MyWave

Inter-comparison of MetOffice and CNMCA ensemble prediction systems for Mediterranean Sea

Reference: MyWave-D3.4

Project N°: FP7-SPACE-2011-284455	**Work programme topic:** SPA.2011.1.5.03 – R&D to enhance future GMES applications in the Marine and Atmosphere areas
Start Date of project: 01.01-2012	**Duration**: 36 Months

WP leader: Luigi Cavaleri **Issue:**

Contributors: Paolo Pezzutto, Christopher Bunney, Luigi Cavaleri

MyWave version scope: all project versions

Approval Date: **Approver:**

Dissemination level: Project

DOCUMENT VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:			
Distribution			
	Christopher Bunney (WP3) Luigi Cavaleri (WP3 leader) Ø. Saetra (Project coordinator) Andrew Saulter (WP4) Stefano Sebastianelli (WP3) Lucio Torrisi (WP3)		

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
0.1	11/09/14	all	First draft of document	Paolo Pezzutto	Luigi Cavaleri,
1.0		all	Document finalization		

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY AND ABREVIATIONS (ORDER!!!)

CNMCA	Centro Nazionale di Meteorologia e Climatologia Aeronautica (National Centre of Aeronautics Weather Science and Climatology)
CORR	Pearson Correlation Coefficient
DF	Deterministic Forecast / Control Member
EM	Ensemble Mean
EPS	Ensemble Prediction System
ES	Ensemble Spread
H_S	Significant Wave Height
MAE	Mean Absolute Error
Q-Q	Quantile-Quantile (plot)
REV	Relative Economic Value
RMSE	Root of Mean Squared Error
ROC	Receiver Operating Characteristic
SI	Scatter Index
SIVAR	Unbiased Scatter Index
SNRMSE	Symmetric Normalized RMSE
SSL	Slope of symmetric linear fit
STD	Standard deviation
T_m	Mean Wave Period
θ_m	Mean Wave Direction
U_{10}	Wind speed (10 m above ground)
u_{10}	Eastward component of wind speed (10 m above ground)

UKMO	U. K. MetOffice
v_{10}	Northward component of wind speed (10 m above ground)
VAR	(Estimate of) variance

APPLICABLE AND REFERENCE DOCUMENTS

Applicable Documents

	Ref	Title	Date / Issue
DA 1	MyWave-A1	MyWave: Annex I – "Description of Work"	September 2011

Reference Documents

	Ref	Title	Date / Issue
DR 1	MyWave-D3.1	Met Office Wave Model Ensemble Prediction Systems in the 'Atlantic-Euro Zone' and USAM-CNMCA 'Nettuno' Ensemble Prediction System in the Mediterranean Sea	11 Jun 2013
DR 2	MyWave-D3.2	Availability of buoy and satellite data for wave comparison in the Italian Seas	20 Feb 2013
DR 3	MyWave-D4.2a	Proposal of metrics for user focused verification of deterministic wave prediction systems	02 Oct 2013
DR 4	MyWave-D4.2b	Proposal of metrics for developer and user focused verification of wave ensemble prediction systems	02 Oct 2013

I INTRODUCTION

Within the framework of MyWave project, work package n°3, U.K. MetOffice and CNMCA (Italian Meteorological Service) have developed and implemented two independent ensemble prediction systems for the Mediterranean Sea. The two systems, later addressed as UKMO and Nettuno, are described in DR 1.

The main purpose of present report is the description of the tasks related to the verification and inter-comparison of Nettuno and UKMO EPS. The validation is based on the observations dataset reported in DR 2, which cover the six month period from July 1 to December 31 , 2013. Proceeding from DR 3 and DR 4, the basic principles on which this verification is based on is that the forecasts are validated "as they are", that is without any calibration or other kind of post processing based on the same verification scores (*e.g.* bias correction). This does not exclude that ensemble members are post-processed in order to account for observation uncertainty, since these are based on knowledge preceding the forecast/validation period.

In chapter II we describe the measurement uncertainty estimations and their use for EPS verification purposes. The aim of chapter III is the assessment of the behaviour of the ensemble mean in the framework of adopting it as a *deterministic* forecast, instead of the control member. Note that Nettuno control member is the actual operational forecast for the Italian Meteorological Service. Chapter IV investigates and compares the ensemble spread characteristics of Nettuno and UKMO systems. The last chapter (V) describes the scores related to the capacity of both EPS in predicting events probability, in term of reliability and resolution, presenting also two metrics particularly suited for forecast users. Finally we draw the conclusions in VI .

II DATA TREATMENT

Observations are affected by errors. In a statistical sense, this means that a single observation is a realization of the convolution of the true climate distribution and the error distribution. The verification of an ensemble system must take this into account. In fact, the exclusion of errors in the validation process would indicate a somehow “fake” reliability for the considered Ensemble Prediction System (EPS). For example, if the EPS were perfectly reliable, then it would lack of reliability if verified against pure observations. Therefore, in the present work, wherever necessary, we post-process the ensemble members in order to account for the related measurement errors. The term “necessary” implies that not all the verification measures presented herein take advantage of this treatment. The verification metrics for which the measurement uncertainty must be accounted for include probabilistic analyses of the whole EPS only, not the verification of control members or ensemble mean.

The mentioned post-processing is made in two ways: adding noise to the ensemble members, or dressing the forecast ensemble distributions. It is expected that the first method converges to the second one increasing the number of data. Anyhow, different verification measures require different approaches. In the next chapters, we indicate which one of the two methods is adopted by appending the words “added noise” or “dressed”. The “dressing” procedure is hereby to be intended member by member (Figure 1). We couple each *EPS* realization with a dedicated *pdf*. The dressed ensemble forecast *pdf* is the average of the single ones (black shape in Figure 1).

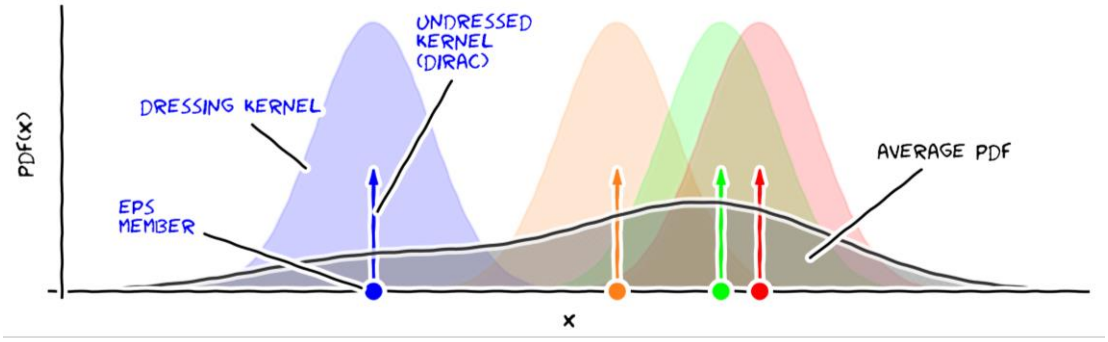


Figure 1: Each EPS *member is dresses with observation uncertainty. Black curve is the average* pdf *which include* ES *and observations uncertainties.*

In the case of EPS dressing the uncertainty for all members is equal and coincides with the one associated to the ensemble mean. All dressing kernels are gaussian. Please note that wherever noise is added to an EPS member, it is a realization of the distribution illustrated in Figure 1. The uncertainty magnitudes, *i.e.* the widths of the Gaussian bells considered in the present work, are described and defined in the following.

II.1 Representation Errors

The representation error conundrum generates when comparing a gridded forecast to the measurements obtained by the observation system (buoys and altimeters, in case of waves).

The issue arises by the difference in the scale of the two systems (see *e.g.* Bidlot and Holt, 2006). As in case of numbers, where the concept representation errors is linked to the number of digits, or the machine precision, in the verification framework, the error has to deal with the observations position which is usually a fraction of the model grid discretization.

Each kind of resampling based on interpolation or best fitting with functions which carry no knowledge of the physics, does undoubtedly invent something. It can be shown that the particular choice among the acceptable methods does not affect the representation errors. Thus we simply choose to couple the nearest-neighbours. A common practice adopted for smoothing representation errors of satellite measurements is the averaging of neighbouring satellite data, obtaining one value per model cell (or grid point). In NETTUNO's case the space grid is on the order of 5 km, while UKMO's one is approximately 7 km. Since the spacing of all the altimeters samples is similar (7 km), then each grid point cannot see more than one neighbour, and we adopt no averaging. We just take care to avoid double counting of samples. If one observation is located at equal distance from two (or more) forecast points, we choose the left one, that is the forecast value corresponding to the point with the lower time, latitude and longitude.

II.2 Observations Errors

A survey has been undertaken to map prior knowledge on instrument by instrument uncertainties. Results are summarized in Table 1 and Table 2 for magnitudes regarding wave and wind measurement, respectively. As can be seen in footnotes, most of the survey results are extracted from user manuals.

It is the writer opinion that some of these reported uncertainties should not be kept as truth. For example, in the case of buoys measurements, Janssen et al. (2003) found errors on the order of 12% of mean significant wave height for the Global Technology System datasets. This value has been used for validating ECMWF wave EPS (Saetra, et al., 2004). The uncertainties indicated in Table 1 are fairly below this order of magnitude. Later on Janssen, et al. (2007) performed triple colocation on ENVISAT, ERS-2 and buoys (GTS) wave height data, covering mostly the northern hemisphere. For the period August 2002-September 2003, they found that the buoys uncertainty was on the order of 8% of the H_S magnitude. For the two altimeters the error were about 6.1% and 6.4%, respectively. Abdalla, et al. (2011) report similar values for Jason-2, Cryosat and Saral-Altika.

In this report a value of 10% is kept for buoy measurements relative error, and 6% for altimeter observation. Mean wave period errors are reported to be on the order of fraction of seconds, or well below the H_S values (Table 1). The same applies to mean wave direction uncertainties. This is hardly to be trusted, since the evaluation of this two quantities implicates ratios of two integrals of the same kind involved in the estimation of H_S (where derived from directional spectra). The mean wave period is evaluated as the ratio of zero-th versus first spectral moments, where the m-th is computed as:

$$I_m = \sum_{i=1}^{N} \sum_{j=1}^{M} f_i^{\,m} E_{ij} \qquad (1)$$

where E_{ij} is the amount of energy density in the i-th frequency band centred at f_i, and j-th direction bin. The only source of uncertainty is the energy content. Thus the uncertainty associated to the spectral moments and to their combinations has to deal with the distribution of energy and its uncertainty.

Table 1: Uncertainties of wave magnitude measurements. Result of survey.

Instrument	Model	Source	Significant wave height [m]	Mean period [s]	Mean direction [deg N]
Altimeter	Jason-2	RADS	0.13 5.4%[1]	-	-
	Cryosat-2	RADS	>0.13 5.4%[1]	-	-
	Saral-Altika	RADS	0.13 5.4%[1]	-	-
Buoy	RON		0.04[2] <2%[3]	0.4[3] 1%[4]	20[3] 1.0[4]
	ARPAL		0.05[5]	0.15[6]	1[6]
	MyOcean Catalogue	OceanSITES	0.5 (≤ 5m) 10% (> 5m)[6]	0.5[9]	10 20[9]
	IPC-CAMERI	Directional Waverider MkIII (DWR-MkIII)	<0.5% <1.0%[10]	-	0.5[7]

An average value of 1s has been trusted by Saetra, et al., (2004) for validating ECMWF-EPS wave periods co-located with GTS buoy data. This magnitude is compatible with zero (down or up) crossing evaluation, where the common instrument time series resolution is about 1Hz. However it does not account for any buoy dynamics influence.

The mean period error would therefore be not systematic nor strictly proportional to T_m. For sake of convenience, we consider a 10% of uncertainty on T_m.

Mean wave direction is evaluated via spectral analysis only, and estimated by a ratio of two integrals. Hence, it is subject to the same consideration made for the mean periods so far. In this case, we consider a pure systematic uncertainty of 10 degrees (see the communicated values in Table 1). According to Table 2 wind speed values are accepted within relative uncertainty ranges of 12% for altimeters and 10% for all the buoy raw series.

[1] Abdalla, et al., (2011) integrated by S. Abdalla personal communication
[2] Work, s.d.
[3] http://64.246.9.130:5555/ron/boe.html
[4] Work, s.d.
[5] Fugro Oceanor, 2014
[6] WMO, 2008
[7] Datawell, (2012) (The first error refers to measured value after calibration, the second to measured value after 3 years)

Table 2: Uncertainties of wind magnitudes measurements. Results of survey.

Instrument	Model	Source	Wind speed [m/s]	Wind direction [deg N]	Wind *u* component [m/s]	Wind *v* component [m/s]
Altimeter	Jason-2	RADS	±1.0 11.9%[1]	-	-	-
	Cryosat-2	RADS	>±1.0 11.9%[1]	-	-	-
	Saral-Altika	RADS	1.0 11.9%[1]	-	-	-
Scatterometer	OSI SAF ASCAT-A Coastal	Metop-A ASCAT	-	-	1.50[8]	1.65
	OSI SAF ASCAT-B 25-km	Metop-B ASCAT	-	-	1.40	1.50
	OSI SAF OSCAT 50-km	Oceansat-2 OSCAT	-	-	1.90	1.34
Buoy	RON		-	-	1.15	1.17
	ARPAL		0.3 1%[9] 2%[10]	3[7]	-	-
	MyOcean Catalogue	OceanSITES	1.0 (≤5m/s) 10% (>5m/s)[9]	15[9]	-	-
	IPC-CAMERI	Directional Waverider MkIII (DWR-MkIII)	0.17[11]	<3	-	-

[8] Marseille, et al., (2014) and Gert-Jan Marseille personal communication.

[9] YOUNG, (2005)

[10] Gill, (2013)

[11] Vasala, (2002)

III BEHAVIOUR OF DETERMINISTIC FORECAST

The aim of this section is the assessment of the behaviour of the ensemble mean (EM) in the framework of adopting it as a *deterministic* forecast. At first we analyse the general behaviour through classical scatter diagrams (III.1). Then we start looking into the details of the forecast sample distributions, comparing their shapes with respect to the measurement ones (III.2). In paragraph III.3 we recall the control member forecasts as a baseline for the evaluation of the EM skill improvement (Taylor diagrams). Finally we propose a number of error measures for the assessment of the skill of control member, ensemble mean and ensemble range along forecast lead time (III.4).

III.1 General

We hereby propose a first comparison of measurements with co-located deterministic model predictions. Scatter-diagrams (usually rendered as two dimensional histograms) provide visual synthetic information on deviances from a desired 1 to 1 relationship. These are a quantitative, but very general, tool for the verification of a *deterministic* forecast. We here assess the behaviour of the ensemble mean (EM).

