanan et al. (2017). For the VI method we use Bayesian inference layers with Gaussian priors. They implement the Flipout estimator Wen et al. (2018) which performs a Monte Carlo approximation of the distribution. During inference for both the Deep Ensemble VI and Deep Ensemble MC dropout we sample 10 times each member of the ensemble (100 samples in total) to estimate the epistemic uncertainty. For the Ensemble DNN and Ensemble Symbolic model we use only the members of the ensemble to estimate the epistemic uncertainty.

Furthermore, we consider the single model version of the DNN and Symbolic methods. Both versions they only capture aleatoric uncertainty. For the VI and MC dropout methods we also consider a simpler version of them by using a single seed model that is sampled 10 times during inference to capture the epistemic uncertainty.They referred as VI Ensemble (instead of Deep Ensemble) and MC dropout Ensemble respectively.

For optimization, we use the negative log likelihood loss function and the Adam optimizer with a learning rate of 10^{-4}. The number of epochs is defined by early stopping, monitoring the mean absolute error (MAE) of the dev_in set. The models are implemented in Tensorflow 2.

D.4 Additional Results

Synthetic dataset The performance metrics for the canonical partitions of the synthetic dataset and generalization set are presented in Tables 23 and 24. A single model metric $mean \pm \sigma$ is computed across the individual metric scores of all members.

For the dev and eval sets, Ensemble DNN has the best predictive performance (Table 23). Model ranking changes remarkably when considering the generalization set, with the Ensemble Symbolic having the best performance scores, showing percentage difference 18.5% in terms of RMSE from the second best model that is the Ensemble DNN. Taking into account that the performance scores on the generalization set cover the whole feature space and are unbiased by construction of the set (i.e uniform sampling eliminates operational preferences and/or spurious correlations among features), the Ensemble Symbolic is expected to be the best candidate model deployed on unseen data, in terms of robustness. Another important observation is that the percentage differences of the scores between the models are significantly higher at the generalization set. This demonstrates that the generalization set can be an useful tool for model research and selection because it amplifies potentially insignificant variations in model performance when tested in a conventional dataset split.

Regarding the metrics that jointly assess robustness and predictive uncertainty (Table 24) it is observed that Ensemble DNN has the best scores in the generalization set. Ensemble DNN is not the best model in terms of robustness (Table 23) and the fact that it takes the first place based on the retention metrics is an indication of considerable improvement of the quality of the uncertainty estimations (i.e better calibration of the predictive uncertainty with the error) in comparison to the Ensemble Symbolic. Same as before it is found that for the generalization set, the model ranking is well defined as there is a clear distinction of the scores across the models. This is not the case though for dev and eval sets, at which the top-2 models (Ensemble DNN and Ensemble VI) appear to have similar performance.