Figure 2 present four scatter-diagrams accounting for forecast significant wave heights matched with buoy (top) and altimeter (bottom) observations. Left panels refer to Nettuno EPS, while right panels account for UKMO system. Both systems are verified against exactly the same dataset. All diagrams cover the whole testing period: 1 July 2013 to 31 December 2013, and provide a cumulative validation for 48 hours. Figure 3 reproduces the same things for wind speed at 10 m above ground, namely U_{10}. In this case forecast is compared with scatterometer observations too. Together with the number of data, each plot is supplied with a set of statistical measures (see Glossary and Annex A for definitions). The variation along forecast lead time of these skill is presented in § III.4 .

In Figure 2 the whole skill set is quite similar for both forecast systems, except for three measures: SSL (symmetric slope), BIAS and FVAR (forecast distribution variance). The first one is the slope of the thick black line, *i.e.* the orthogonal least squares fitting assuming a direct proportionality between observations and forecast. In all cases Nettuno presents higher SSL than UKMO, and higher than unity. This indicates a tendency to overestimate significant wave heights (Figure 2), whereas UKMO tendency is somehow opposite. This difference is more pronounced at the buoy sites, and almost negligible if the two models are compared against altimeter measurements.

BIAS is the average difference of forecast and observations. Nettuno's values remain on the order of -2 cm for both observation systems, while the UKMO -1 cm against altimeters drops down to -8 cm against buoys. Last important thing to notice is the width of the forecast distribution (FVAR) with respect to the observation ones (OVAR). The UKMO ensemble mean variability is slightly lower than the measurement one, the ratios being on the order of 0.96. On the other hand, the width of the Nettuno wave height distribution is a bit higher than expected with ratios of approximately 1.25 and 1.15, respectively against buoy and altimeters.

The differences between the two forecast systems are more pronounced at the buoy locations. Note that altimeter data are taken on the whole Mediterranean basin, while all the buoys are distributed along coastal regions. For this reason, it can be stated that the two forecasts behave in similar manner in the open sea, while their differences are somehow emphasized close to the coastal regions.

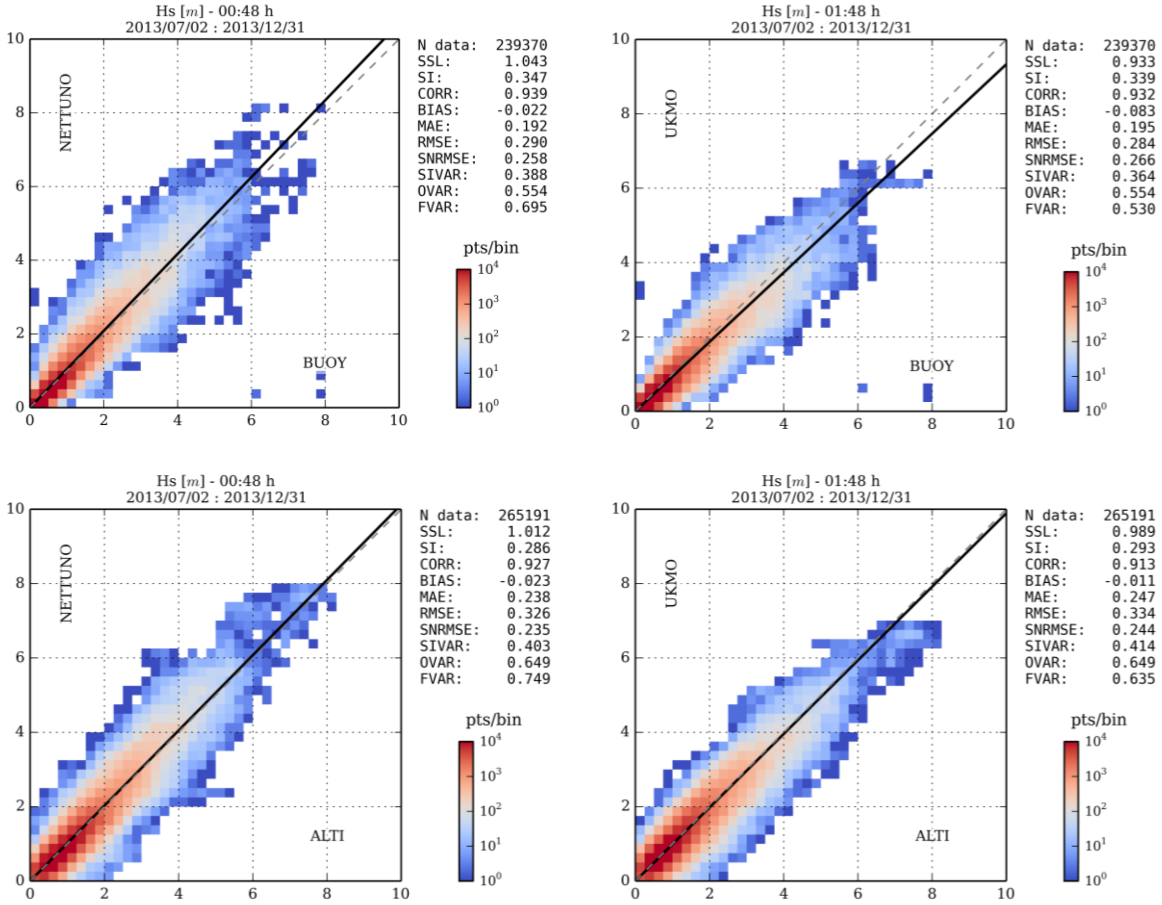


Figure 2: Ensemble Mean Forecast Vs. Observations histograms for significant wave height [m]. Upper panels: buoy locations, lower panel: altimeters. Left: Nettuno; right: UKMO. Dashed line: main diagonal; continuous line: best symmetric fit. Coverage: Mediterranean Sea.

Validation against altimeters and buoy wind speed (see Figure 3) depicts similar tendencies in model discrepancies. That is to say that Nettuno generally overestimates wind speed, while UKMO presents SSL close to unity, and Nettuno EM variability is more pronounced than UKMO's. It must be noticed that in this case, the FVAR/OVAR ratios are on the order of 1.1 (Nettuno) and 0.9 (UKMO). Therefore, with respect to observations, both wave models have the general effect of transforming the input (wind) distribution into a wider one for the waves (ratios: 1.2 Nettuno, 0.96 UKMO). All wind speed biases are positive, the higher ones corresponding to Nettuno: 0.3 (altimeters) - 0.5 (buoys) m/s versus 0.2-0.1 m/s for UKMO.

Things appear to be slightly different when comparing the forecast against scatterometer wind speed observations (Figure 3, last two plots). The width of Nettuno distribution matches the observations one, while UKMO underestimates the measurements variability with a ratio of approximately 0.8. Both BIAS are negative, the stronger corresponding to UKMO input, which generally underestimates the observations (UKMO SSL is 0.94, while Nettuno's is 0.97).

The number of scatterometer data is one order of magnitude higher than the rest of the observation dataset. This roughly implies higher consistency of the verification results. On the other hand, the distribution of scatterometers Metop-A and Metop-B data is uneven in time. The observation time window is limited, hence the comparison with other instruments verifications is not straightforward.

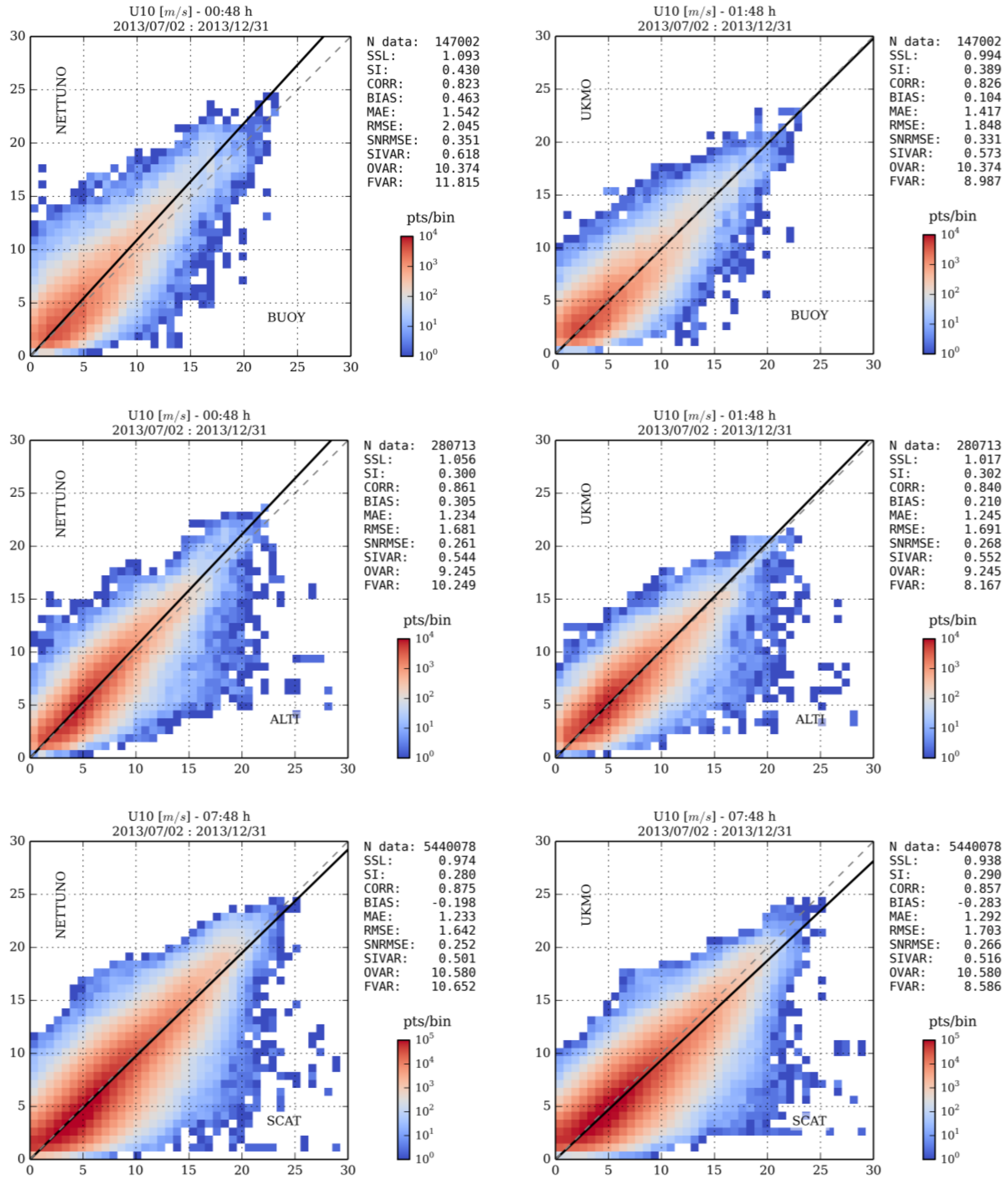


Figure 3: Ensemble Mean Forecast Vs. Observations histograms for 10 m wind speed [m/s]. Upper panels: buoy locations, mid panel: altimeters, bottom: scatterometers. Left: Nettuno; right: UKMO. Dashed line: main diagonal; continuous line: best symmetric fit. Coverage: Mediterranean Sea.

III.2 Reproduction of Statistical Features (Q-Q plots)

A useful way of comparing distributions is to consider a common set of quantiles and graph the corresponding couples (see *e.g.* Figure 4 where forecasts are compared with reference measurements in abscissa). In case of two identical distributions, the points of the graph lie on the main diagonal, *i.e.* the line with null intercept and unitary slope. From the q-q curve one obtains useful information. For example, if the points lie approximately on a line whose slope is lower (greater) than unity, it means that the tested distribution - ordinate - is narrower (wider) than the reference one - abscissa. The fact that all points lie above the diagonal and can be fitted with a unitary slope line is a symptom of positive bias.

Figure 4 compares the significant wave height distributions of both Nettuno (top) and UKMO (bottom) forecast against buoy observations in terms of percentiles. The right column plots are detailed zoom of the distribution tails, which are defined with respect to the 95th percentile of the observations. Results for altimeter measurements are depicted in Figure 5.

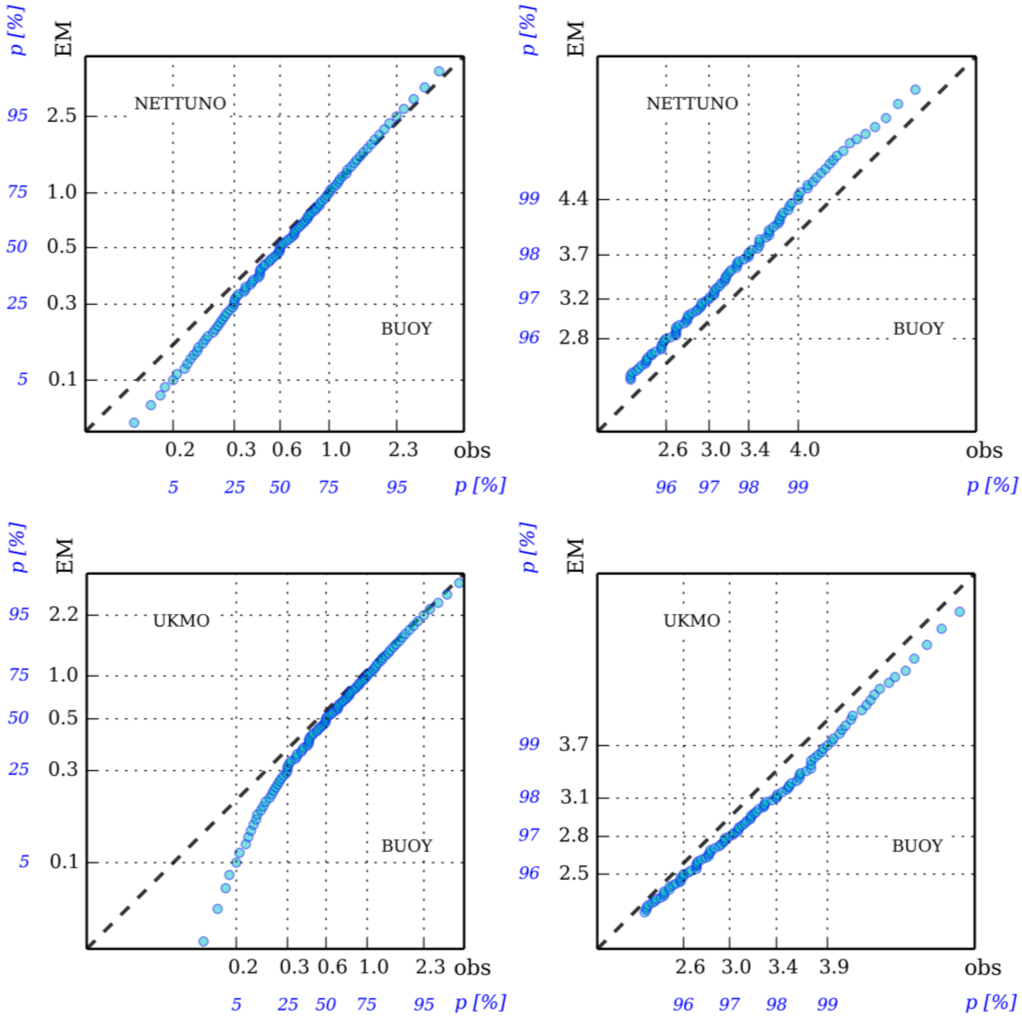


Figure 4: QQ plot for forecast (EM) H_S [m] Vs. buoy observations. Left: whole distribution; right: tail (obs > 95th percentile). Time window: 1/7 to 31/12 2013. Cumulative for lead time up to 24 h. Coverage: Mediterranean Sea. Upper panels: Nettuno, lower: UKMO.

Both model outcomes touch the main diagonal. This occurs at approximately 75th percentile for Nettuno, and 85th percentile for UKMO. The body of both distributions are thus wider than the reference sample. While Nettuno crosses the diagonal, UKMO does not; that is Nettuno's tail is flatter and wider than reference, whereas UKMO's decays faster. Hence, the former slightly exaggerates the occurrence of extreme cases, and the latter slightly underpredicts it. Similar behaviour is observed for the altimeter locations (Figure 5), but there a lower degree of discrepancy with reference climate, especially for UKMO wave heights. The analysis of the two distribution tails depicts a somehow smaller, but still significant, overprediction for Nettuno. UKMO underprediction appears at the very end of the tail.

It is worth to mention that the tail discrepancies do not underline better or worse extremes prediction for one of the two models. These aspects just indicate that, especially with respect to buoy measurements, extreme events tend to be overestimated by Nettuno ensemble mean (0.3 m), and underestimated by UKMO (0.2 m).

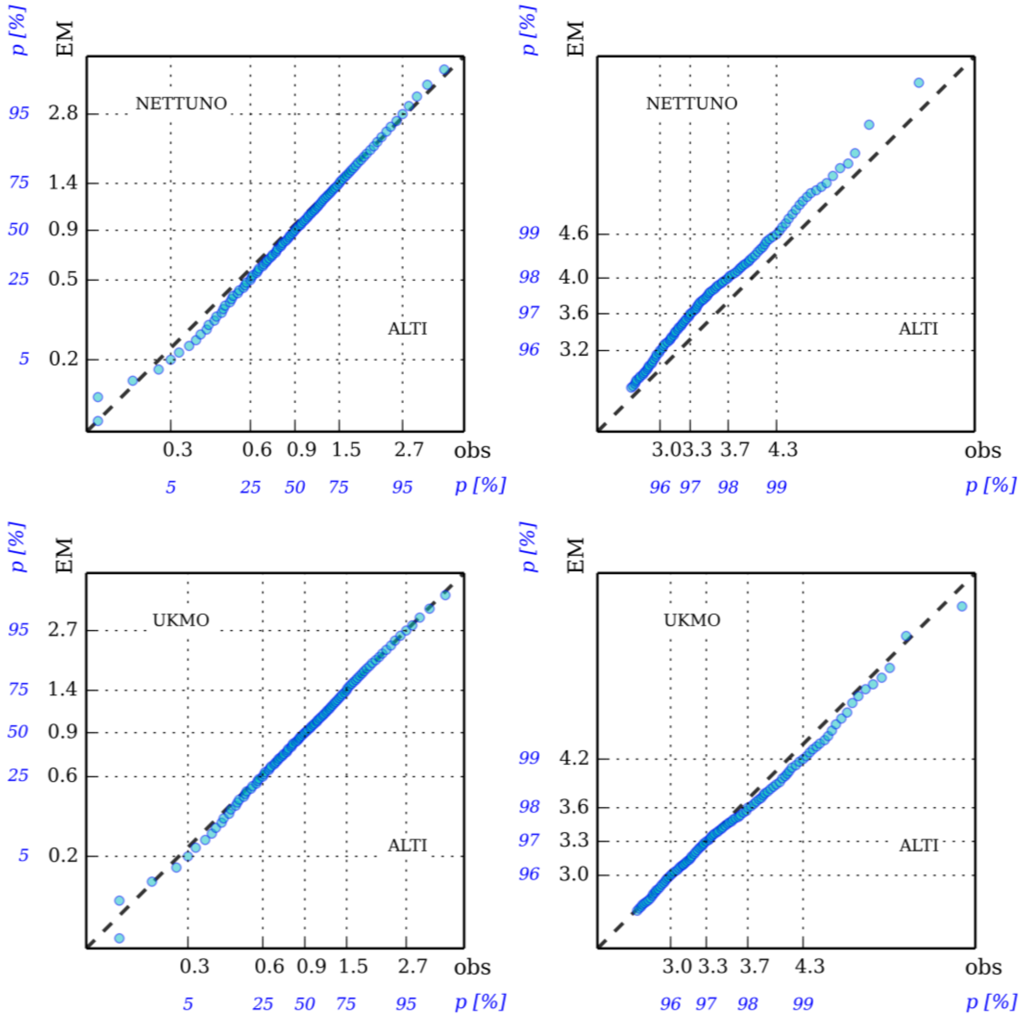


Figure 5: same as Figure 4 but for altimeters.

III.3 Taylor Diagrams

Taylor diagram (see *e.g.* Figure 6) provide a synthetic visualization of the goodness of a *deterministic* forecast, allowing comparison with multiple forecasts. The two variables, namely STD (standard deviation) and the arc-cosine of Pearson correlation (CORR) are here polar coordinates. We here propose the normalized version of the representation, with RSTD (STD normalized with the observation sample value) as the radius. The normalization is useful for comparison of model performances with respect to different variables. The centred pattern RMS difference (Taylor, 2001) is contoured as circles centred at the reference point with unitary RSTD and correlation. The closer the points to this reference, the better the skill of the model.

In Figure 6 we propose the synthetic comparison of both wave model input (U_{10}) and output (H_S, T_m) skills with reference to buoy dataset. Figure 7 accounts for altimeter measurements. The right plots depict ensemble mean performances. Left plots are given as reference: they represent the Taylor skills of the control members (DF - *deterministic* forecast). In the case of Nettuno the control member coincides with the Italian Meteorological service operational forecast. Note that the analysis is limited to the first 24 forecast hours. For the inspection of some error measures along forecast time, please refer to §III.4 . We notice that in general, and for both forecast systems, the ensemble mean is more skilled than the control member (please remember that there's no grid resolution difference with respect to the EPS). The improvement is much visible for the input (wind speed) than for the wave models output, and, between the two, for Nettuno's set. In particular, both Nettuno and UKMO wind variability consistently reduces, but the former denotes much improved correlation with the signal. Anyhow, UKMO wind speed turns out to be slightly more skillfull with respect to buoy measurements, and the two systems equate in the comparison with altimeter data.

There's not strong variation in the significant wave heights prediction skill, except for Nettuno at altimeter sites. According to Taylor's point of view, and looking only to these synthetic variables, there seems to be not much correlation between the improvement in the model inputs and their outputs.

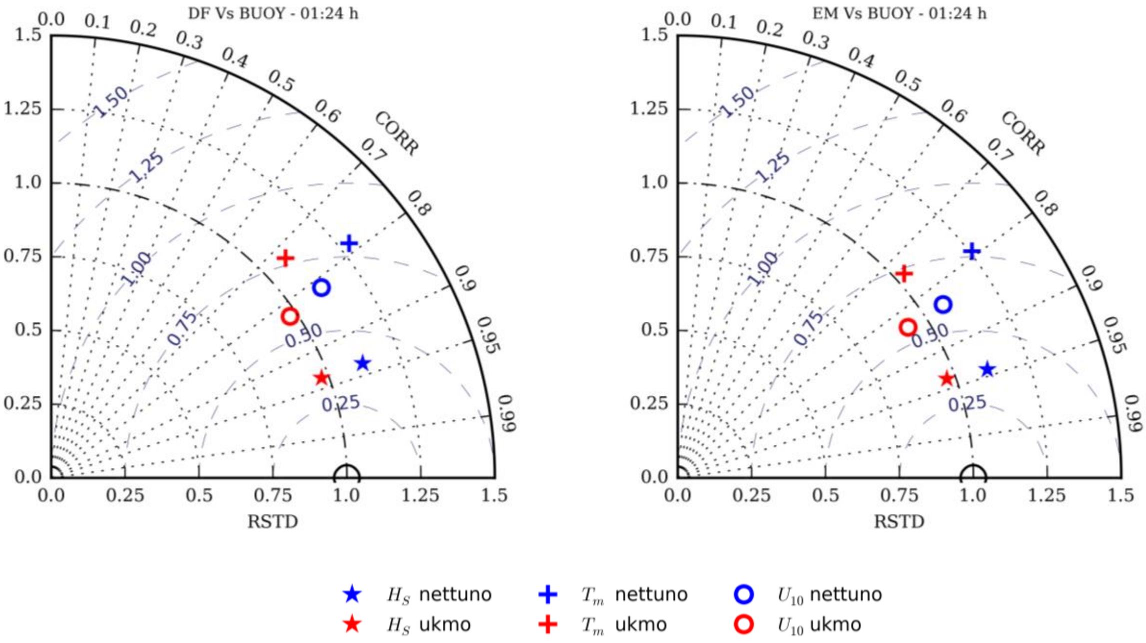


Figure 6: Taylor diagram for control member (left) and ensemble mean (right) referred to buoy measurements. Comparison for the first 24 forecast hours. Standard deviation and centred pattern RMS are normalized with respect to reference. Time window: 1/7 to 31/12 2013. Coverage: Mediterranean Sea.

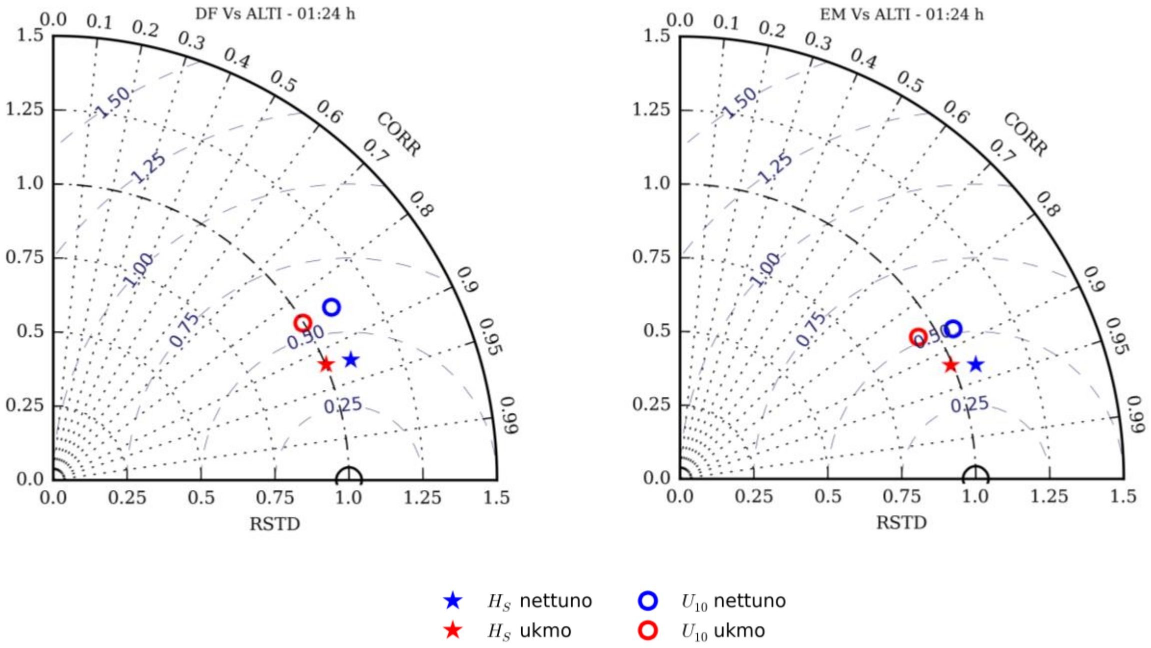


Figure 7: same as Figure 6 but referred to altimeters.

III.4 Deterministic Errors

In this paragraph we give an inspection of how some error measures vary along forecast lead time. At first, we present some common measures of deterministic error, defined as the difference between observations and co-located forecasts. Then we assess the model skills through a set of non-dimensional indices. These allow relative comparisons variable by variable. Data are represented at six hour synoptic UTC intervals (00, 06, 12, 18), meaning that the results are averaged over the preceding six hours collection.

As dimensional measures we consider here only the average error (BIAS) and its root mean square, namely:

$$BIAS = \langle \mathbf{f} - \mathbf{o} \rangle \qquad (2)$$

$$RMSE = \left\| \mathbf{f} - \mathbf{o} \right\|_2 \qquad (3)$$

where **f** is the vector of forecast ordered with respect to observations **o**, <·> denotes average and $||\cdot||_2$ is Euclidean norm.

In Figure 8 we show the errors for wind speed along lead time, subsampling every 6 hours, for both models with respect to buoy measurements. In Figure 9 and Figure 10 we give the same kind of results with respect to altimeter and scatterometer data, respectively. Together with the errors of control member (dashed lines) and ensemble mean (continuous lines), each plot depicts also the range of error corresponding to the ensemble range (shading).