Dataset	Method	Model	RMSE (kW)				MAE (kW)				MAPE (%)			
			In	Out	Full	Gen	In	Out	Full	Gen	In	Out	Full	Gen
Dev	MC dropout	Deep ensemble	1082	1064	**1073**	1498	825	834	**830**	1000	**6.16**	7.63	6.91	23.00
	VI	Deep ensemble	**1081**	1068	1075	1446	**823**	838	**830**	975	6.38	7.70	7.04	23.19
	DNN	Ensemble	1088	**1062**	1075	1427	827	**832**	**830**	953	6.24	**7.49**	**6.87**	**21.21**
	Symbolic	Ensemble	1132	1126	1129	**1204**	851	864	858	**873**	7.32	8.94	8.13	27.33
	MC dropout	Ensemble	$1091_{\pm 3}$	$1074_{\pm 4}$	$1082_{\pm 3}$	$1526_{\pm 84}$	$832_{\pm 2}$	$842_{\pm 3}$	$837_{\pm 3}$	$1023_{\pm 35}$	$6.35_{\pm 0.20}$	$7.69_{\pm 0.09}$	$7.03_{\pm 0.13}$	$24.54_{\pm 1.04}$
	VI	Ensemble	$1085_{\pm 4}$	$1074_{\pm 6}$	$1079_{\pm 4}$	$1458_{\pm 38}$	$825_{\pm 3}$	$842_{\pm 5}$	$833_{\pm 2}$	$985_{\pm 22}$	$6.42_{\pm 0.18}$	$7.75_{\pm 0.07}$	$7.08_{\pm 0.12}$	$23.72_{\pm 1.26}$
	DNN	Single	$1096_{\pm 7}$	$1081_{\pm 10}$	$1089_{\pm 8}$	$1487_{\pm 52}$	$834_{\pm 5}$	$846_{\pm 6}$	$840_{\pm 6}$	$1008_{\pm 29}$	$6.34_{\pm 0.26}$	$7.67_{\pm 0.10}$	$7.01_{\pm 0.15}$	$23.66_{\pm 1.62}$
	Symbolic	Single	$1134_{\pm 2}$	$1129_{\pm 5}$	$1132_{\pm 3}$	$1213_{\pm 27}$	$853_{\pm 1}$	$866_{\pm 4}$	$860_{\pm 2}$	$879_{\pm 19}$	$7.33_{\pm 0.05}$	$8.96_{\pm 0.06}$	$8.15_{\pm 0.05}$	$27.54_{\pm 1.73}$
Eval	MC dropout	Deep ensemble	**1069**	1111	1090	1498	814	859	837	1000	6.26	6.95	6.59	23.00
	VI	Deep ensemble	**1069**	1104	**1086**	1446	**813**	854	**834**	975	6.24	6.92	6.58	23.19
	DNN	Ensemble	1076	**1099**	1087	1427	818	**851**	**834**	953	**6.13**	**6.91**	**6.52**	**21.21**
	Symbolic	Ensemble	1117	1133	1125	**1204**	841	866	854	**873**	7.25	7.29	7.27	27.33
	MC dropout	Ensemble	$1078_{\pm 4}$	$1122_{\pm 6}$	$1100_{\pm 5}$	$1526_{\pm 84}$	$822_{\pm 3}$	$868_{\pm 5}$	$845_{\pm 4}$	$1023_{\pm 35}$	$6.34_{\pm 0.08}$	$7.05_{\pm 0.08}$	$6.70_{\pm 0.07}$	$24.54_{\pm 1.04}$
	VI	Ensemble	$1072_{\pm 3}$	$1109_{\pm 6}$	$1090_{\pm 4}$	$1458_{\pm 38}$	$815_{\pm 2}$	$858_{\pm 4}$	$837_{\pm 3}$	$985_{\pm 22}$	$6.28_{\pm 0.13}$	$6.96_{\pm 0.06}$	$6.62_{\pm 0.09}$	$23.72_{\pm 1.26}$
	DNN	Single	$1084_{\pm 6}$	$1116_{\pm 18}$	$1100_{\pm 12}$	$1487_{\pm 52}$	$825_{\pm 5}$	$864_{\pm 14}$	$844_{\pm 10}$	$1008_{\pm 29}$	$6.24_{\pm 0.13}$	$7.04_{\pm 0.17}$	$6.64_{\pm 0.14}$	$23.66_{\pm 1.62}$
	Symbolic	Single	$1120_{\pm 2}$	$1137_{\pm 5}$	$1128_{\pm 3}$	$1213_{\pm 27}$	$843_{\pm 1}$	$869_{\pm 4}$	$856_{\pm 2}$	$879_{\pm 19}$	$7.26_{\pm 0.04}$	$7.30_{\pm 0.03}$	$7.28_{\pm 0.03}$	$27.54_{\pm 1.73}$

Table 23: Predictive performance for the canonical partitions of the synthetic dataset and the generalization set. One standard deviation is quoted for the single seed results.