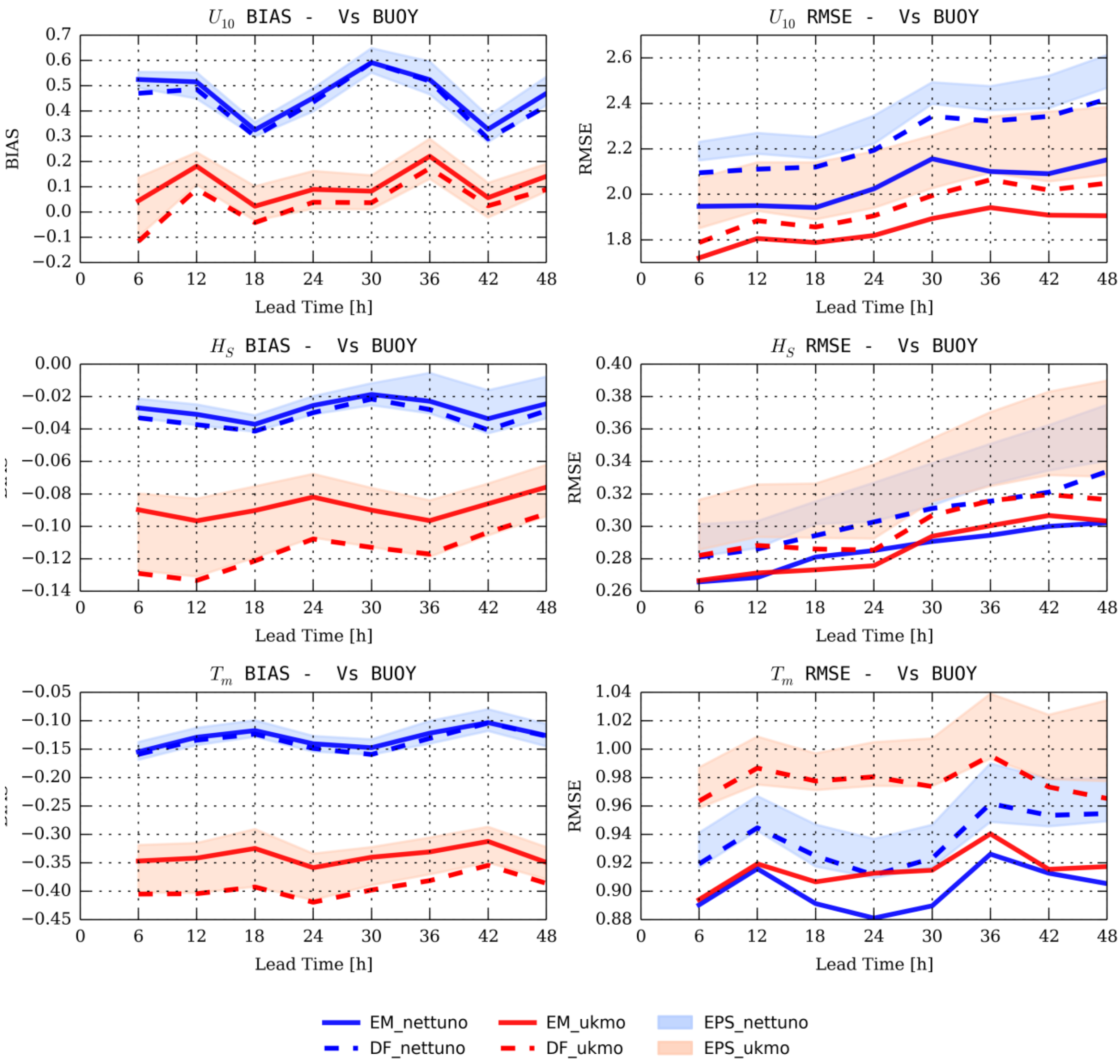


Figure 8: BIAS (left) and RMSE (right) for Nettuno (blue) and UKMO (red) at buoy locations. Top row: wind speed [m/s]; middle: significant wave height [m]; bottom: mean wave period [s].

At first let us consider models BIAS against buoy and altimeter observations. It can be noticed that there is no appreciable trend and all values fluctuate around the mean values which are assessed at the beginning of this chapter (§III.1 and related figures). In all cases, when passing from control member to ensemble mean there's an increase in bias, which mean that there's a general increase in energy content. This moves the bias of the forecast in positive direction. For the model input (wind speed) the effect is negative since the bias is already positive for both Nettuno and UKMO control member. For significant wave height, which control bias is negative, the energy increase determines an improvement. The models behaviour with respect to scatterometer data (Figure 10) does not show the tendencies observed for the remaining part of the verification dataset. Please note that a detailed analysis along forecast lead time cannot be pursued, since scatterometers scan the Mediterranean sea only at certain hours (this is the reason for the missing points in Figure 10). We must add that the EPS range width differences (model to model) is much more relevant in the wave parameters than in the model input. It seems that UKMO wave model is a little more sensitive to wind speed perturbations than Nettuno.

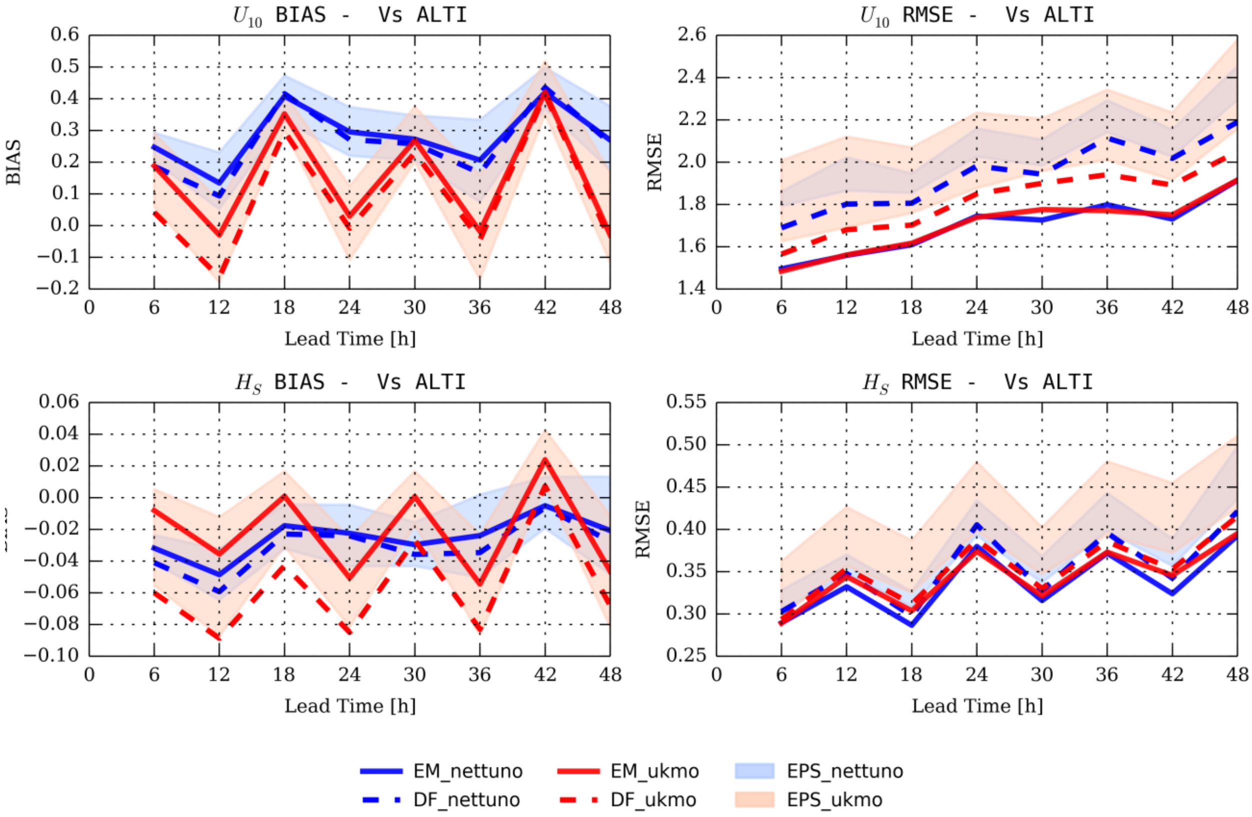


Figure 9: Same as Figure 8 but wind speed [m/s] (top) and significant wave height [m] (bottom) at altimeter locations.

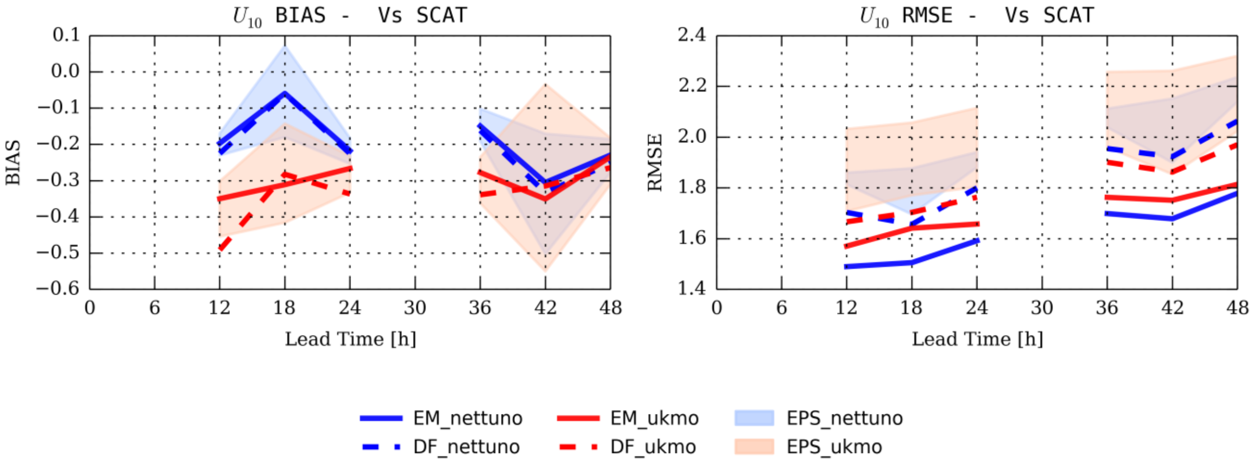


Figure 10: Same as Figure 8 but for scatterometer wind speed [m/s].

The variations so far are barely noticeable for Nettuno, and relevant for UKMO. The reason for that is a wider spread of the latter since the start. However, Nettuno is the more skilled (in terms of bias) for wave magnitudes at buoy locations, whereas it generally overestimates wind speed (Figure 8). This overestimation is still visible in the comparison against altimeter data, but it is less significant (Figure 9). In this case the model to model comparison is a little bit difficult: while the Nettuno bias is somehow stable along the forecast lead time, both

UKMO U_{10} and H_S show strong variations. The main reason is the altimeter dataset, which is unevenly distributed in space during the day. In particular, there are no Saral-Altika data (except for some points in the Alboran Sea) between 00:00 and 06:00 UTC, and between 12:00 and 18:00 UTC. The relationship between one model behaviour and one altimeter depends on both. As an example, consider Figure 11 which depicts the significant wave height BIAS for both models against the three altimeters, and note the strong zonal differences especially in the comparison with CryoSat measurements.

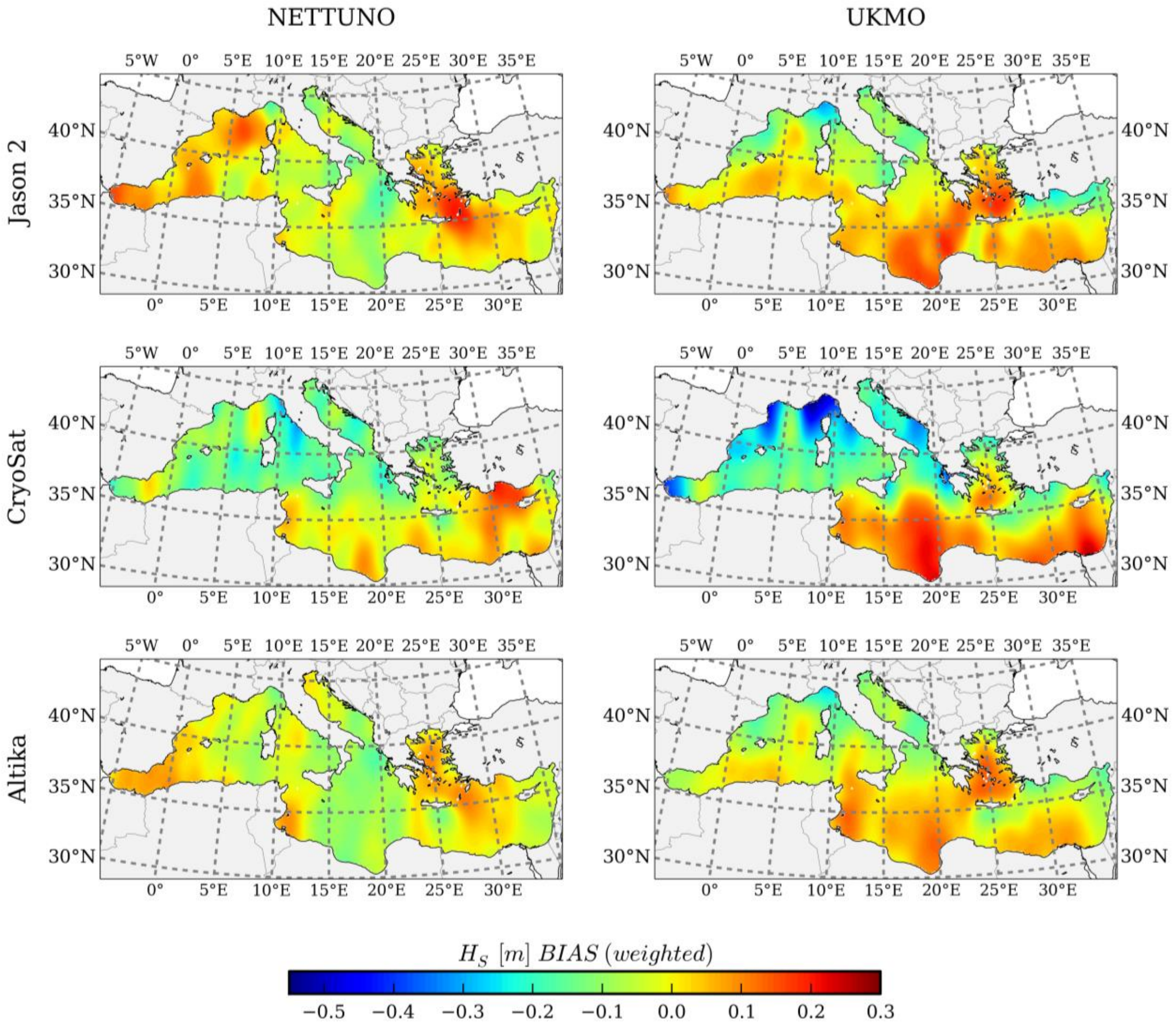

NETTUNO
UKMO
Jason 2
CryoSat
Altika
5°W 0° 5°E 10°E 15°E 20°E 25°E 30°E 35°E
40°N
35°N
30°N
0° 5°E 10°E 15°E 20°E 25°E 30°E
H_S [m] BIAS (weighted)
−0.5 −0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3

Figure 11: Map of H_S [m] BIAS for Nettuno (left) and UKMO (right) ensemble mean against the three altimeters. Results are gaussian weighted for smooth rendering.

Let now spend some words about RMSE (see the right column plots of each figure). In general, there is a net reduction of error if one considers the ensemble mean forecast, instead of the control member. For each variable considered herein, and for each instrument, the RMSE increases with forecast lead time. Since BIAS is almost stable, this means that there's an increase of error variance. Except for wind speed at buoy locations (Figure 8, top right panel), in all other cases the EM of both models show similar RMSE values.

Dimensional error measures do not allow cross comparison of model performance with respect to different output fields and different verification baselines (*e.g.* against buoy and altimeter observations). This is the main reason for proposing a set of non-dimensional verification metrics. Figure 12 introduces two classic measures which quantify the slope and the dispersion of the scatter between forecast and observed significant wave height (see III.1): symmetric slope (SSL) and scatter index (SI). Symmetric slope (SSL, top row) is defined as the slope of the best fitting line, with zero intercept, which minimizes the orthogonal errors in the scatter data. It can be retrieved with the following equation:

$$SSL = \left(A + \sqrt{A^2 + B^2}\right)B^{-1} \tag{4}$$

with

$$\begin{aligned} B &= 2\mathbf{f}^T\mathbf{o} \\ A &= \left\langle \mathbf{f}^2 - \mathbf{o}^2 \right\rangle \end{aligned} \tag{5}$$

where: ***f*** and ***o*** are co-located forecast and observations, and the superscript *T* stands for *transposed*. Values of SSL equal to unity indicate a perfect *average* one to one relationship between forecast and observations. The scatter index (SI, bottom) is defined as

$$SI = RMSE \cdot \langle \mathbf{o} \rangle^{-1} \tag{6}$$

is a measure of the dispersion of the scatter, *i.e.* lower SI indicates less dispersed data scatter.

Nettuno SSL is constant along forecast lead time, and slightly higher that one (Figure 12, top). There are no differences in taking EM or DF as forecast. On the other hand UKMO SSL ranges from about 0.93 against buoy observations and an average 0.98 against altimeter ones. In this case, the value varies approximately from 0.97 and 1.00 whether the verification includes or not the measurements of Saral-Altika. Moreover UKMO EM improves the skill of the control member at early lead times. As the lead time increases the SI of both systems increases in the same manner (Figure 12, bottom) and their skill is practically indistinguishable. Both EM slightly improve the skill of the correspondent control members.

An alternative formulation for scatter index, which is based on the hypothesis that both observations and forecast are error affected, is the symmetric normalized RMSE:

$$SNRMSE = RMSE \cdot \left(\mathbf{f}^T\mathbf{o}\right)^{-\frac{1}{2}} \tag{7}$$

Equation *(7)* represents the correct representation of the dispersion of the plume around the scatter diagram main diagonal, since SNRMSE is a measure of the orthogonal dispersion of the forecast-observation couples. Figure 13 indicates that, by this point of view, this dispersion is slightly lower for Nettuno H_S, mainly in the comparison against buoy measurements (left panel). The observed similarities (and differences) in SI are due to a composition of elementary quantities. In fact, the square of the scatter index is proportional to MSE, and therefore to the sum of error standard deviation and BIAS. The latter has already been analysed in the previous paragraph. It is therefore important to quantify the variability of the error after that the bias is removed. Normalization of error variance with respect to observations variance, instead of their mean, allow the following decomposition of scatter index:

$$SIVAR = RVAR - 2\,CORR \cdot RVAR^{\frac{1}{2}} + 1 \tag{8}$$

where CORR is Pearson correlation between forecast and observations, and RVAR is the ratio of forecasts variance over observations variance.

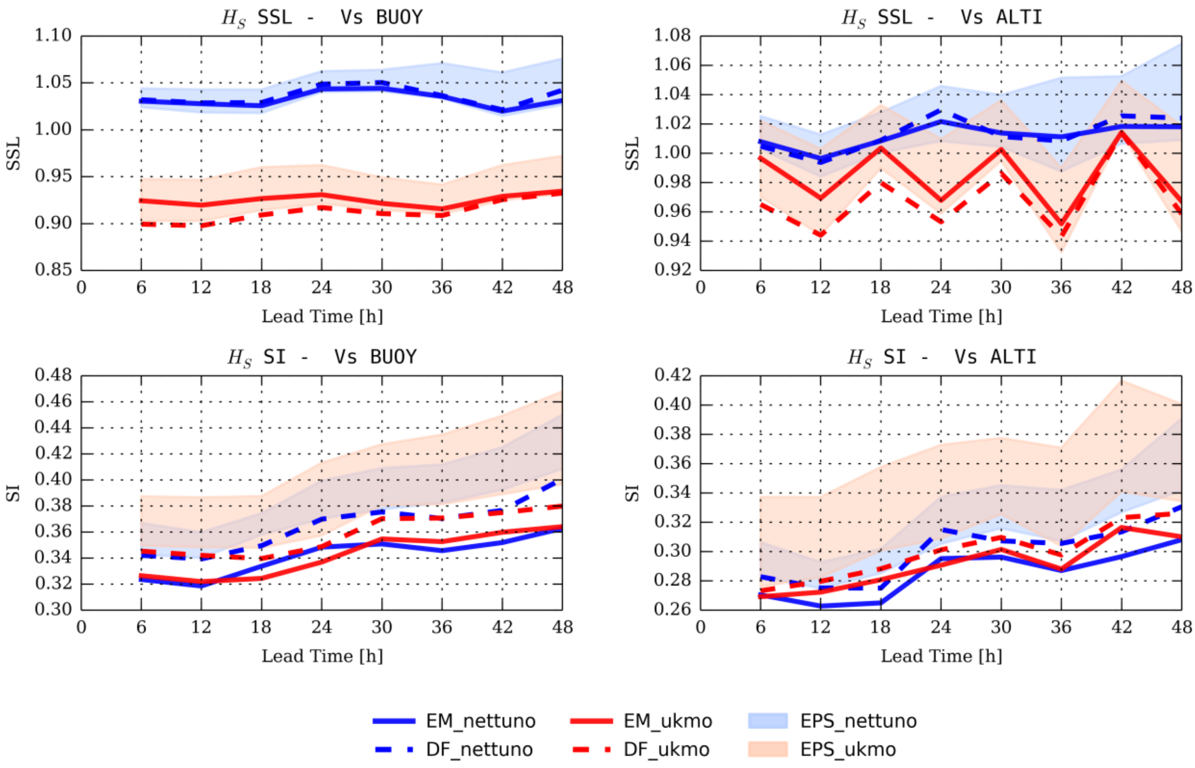


Figure 12: Classic non-dimensional measures: symmetric slope (top) and scatter index (bottom). Verification of forecast H_S against buoy (left) and altimeter (right) observations.

Equation *(8)* is a measure of pure scatter, since it does not account for bias. By definition, the absolute value of CORR is never higher than one, and RVAR is the square of a positive quantity. Therefore SIVAR is pure positive quantity, reaching its minimum (zero, best skill) at RVAR and CORR both equal to one. Results for this breakdown are presented in Figure 14 for H_S forecast against buoy (left) and altimeter (right) measurements.

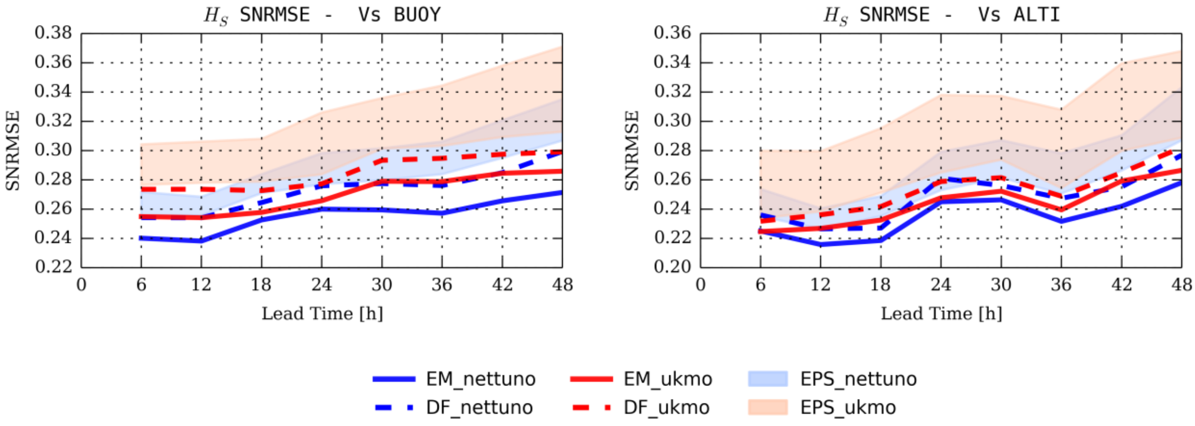


Figure 13: Symmetric Normalized RMSE for forecast H_S against buoy (left) and altimeter (right) observations.

In general, the correlation between model and observation (top row) decreases with lead time, and the ratio between model and observations sample variability (middle row) is somehow constant. Therefore the scatter, measured as SIVAR (bottom) increases with forecast lead time. For both elementary indices (RVAR, CORR) we observe no variation between UKMO ensemble mean and control member, whereas the improvement for Nettuno is noticeable. For this reason, there's sensible improvement in SIVAR for Nettuno EM with respect to DF, and no (or negligible) improvement for UKMO. In the comparison against buoy data (left) the best Nettuno correlation does not compensate for the high forecast variability, and the UKMO SIVAR is slightly better. In the matching with altimeter measurements, this compensation occurs, and the skill of both EM are very close to each other.

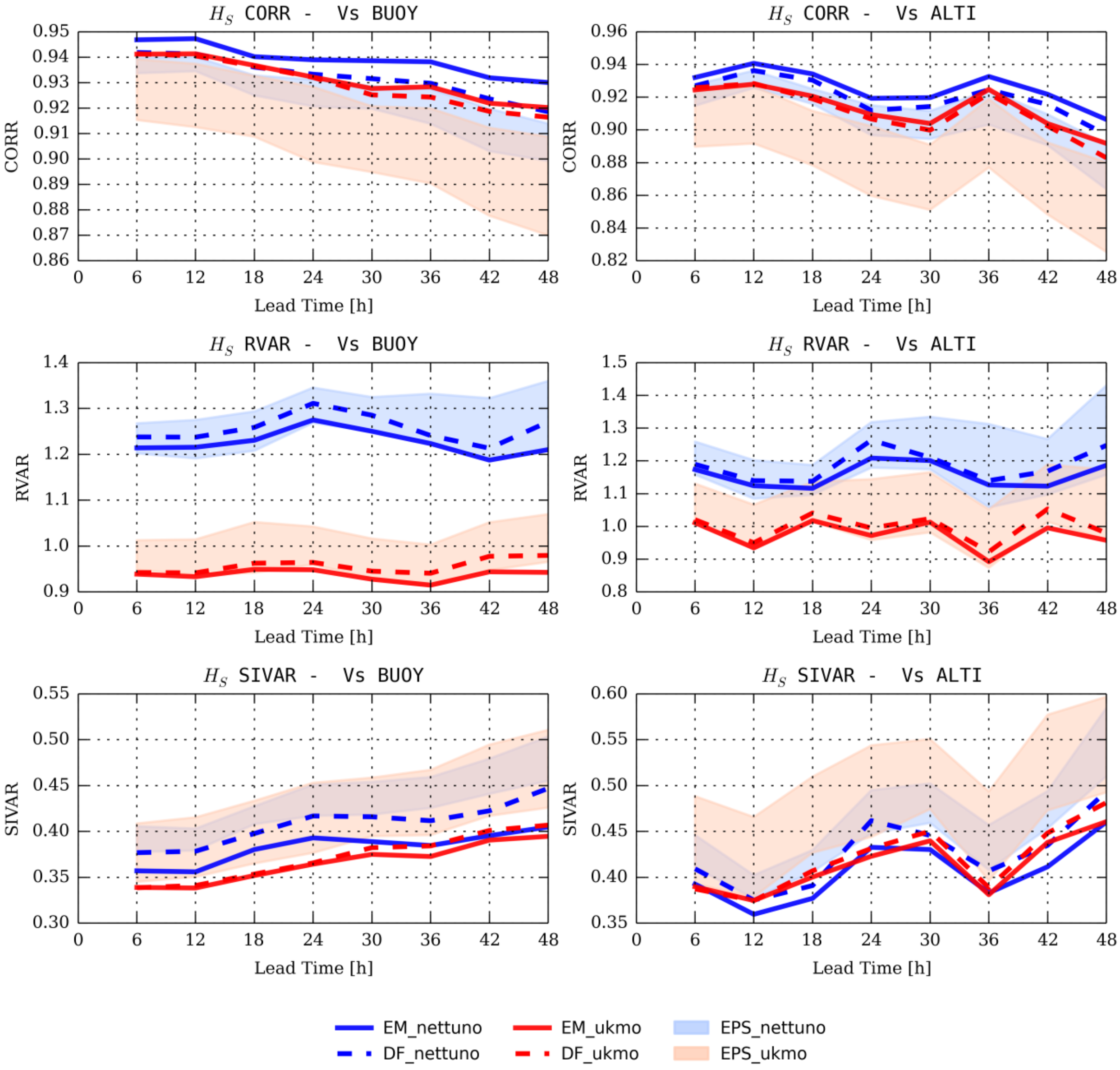


Figure 14: Modified scatter index (SIVAR, bottom) and breakdown (CORR, top and RVAR, middle) for forecast H_S against buoy (left) and altimeter (right) observations.

III.5 Summary

In this section we have investigated on the ensemble mean ability to represent the observations sample statistics, and on possible added skill with respect to the control member.

- Results based on the non-dimensional (mostly unbiased) verification of UKMO and Nettuno EPS confirm what is already assessed in literature, that is that ensemble mean is generally more skilled than a *deterministic* forecast, if the EPS is built with the same model grid resolution.

- On the other hand, the analysis of both wave model input and output bias shows that the EMs are more energetic than the control members. This is partly due to the way in which the results are post-processed to get the EPS statistics. For example, the straightforward average of significant wave height is higher than the wave height associated to the wave model energy average.

- The main topic of this report is the UKMO and Nettuno EPS inter-comparison. From this point of view we observed similar skills for both systems. The main difference is related to the overall co-located sample distribution of forecasts. In particular Nettuno distribution tails are wider than the observed climatology, while UKMO are slightly shorter.

- We noted in particular that UKMO bias is more sensitive than Nettuno one to the comparison with different altimeters. In terms of Taylor skills, it turns out that UKMO wave predictions are better balanced than Nettuno ones at buoy locations, and slightly less skilled with respect to altimeter observations. Nettuno is more correlated to the observations than UKMO, and this aspect compensates the skill gap due to the higher variability.

IV ENSEMBLE SPREAD

In this report we generally address to (ensemble) spread, or ES, to the standard deviation of the ensemble realizations. In some cases, but only where stated, ES is evaluated as the inter quartile range of the ensemble realizations.