Dataset	Method	Model	R-AUC $*10^5$				F1-AUC				F1@95%			
			In	Out	Full	Gen	In	Out	Full	Gen	In	Out	Full	Gen
Dev	MC dropout	Deep ensemble	4.17	4.66	4.40	4.97	0.479	0.427	0.454	0.477	0.576	0.545	0.561	0.576
	VI	Deep ensemble	**4.03**	4.54	**4.26**	4.32	0.491	**0.433**	**0.465**	0.506	0.579	0.544	0.563	0.582
	DNN	Ensemble	4.08	**4.50**	**4.26**	**4.20**	**0.492**	**0.433**	**0.465**	**0.509**	**0.581**	**0.549**	**0.565**	0.595
	Symbolic	Ensemble	4.90	5.49	5.17	4.41	0.475	0.423	0.452	0.494	0.571	0.539	0.555	**0.596**
	MC dropout	Ensemble	$4.23_{\pm 0.05}$	$4.70_{\pm 0.04}$	$4.45_{\pm 0.04}$	$5.39_{\pm 0.45}$	$0.485_{\pm 0.002}$	$0.428_{\pm 0.002}$	$0.459_{\pm 0.001}$	$0.481_{\pm 0.013}$	$0.573_{\pm 0.002}$	$0.541_{\pm 0.003}$	$0.557_{\pm 0.003}$	$0.560_{\pm 0.012}$
	VI	Ensemble	$4.05_{\pm 0.02}$	$4.58_{\pm 0.04}$	$4.29_{\pm 0.02}$	$4.53_{\pm 0.31}$	$0.490_{\pm 0.001}$	$0.432_{\pm 0.001}$	$0.464_{\pm 0.002}$	$0.499_{\pm 0.011}$	$0.578_{\pm 0.001}$	$0.544_{\pm 0.002}$	$0.562_{\pm 0.001}$	$0.578_{\pm 0.012}$
	DNN	Single	$4.16_{\pm 0.06}$	$4.68_{\pm 0.08}$	$4.41_{\pm 0.07}$	$5.27_{\pm 0.63}$	$0.488_{\pm 0.003}$	$0.430_{\pm 0.002}$	$0.461_{\pm 0.003}$	$0.471_{\pm 0.019}$	$0.576_{\pm 0.003}$	$0.544_{\pm 0.002}$	$0.560_{\pm 0.003}$	$0.568_{\pm 0.012}$
	Symbolic	Single	$4.92_{\pm 0.03}$	$5.52_{\pm 0.06}$	$5.20_{\pm 0.04}$	$4.55_{\pm 0.26}$	$0.475_{\pm 0.001}$	$0.423_{\pm 0.001}$	$0.452_{\pm 0.001}$	$0.490_{\pm 0.003}$	$0.570_{\pm 0.001}$	$0.538_{\pm 0.001}$	$0.554_{\pm 0.001}$	$0.594_{\pm 0.006}$
Eval	MC dropout	Deep ensemble	4.11	4.80	4.47	4.97	0.487	0.432	0.459	0.477	0.587	0.548	0.568	0.576
	VI	Deep ensemble	**3.97**	**4.59**	**4.29**	4.32	**0.497**	**0.441**	**0.470**	0.506	**0.589**	**0.549**	**0.570**	0.582
	DNN	Ensemble	4.02	4.60	4.32	**4.20**	**0.497**	0.439	0.469	**0.509**	0.588	**0.549**	0.569	0.595
	Symbolic	Ensemble	4.82	5.42	5.09	4.41	0.484	0.426	0.458	0.494	0.579	0.548	0.564	**0.596**
	MC dropout	Ensemble	$4.17_{\pm 0.05}$	$4.87_{\pm 0.06}$	$4.54_{\pm 0.06}$	$5.39_{\pm 0.45}$	$0.491_{\pm 0.001}$	$0.435_{\pm 0.001}$	$0.463_{\pm 0.001}$	$0.481_{\pm 0.013}$	$0.582_{\pm 0.002}$	$0.542_{\pm 0.003}$	$0.561_{\pm 0.003}$	$0.560_{\pm 0.012}$
	VI	Ensemble	$4.00_{\pm 0.02}$	$4.64_{\pm 0.06}$	$4.33_{\pm 0.04}$	$4.53_{\pm 0.31}$	$0.496_{\pm 0.002}$	$0.440_{\pm 0.002}$	$\mathbf{0.470_{\pm 0.002}}$	$0.499_{\pm 0.011}$	$0.588_{\pm 0.002}$	$0.546_{\pm 0.002}$	$0.568_{\pm 0.002}$	$0.578_{\pm 0.012}$
	DNN	Single	$4.09_{\pm 0.05}$	$4.84_{\pm 0.19}$	$4.49_{\pm 0.14}$	$5.27_{\pm 0.63}$	$0.493_{\pm 0.004}$	$0.433_{\pm 0.005}$	$0.464_{\pm 0.005}$	$0.471_{\pm 0.019}$	$0.584_{\pm 0.004}$	$0.543_{\pm 0.007}$	$0.564_{\pm 0.005}$	$0.568_{\pm 0.012}$
	Symbolic	Single	$4.84_{\pm 0.02}$	$5.46_{\pm 0.05}$	$5.13_{\pm 0.03}$	$4.55_{\pm 0.26}$	$0.483_{\pm 0.001}$	$0.425_{\pm 0.001}$	$0.457_{\pm 0.0}$	$0.490_{\pm 0.003}$	$0.579_{\pm 0.0}$	$0.547_{\pm 0.002}$	$0.563_{\pm 0.001}$	$0.594_{\pm 0.006}$

Table 24: Retention performance for the canonical partitions of the synthetic dataset and the generalization set. One standard deviation is quoted for the single seed results.