This chapter starts with a brief description of the overall spread evolution along forecast time (§IV.1). Then we present synthetic diagnostic diagrams which accounts for the EPS capacity in capturing the observations (Bounding Boxes, §IV.2) and for their average distribution inside the ensemble cloud (paragraph IV.3 and Rank Histograms). Finally, in §IV.4 we investigate the average relationships between EPS spread and deterministic errors in the six months verification sample.

It is worth to recall that BBs and RHs are evaluated following Saetra et al. (2004), adding noise to the *EPS* members. The amount of uncertainty and the noise properties have been introduced in §II.2 .

IV.1 EPS Spread Climate

Here we propose a comparison between the two models based on the development along forecast time of the ensemble spread for a number of variables. Figure 15 depicts this comparison for wave model input (U_{10}, top left) and output, namely H_S (top right), T_m (bottom left) and θ_m (bottom right). Nettuno and UKMO systems differs in the technique with which the ensemble members are generated and propagated in time. UKMO EPS has an overlapping structures which keep some memory of past spread. This causes UKMO wind speed initial distribution to be wider than Nettuno one (Figure 15, top left). Anyhow, while propagating in time, Nettuno U_{10} ES grows faster, reaching its counterpart at the end of the 48h.

Effects of the higher UKMO initial input spread are reflected on the wave model output magnitudes (e.g. H_S, top right). However, the wave fields are subject to a long (from hours to days) hysteresis. This causes much lower growth rates for significant wave height spread, for example. Therefore Nettuno-UKMO H_S spread difference begins to reduce with some delay, with respect to what happens in the wind field. This aspect is more visible in the bottom left panel of Figure 15: both mean wave period spread growth and its system-system difference reduction start to be relevant after approximately 30h of forecast. On this basis, we expect a total reduction in EPS spread difference in the interval between 72h and 80h forecast time.

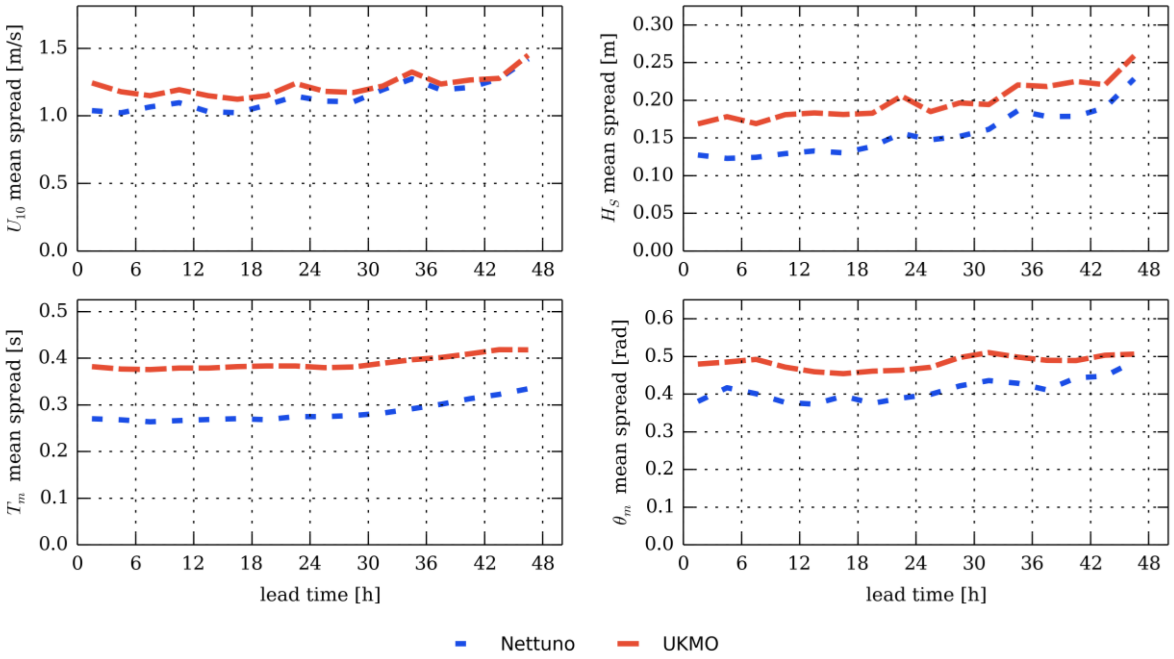


Figure 15: EPS spread climate for U_{10} (top left), H_S (top right), T_m (bottom left) and θ_m (bottom right). Spread is averaged (RMS) each 3h of forecast lead time. Time window: 1/7 to 31/12 2013. Coverage: Mediterranean Sea. Estimations made accounting for co-located forecasts only.

IV.2 Frequency of Missed Observations (Bounding Boxes)

In this paragraph we introduce the first validation of Nettuno and UKMO EPS spread against observations. This is done considering the relative frequency of empty bounding boxes (BBs). For a given measurement, the bounding box is defined by the extreme realizations of the co-located ensemble forecast. If the observation falls inside the BB, then its value is “0”. Otherwise the empty BB is labelled with “1”. The relative frequency of empty BBs, is the sum over the considered observations sub-sample of the BB emptiness, divided by the number of entries in the sub-sample. A BB can easily be generalized to any number of dimensions.

Let us consider the top left panel of Figure 16. Each point of each line represents the relative frequency of empty BB for a 3 hours lead time sub-sample. For example, the 30% of significant wave height buoy measurements data are out of Nettuno EPS extremes. That is, only the 70% of these observations take values that are bounded by the (local) extreme ensemble member values. The four panels in Figure 16 depict the results for significant wave height, wind speed, mean wave period and mean wave direction with reference to buoy measurements. Figure 17 gives the same results, but only for the cases in which H_S is higher than the 50th percentile of the underlying (buoy) climatology (i.e. H_S>0.6m). In Figure 18 we report results for multi-dimensional (multi-variable) BBs at buoy locations. The frequency of missed events at altimeter locations is given in Figure 19. Please note that all diagrams state “+ noise”, which means that the ensemble members have been post processed (prior to verification) adding some noise (see chapter II).

From a general point of view (Figure 16 to Figure 19), Nettuno and UKMO systems miss the same amount of observations, and the relative number of empty BBs is practically constant along the forecast lead time. Anyhow, some more or less macroscopic issues are to be described. For example, both EPS capture the same amount of wind speed buoy measurements (20%, Figure 16). Let us recall that, especially at early lead times and at buoy locations, UKMO spread is higher than the Nettuno one (see. Figure 15 and Figure 8), and that Nettuno bias for buoy U_{10} observations is much higher than its counterpart error (Figure 8). Top right panel of Figure 16 is therefore a clear example of similar skill due to compensation between two distinct features (the width of the ensemble and the model bias). This stresses the fact that no one of the verification measures should be taken as universal goodness parameters. However, from the present point of view (BBs) the Nettuno EPS wave model input tends to capture more observations than UKMO ensemble wind fields, especially when accounting for their behaviour against altimeter measurements (Figure 19).

Now let us consider the models output, and significant wave height in particular. It turns out, observing Figure 16 and Figure 19 (top left panels), that the UKMO skill is 3% to 5% lower (better) than the Nettuno ones. Note that the BBs do not account for variables values, but only for the relative position of the measurements with respect to the EPS extrema. In order to understand to which kind of subset (if any) the higher skill differences should be deputed, we perform the analysis isolating the upper half of the H_S measurement distribution. By comparing Figure 17 with Figure 16, one can appreciate that most part of the skill difference is due to the lower half part of the distribution (H_S<0.6m), *i.e.* to a subset of generally non interesting sea states. An interesting outcome is given by the fact that both systems miss approximately half of the wave triplets measured at buoy locations (Figure 18, bottom right).

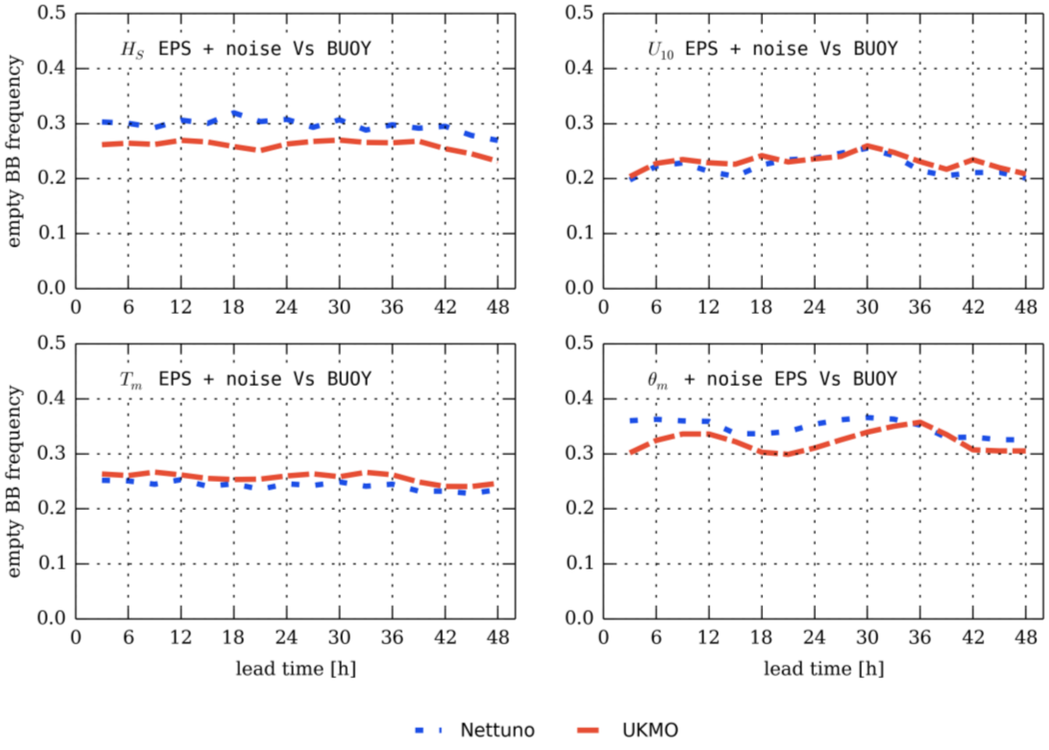


Figure 16: Frequency of empty bounding boxes per lead time (3h cumulated values). Results for verification against buoy observations: H_S (top left panel), U_{10} (top right), T_m (bottom left) and θ_m (bottom right). Time window: 1/7 to 31/12 2013. Coverage: Mediterranean Sea.

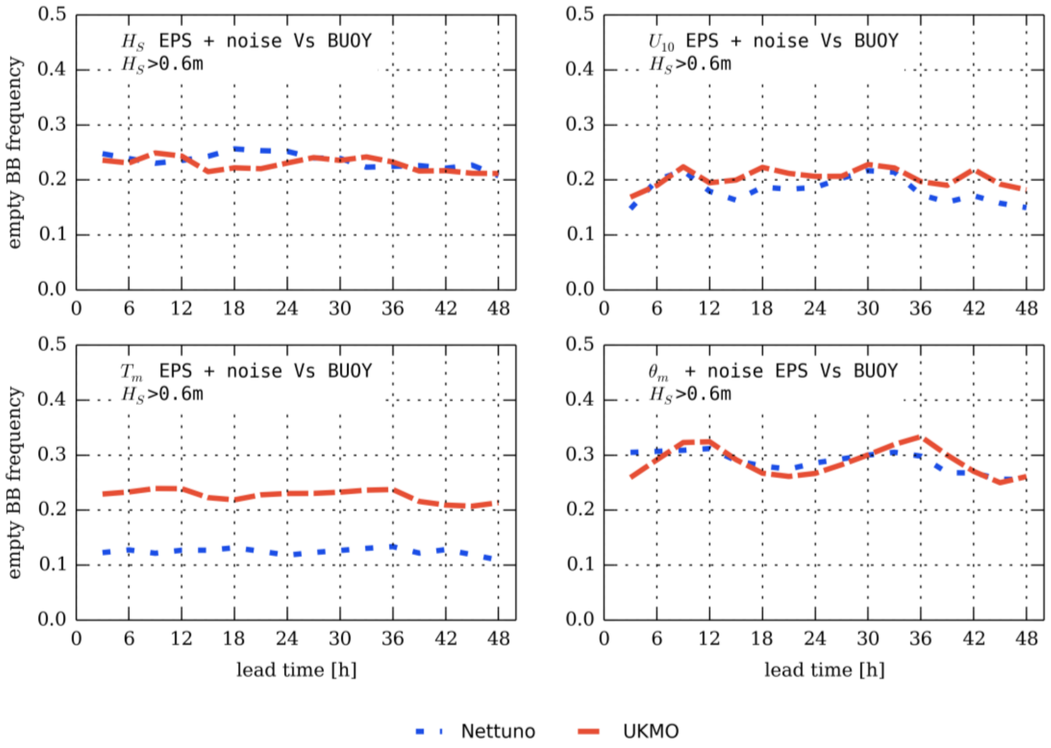


Figure 17: Same as Figure 16 but including only the cases in which H_S is higher than the median of the underlying climatology.

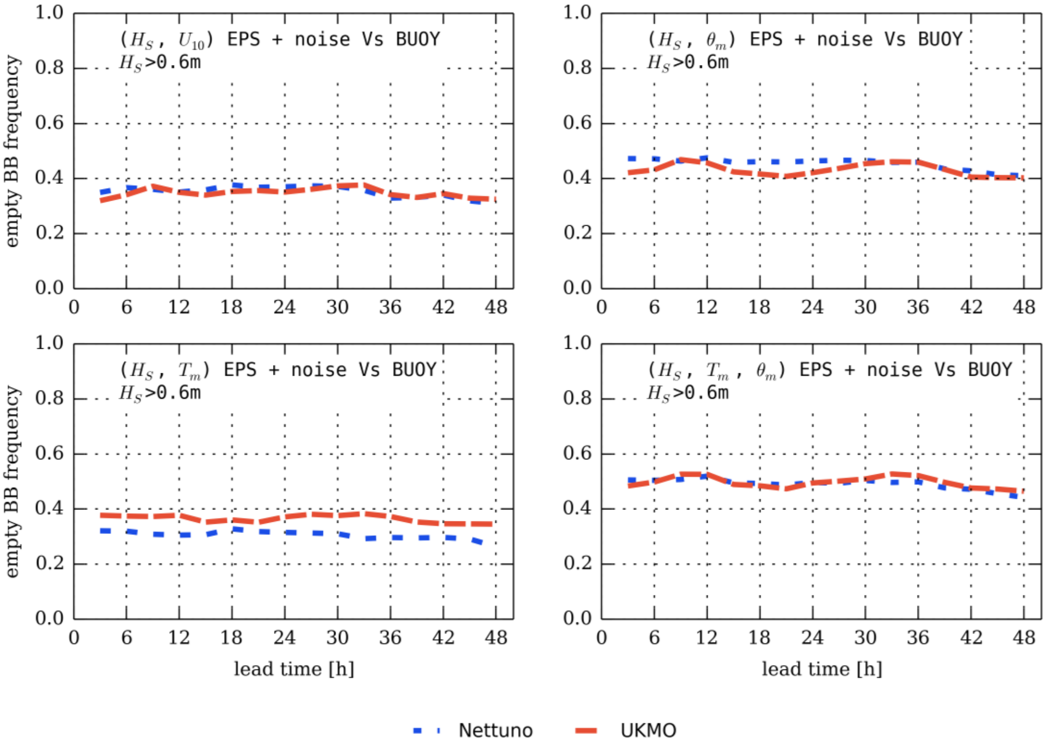


Figure 18: Frequency of empty multi-dimensional BBs for buoy observations, and for the cases in which H_S is higher than the median of the underlying climatology. The rest of notation is the same as Figure 16.