Real dataset Due to a lack of knowledge of the actual data generation process, it is not possible to generate an analogous test set to the generalization set, which is only relevant in the 'synthetic world.' As a result, model evaluation is limited to canonical partitions. Furthermore, we compare the results for the two datasets, namely synthetic and real ones, and we provide an interpretation of the observed behaviors.

The performance scores are shown in the Tables 25 and 26. Deep Ensemble VI has the best performance across all metrics. This is not the case for the synthetic dataset at which Ensemble DNN has the best scores overall. This outcome is of no surprise, as both methods (along with Deep Ensemble MC dropout) are similar therefore deviations on their ranking are to be expected when working with different datasets.

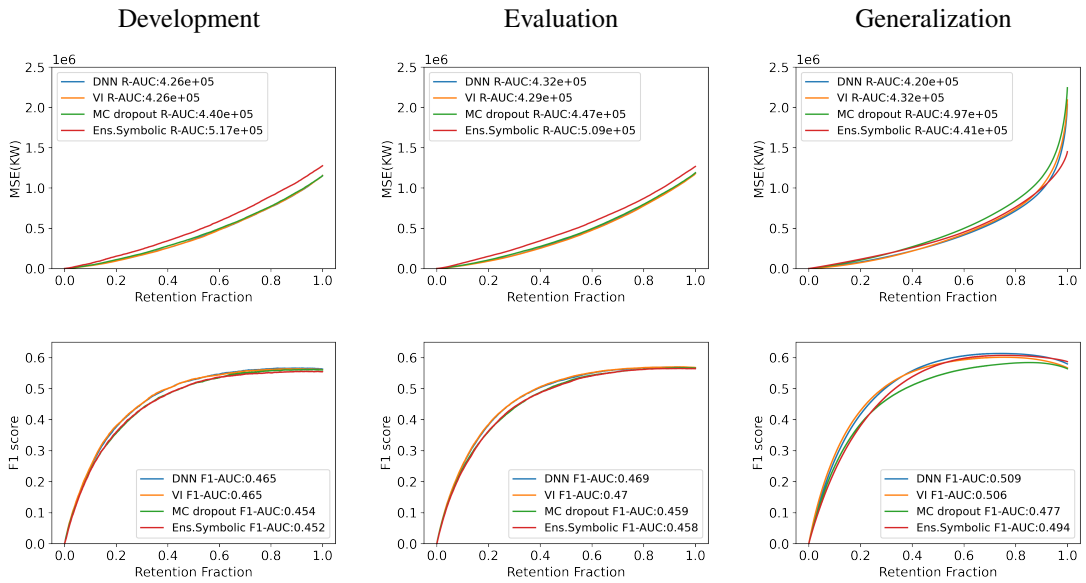


Figure 15: Retention curves for the synthetic development, evaluation and generalization sets. VI and MC dropout refer to the deep ensemble technique while DNN and Symbolic correspond to the ensemble setting.

For the Ensemble Symbolic, it is worth noting that it has the lowest predictive performance in the dev and eval sets, which is consistent with the results in the synthetic data. Furthermore, when compared with the corresponding synthetic partitions, it is discovered to have a larger performance drop compared to the best model. This outcome is attributed to the limited expressivity of the Ensemble Symbolic, resulting in a more pronounced performance degradation in the real data. On the other hand, the synthetic generalization set revealed that the Ensemble Symbolic is the best candidate in terms of robustness across all possible operational conditions. Such domain constrained models have the advantage of unbiased performance across all possible conditions (operational or weather) making them good candidates for active performance optimization tasks (such as vessel-specific weather routing).