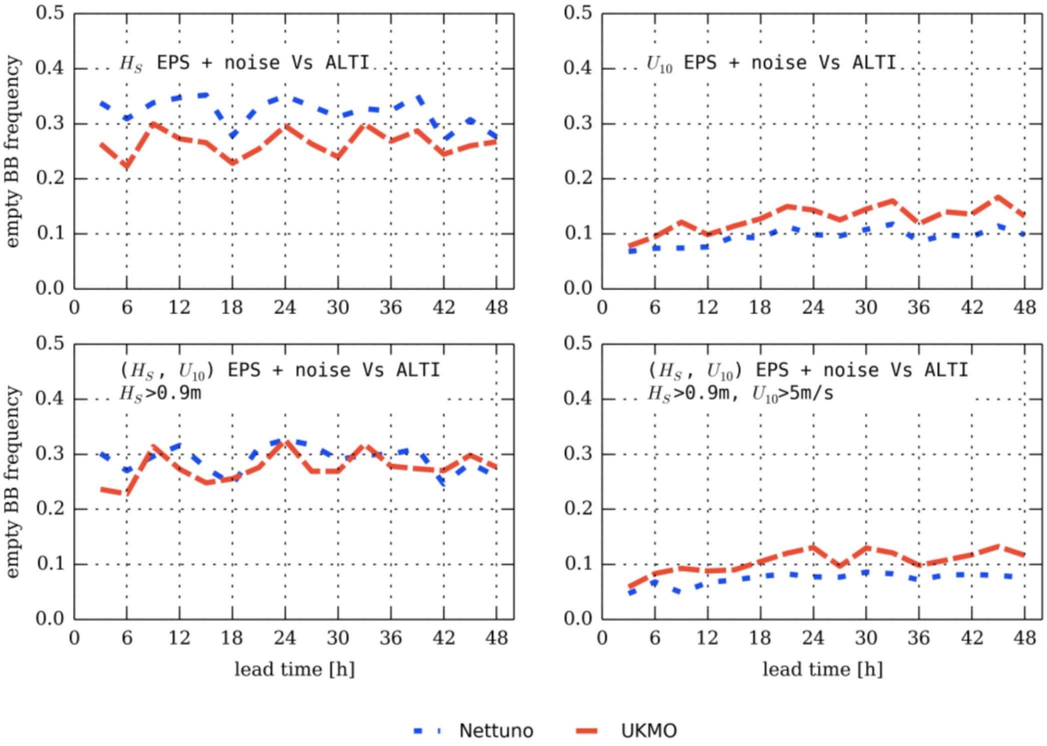


Figure 19: Frequency of empty mono- and multi-dimensional BBs for altimeter observations. H_S (tope left), U_{10} (top right), H_S and U_{10} with H_S higher than the median of the underlying climatology (bottom left), with both H_S and U_{10} higher than the median of the relative underlying climatology (bottom right). The rest of notation is the same as Figure 16.

IV.3 Distribution of Observations in the EPS

Rank Histograms (RH) were independently introduced by Anderson (1996), Talagrand et al. (1997) and Hamill and Colucci (1998). These histograms are generated accounting for the relative occurrence of ranked observations, where rank *n* is defined for an observation falling between the *n*-th and the *(n+1)*-th *EPS* member. If *N* is the total number of EPS members, then RH is build up with *N+1* bars. The leftmost (rightmost) one corresponds to rank *0* (*N*), and represents the relative number of observations falling below (above) the minimum (maximum) EPS outcome. The sum of the two extreme bars equals the frequency of empty BB. The ideal height of all the bars is *1/(N+1)*.

The two EPS systems considered in this report have different number of members which causes the classical RH inter-comparison practically impossible. Therefore we have done some adjustments to the original representation (see e.g. Figure 20). First of all, abscissa, representing the rank, is normalized by N. In order to make the vertical scale comparable, the ordinates are rescaled so that the perfect skill line is in common. By doing that, we have two different vertical scales, centred at the best skill: the left one (blue) is for Nettuno EPS, and the right one (red) is for UKMO.

Here we present verification diagrams for wind speed and significant wave height only. Results for comparison against buoy measurements are given in Figure 20, while Figure 21 depicts results related to altimeter observations.

Left and right plots give separate analysis for the first and second day of forecast. Please note that, passing from day 1 to day 2, there are very small differences in the distributions. This result copes with the stability (along forecast time) of the number of empty BBs (§IV.2). Since the relative frequency of missed observations is always greater than 1/(N+1), then RH have shapes that indicate that both EPS are generally under-spread.

UKMO significant wave height relative high bias (negative bias) causes the RH at buoy location to be right shifted, while Nettuno RH is a bit more symmetric (Figure 20, top panels). At altimeter locations the bias is nearly negligible, and both Nettuno and UKMO RH present the under-spread EPS characteristic U-shape. Anyhow, UKMO’s U-shape is slightly less pronounced. The wave model input (U_{10}) is under-spread at buoy location for both EPS systems (Figure 20, bottom). At altimeter locations (Figure 21, bottom), except for the bias, the almost flat distributions and the low empty BBs percentage suggest that the wind speed EPS is well calibrated for both systems.

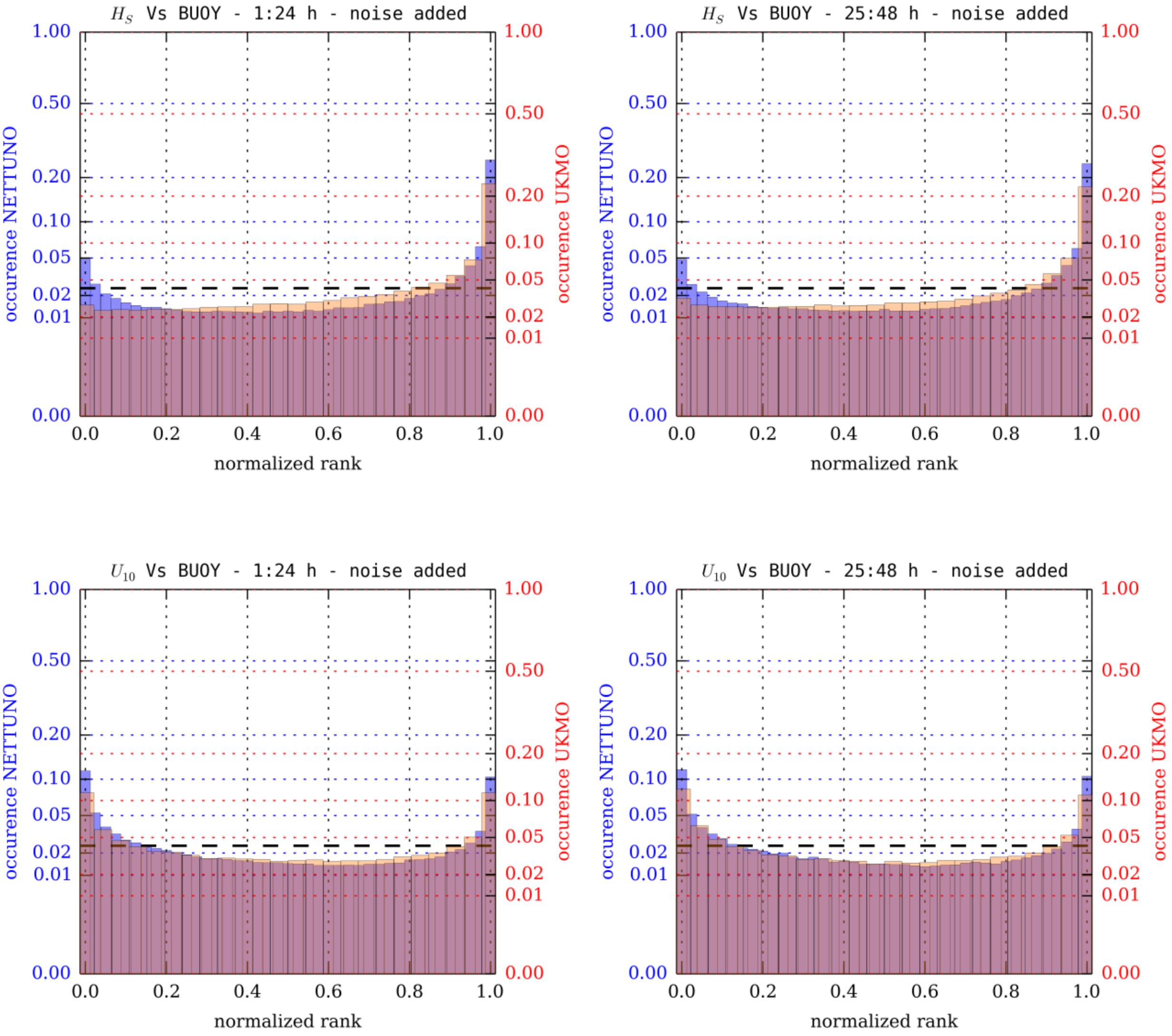


Figure 20: Rank Histogram for H_S (top) and U_{10} (bottom) at buoy locations. The horizontal dashed line indicates the perfect performance (Nettuno: 0.024; UKMO: 0.042). Time window: 1/7 to 31/12 2013. Coverage: Mediterranean Sea. Forecasts include observation uncertainties.

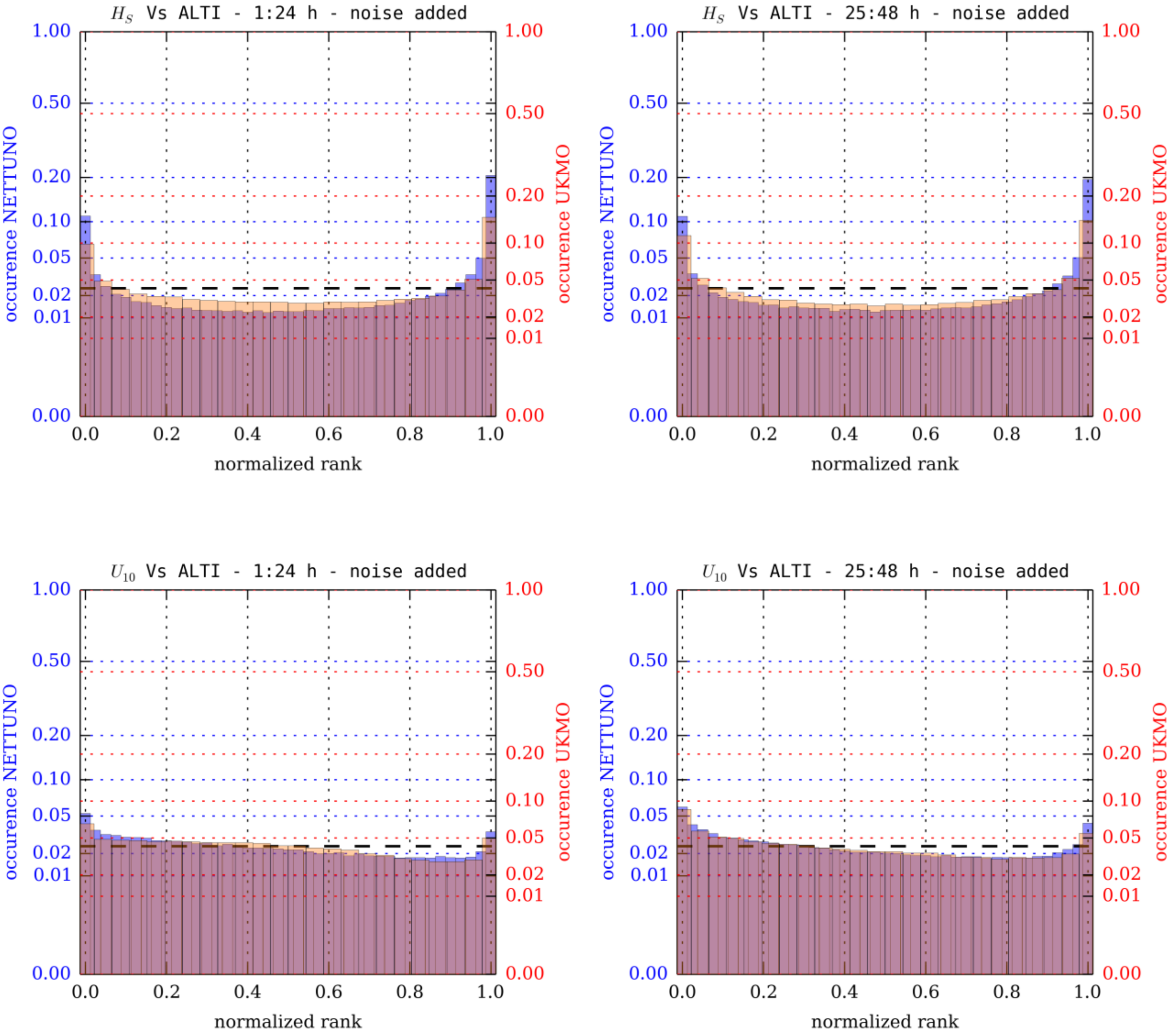


Figure 21: Same as Figure 20 but for altimeter observations.

IV.4 Spread-Skill Relationships

In principle, ensemble spread should replicate – but not coincide with (Hersbach, 2000) – the estimation of uncertainty given by the deterministic errors. We present here extensive graphical description of the existing relationships (if any) between ES and the ensemble mean deterministic error. We follow here two different approaches. The “continuous” one considers ES as the standard deviation of the EPS outcomes (see e.g. Scherrer et al. 2004), while the in “discrete” approach identifies the ES with the interquartile range (IQR) of the ensemble members (Saetra and Bidlot, 2004). The ensemble members are enriched with noise, in order to mimic the uncertainty of observation errors.