Dataset	Method	Model	RMSE (kW)			MAE (kW)			MAPE (%)		
			In	Out	Full	In	Out	Full	In	Out	Full
Dev	MC dropout	Deep ensemble	1269	1501	1389	855	1067	**961**	5.42	**7.39**	6.40
	VI	Deep ensemble	**1264**	1514	1395	**848**	1074	**961**	**5.29**	7.44	**6.37**
	DNN	Ensemble	1285	**1484**	**1388**	868	**1066**	967	5.43	7.49	6.46
	Symbolic	Ensemble	1405	1630	1522	971	1181	1076	6.41	8.98	7.69
	MC dropout	Ensemble	$1291_{\pm 21}$	$1540_{\pm 56}$	$1422_{\pm 30}$	$874_{\pm 17}$	$1098_{\pm 39}$	$986_{\pm 17}$	$5.63_{\pm 0.22}$	$7.80_{\pm 0.29}$	$6.72_{\pm 0.14}$
	VI	Ensemble	$1276_{\pm 12}$	$1537_{\pm 41}$	$1413_{\pm 24}$	$858_{\pm 9}$	$1093_{\pm 26}$	$975_{\pm 14}$	$5.39_{\pm 0.07}$	$7.65_{\pm 0.19}$	$6.53_{\pm 0.11}$
	DNN	Single	$1318_{\pm 51}$	$1547_{\pm 63}$	$1438_{\pm 44}$	$893_{\pm 41}$	$1113_{\pm 46}$	$1003_{\pm 30}$	$5.74_{\pm 0.32}$	$8.03_{\pm 0.42}$	$6.88_{\pm 0.27}$
	Symbolic	Single	$1416_{\pm 8}$	$1654_{\pm 53}$	$1540_{\pm 25}$	$980_{\pm 11}$	$1199_{\pm 54}$	$1089_{\pm 23}$	$6.46_{\pm 0.10}$	$9.09_{\pm 0.38}$	$7.77_{\pm 0.18}$
Eval	MC dropout	Deep ensemble	1248	1925	1622	850	1389	1119	5.54	8.29	6.91
	VI	Deep ensemble	**1243**	**1895**	**1602**	**842**	**1356**	**1098**	**5.38**	**8.09**	**6.73**
	DNN	Ensemble	1264	1928	1630	863	1414	1138	5.48	8.75	7.12
	Symbolic	Ensemble	1393	2341	1926	964	1744	1354	6.37	10.66	8.52
	MC dropout	Ensemble	$1271_{\pm 18}$	$1954_{\pm 47}$	$1649_{\pm 26}$	$868_{\pm 16}$	$1416_{\pm 39}$	$1142_{\pm 20}$	$5.72_{\pm 0.20}$	$8.54_{\pm 0.33}$	$7.13_{\pm 0.20}$
	VI	Ensemble	$1255_{\pm 11}$	$1916_{\pm 32}$	$1620_{\pm 20}$	$852_{\pm 9}$	$1377_{\pm 39}$	$1114_{\pm 22}$	$5.46_{\pm 0.06}$	$8.26_{\pm 0.33}$	$6.86_{\pm 0.18}$
	DNN	Single	$1296_{\pm 47}$	$1985_{\pm 111}$	$1677_{\pm 61}$	$887_{\pm 39}$	$1462_{\pm 91}$	$1175_{\pm 47}$	$5.74_{\pm 0.28}$	$9.15_{\pm 0.62}$	$7.44_{\pm 0.38}$
	Symbolic	Single	$1403_{\pm 8}$	$2366_{\pm 304}$	$1948_{\pm 179}$	$973_{\pm 10}$	$1770_{\pm 253}$	$1371_{\pm 123}$	$6.42_{\pm 0.10}$	$10.84_{\pm 1.42}$	$8.63_{\pm 0.70}$

Table 25: Predictive performance for the canonical partitions of the real dataset. One standard deviation is quoted for the single seed results.

Dataset	Method	Model	R-AUC $*10^5$			F1-AUC			F1@95%		
			In	Out	Full	In	Out	Full	In	Out	Full
Dev	MC dropout	Deep ensemble	4.52	7.21	6.05	0.510	**0.469**	0.486	0.618	0.536	0.577
	VI	Deep ensemble	**4.29**	7.22	**5.81**	**0.521**	0.460	**0.493**	**0.625**	**0.541**	**0.584**
	DNN	Ensemble	4.71	**6.80**	5.84	0.514	0.441	0.477	0.616	0.521	0.570
	Symbolic	Ensemble	6.55	10.80	8.48	0.453	0.362	0.419	0.557	0.472	0.511
	MC dropout	Ensemble	$4.75_{\pm 0.32}$	$8.58_{\pm 0.99}$	$6.84_{\pm 0.48}$	$0.510_{\pm 0.009}$	$0.450_{\pm 0.014}$	$0.477_{\pm 0.006}$	$0.610_{\pm 0.009}$	$0.525_{\pm 0.012}$	$0.568_{\pm 0.005}$
	VI	Ensemble	$4.40_{\pm 0.17}$	$7.85_{\pm 0.52}$	$6.18_{\pm 0.24}$	$0.518_{\pm 0.007}$	$0.448_{\pm 0.012}$	$0.484_{\pm 0.004}$	$0.620_{\pm 0.005}$	$0.535_{\pm 0.009}$	$0.579_{\pm 0.003}$
	DNN	Single	$5.00_{\pm 0.57}$	$7.84_{\pm 0.62}$	$6.61_{\pm 0.51}$	$0.505_{\pm 0.016}$	$0.416_{\pm 0.020}$	$0.459_{\pm 0.012}$	$0.605_{\pm 0.020}$	$0.503_{\pm 0.019}$	$0.555_{\pm 0.012}$
	Symbolic	Single	$6.99_{\pm 0.22}$	$11.33_{\pm 1.14}$	$8.99_{\pm 0.53}$	$0.449_{\pm 0.006}$	$0.355_{\pm 0.019}$	$0.412_{\pm 0.009}$	$0.552_{\pm 0.010}$	$0.465_{\pm 0.019}$	$0.509_{\pm 0.009}$
Eval	MC dropout	Deep ensemble	4.34	**13.59**	9.28	0.513	0.394	0.451	0.621	0.459	0.544
	VI	Deep ensemble	**4.15**	14.07	**9.13**	**0.525**	**0.398**	**0.467**	**0.627**	**0.477**	**0.557**
	DNN	Ensemble	4.50	14.00	9.52	0.517	0.387	0.451	0.616	0.428	0.528
	Symbolic	Ensemble	6.39	22.56	13.56	0.455	0.267	0.383	0.558	0.320	0.447
	MC dropout	Ensemble	$4.57_{\pm 0.27}$	$14.46_{\pm 1.28}$	$10.00_{\pm 0.59}$	$0.511_{\pm 0.010}$	$0.383_{\pm 0.022}$	$0.441_{\pm 0.013}$	$0.610_{\pm 0.010}$	$0.441_{\pm 0.026}$	$0.530_{\pm 0.015}$
	VI	Ensemble	$4.26_{\pm 0.16}$	$14.58_{\pm 1.02}$	$9.57_{\pm 0.56}$	$0.521_{\pm 0.007}$	$0.383_{\pm 0.017}$	$0.455_{\pm 0.008}$	$0.621_{\pm 0.006}$	$0.467_{\pm 0.027}$	$0.549_{\pm 0.014}$
	DNN	Single	$4.78_{\pm 0.52}$	$15.68_{\pm 2.16}$	$10.97_{\pm 1.21}$	$0.506_{\pm 0.017}$	$0.364_{\pm 0.020}$	$0.425_{\pm 0.024}$	$0.603_{\pm 0.021}$	$0.416_{\pm 0.033}$	$0.515_{\pm 0.023}$
	Symbolic	Single	$6.81_{\pm 0.23}$	$27.85_{\pm 7.33}$	$17.51_{\pm 4.27}$	$0.449_{\pm 0.006}$	$0.254_{\pm 0.037}$	$0.360_{\pm 0.027}$	$0.553_{\pm 0.009}$	$0.322_{\pm 0.068}$	$0.447_{\pm 0.028}$

Table 26: Retention performance for the canonical partitions of the real dataset. One standard deviation is quoted for the single seed results.

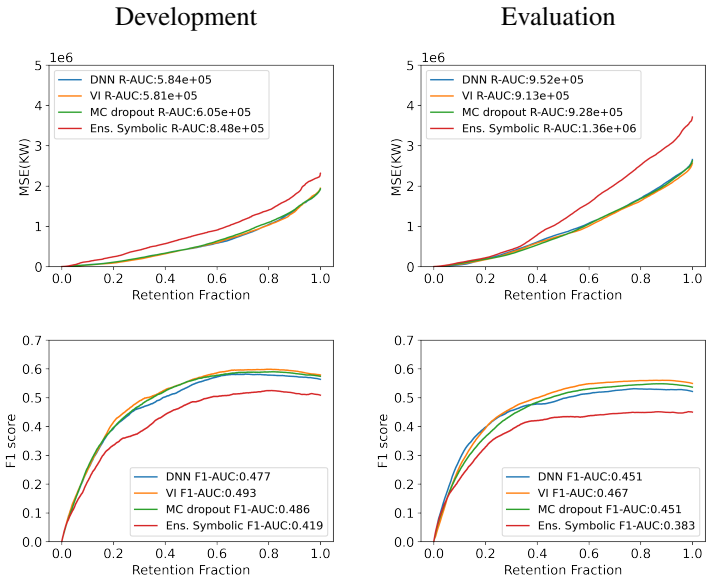


Figure 16: Retention curves for the real development, evaluation sets. VI and MC dropout refer to the deep ensemble technique while DNN and Symbolic correspond to the ensemble setting.