If pursued via scatter diagrams, the analysis would produce cumbersome results. As done by a number of authors, the investigation is carried on comparable dataset sub-samples. For a given variable, the co-located set of spread and skill is sub-sampled according to a number of ES bins. In the "discrete" approach, the error measure corresponding to the ES (IQR) bin is taken as the 75th percentile of the EM absolute errors populating the bin. In the "continuous" approach we take the standard deviation of the EM-observation couples (not their RMS which is an overestimation since it depends also on the bias). For example, Figure 22 (top left panel) depicts the results of the "continuous" approach for modelled H_S against altimeter measurements corresponding to the first 12h of forecast. The blue and read lines are the standard deviation of errors in the corresponding ES bin for Nettuno and UKMO EPS, respectively. The background of the plot shows the histograms (referring to the right plot vertical axis) of the ES distribution. Note that too little populated bins are not taken into account.

Results for altimeter measurements are given in Figure 22 ("continuous") and Figure 23 ("discrete"). Top row relates to wave model output (H_S), and the wave model input (U_{10}) spread-skill relationships are depicted in the bottom plots. In each row there are four distinct graphs, one each 12h of forecast time, with time increasing from left to right. Figure 24 and Figure 25 report the same results for buoy sites. Finally, in Figure 26 and Figure 27 we add the description of spread-skill for mean wave period and direction. The tables following this paragraph (from Table 3 to Table 8) resume the parameters of the following linear relation for deterministic error (DE) as a *function* of ES:

$$DE = a + b \cdot ES \qquad (9)$$

According to previous paragraph, both wind speed and significant wave height ensemble members distributions are somehow too narrow. And this holds for both Nettuno and UKMO EPS. In this paragraph we look at ES relationship with related deterministic error, and, from this point of view, it appears that the ensemble systems are under-spread only under some circumstances. The reader can appreciate that in all the figures that follow these lines there's always an initial under-spread at low ES (and error) values, but a general tendency to overestimate the uncertainty at high values. This is much visible in the figures reporting the "discrete" results.

In the comparison against altimeter observations (Figure 22 and Figure 23) we see that Nettuno EPS lines are steeper than UKMO's ones (see also Table 3 to Table 8), with the latter tending to underestimate less than Nettuno in the low range, and overestimate more in the high range. In the analysis limited to buoy locations (Figure 24 to Figure 27) the behaviour of the two EPS systems is quite similar: spread-skill lines are approximately parallel almost everywhere. However, UKMO lines are always below the Nettuno one, which means that at low error values Nettuno tends to underestimate the uncertainty more than UKMO, while the latter tends to exaggerate the ES in the high uncertainty ranges.

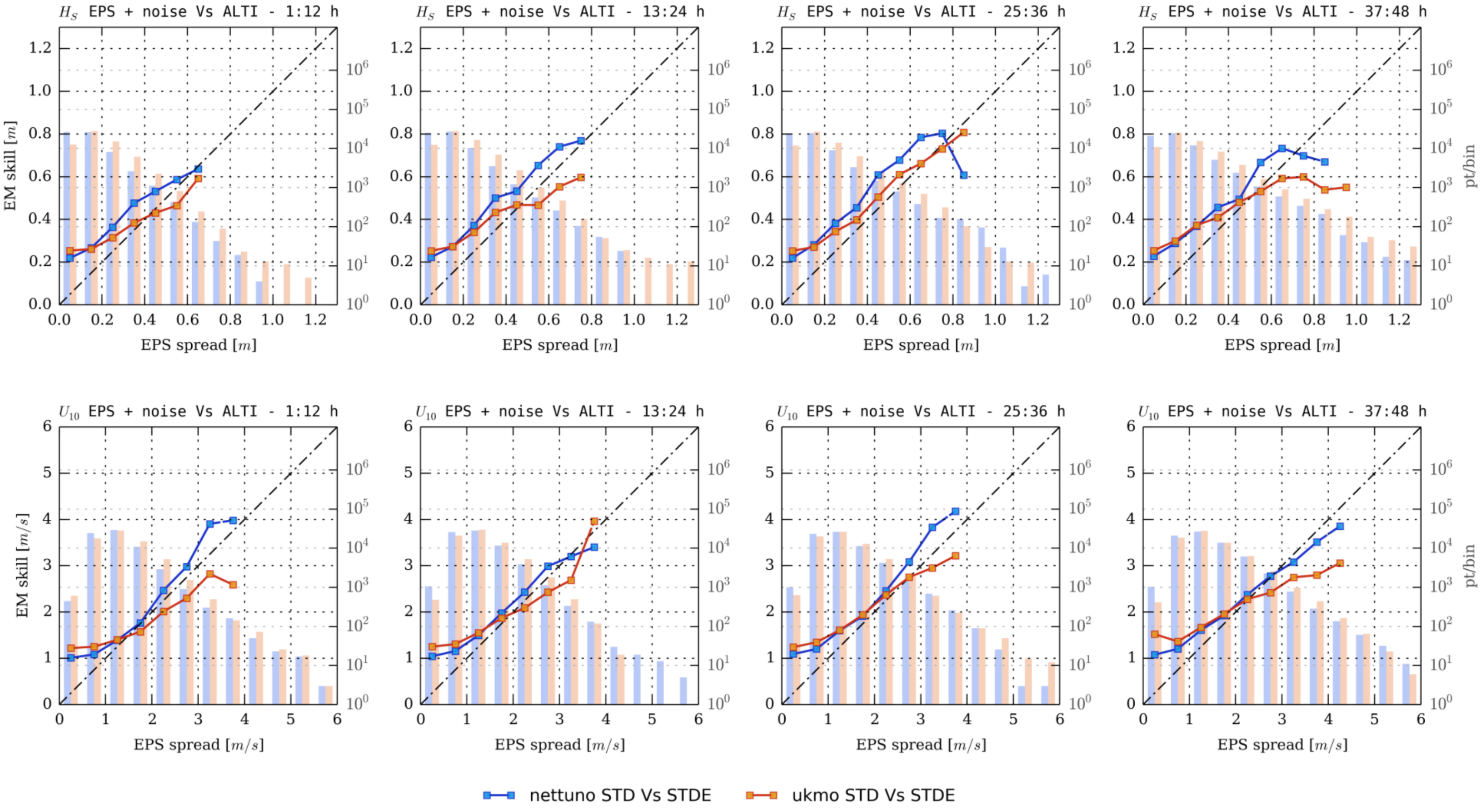


Figure 22: Spread-Skill relationships for H_S (top) and U_{10} (bottom) at altimeters location for different lead time ranges (left to right). Points: standard deviation of errors in corresponding ES bin; histograms: number of points per bin. Time window: 1/7 to 31/12 2013, coverage: Mediterranean Sea.

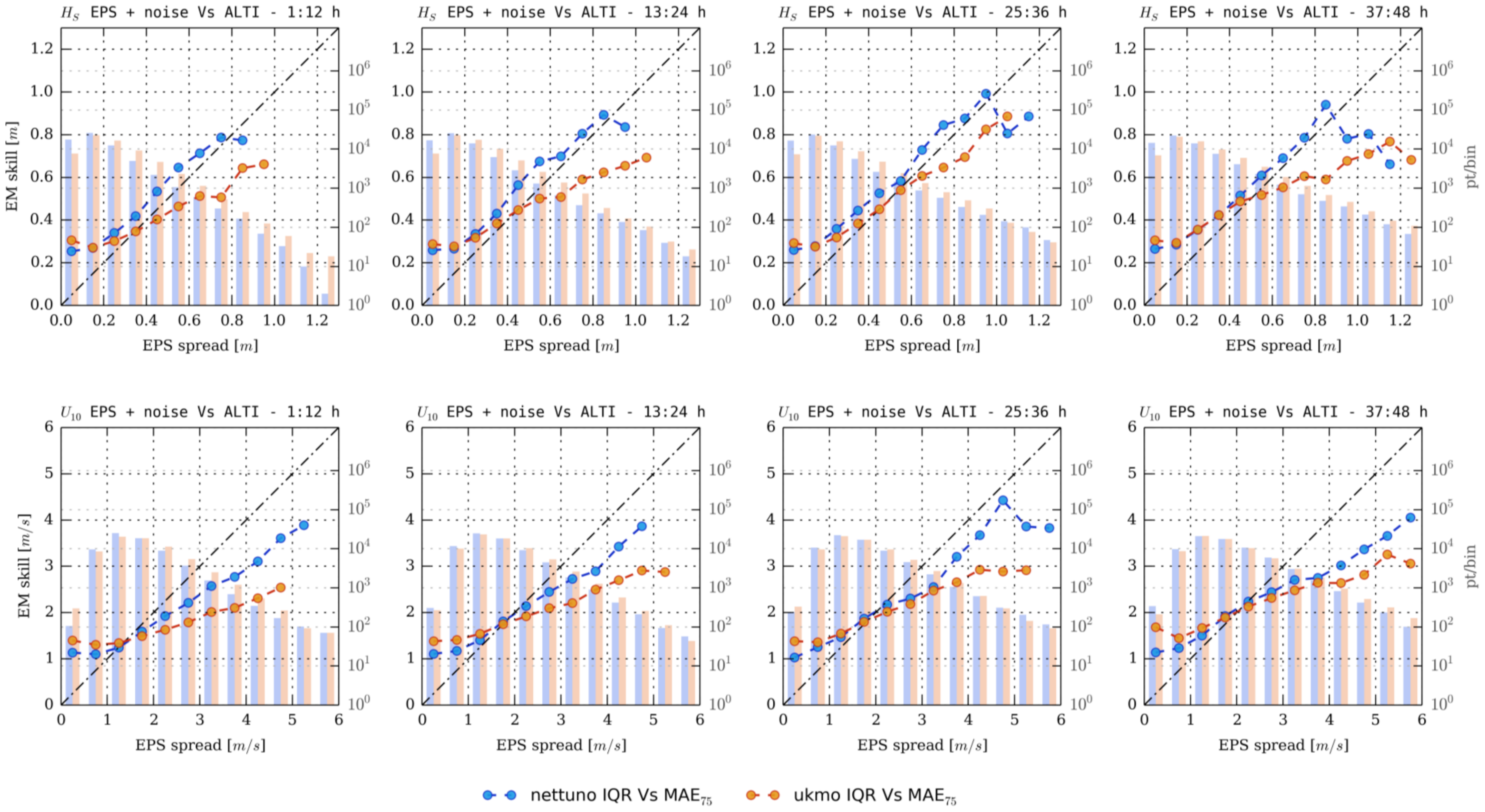


Figure 23: Same as Figure 22 but circles: 75th percentiles of absolute error in EPS$_{IQR}$ bin; histograms: number of points per bin.

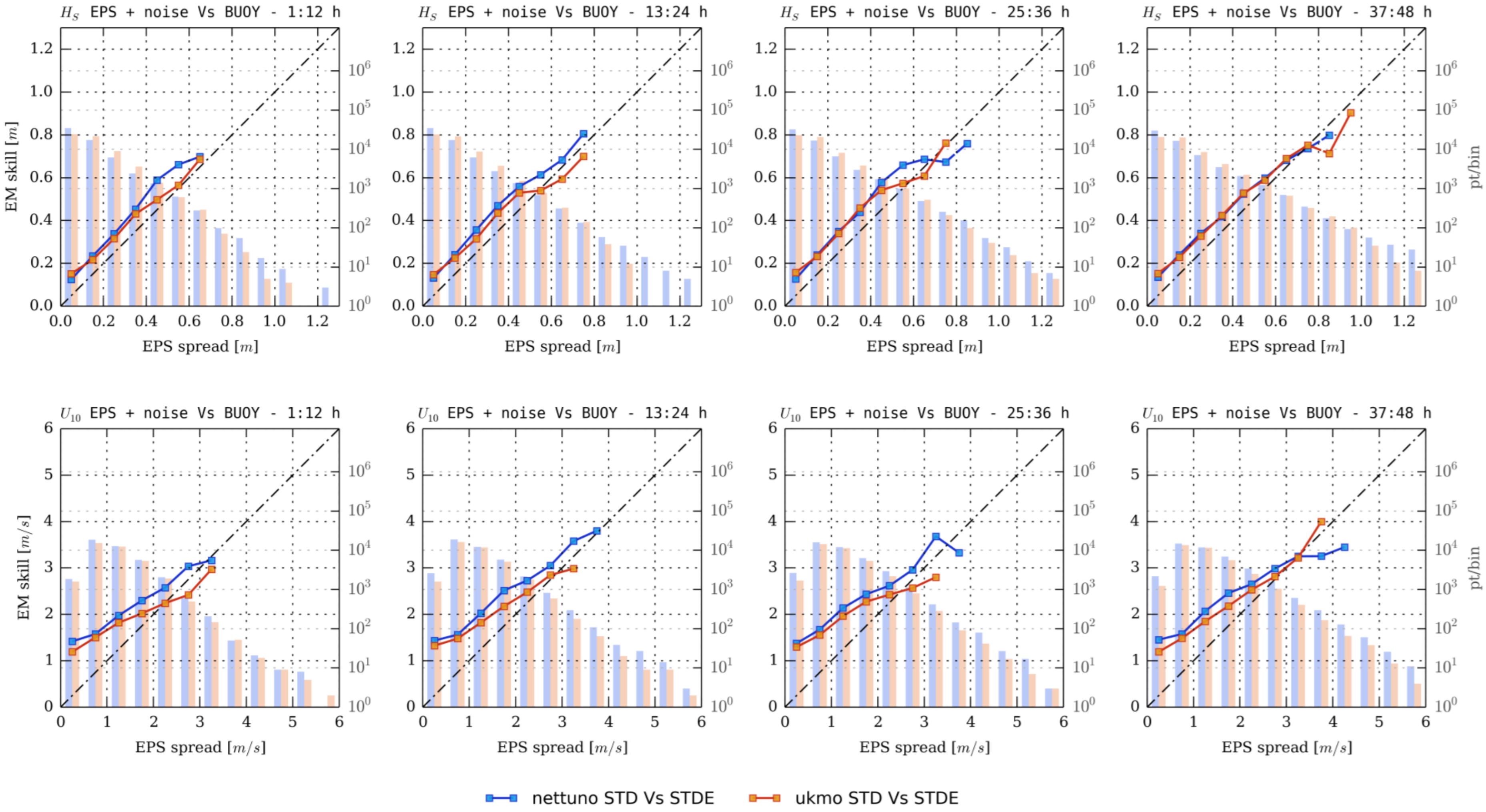


Figure 24: same as Figure 22 but for buoy observations.

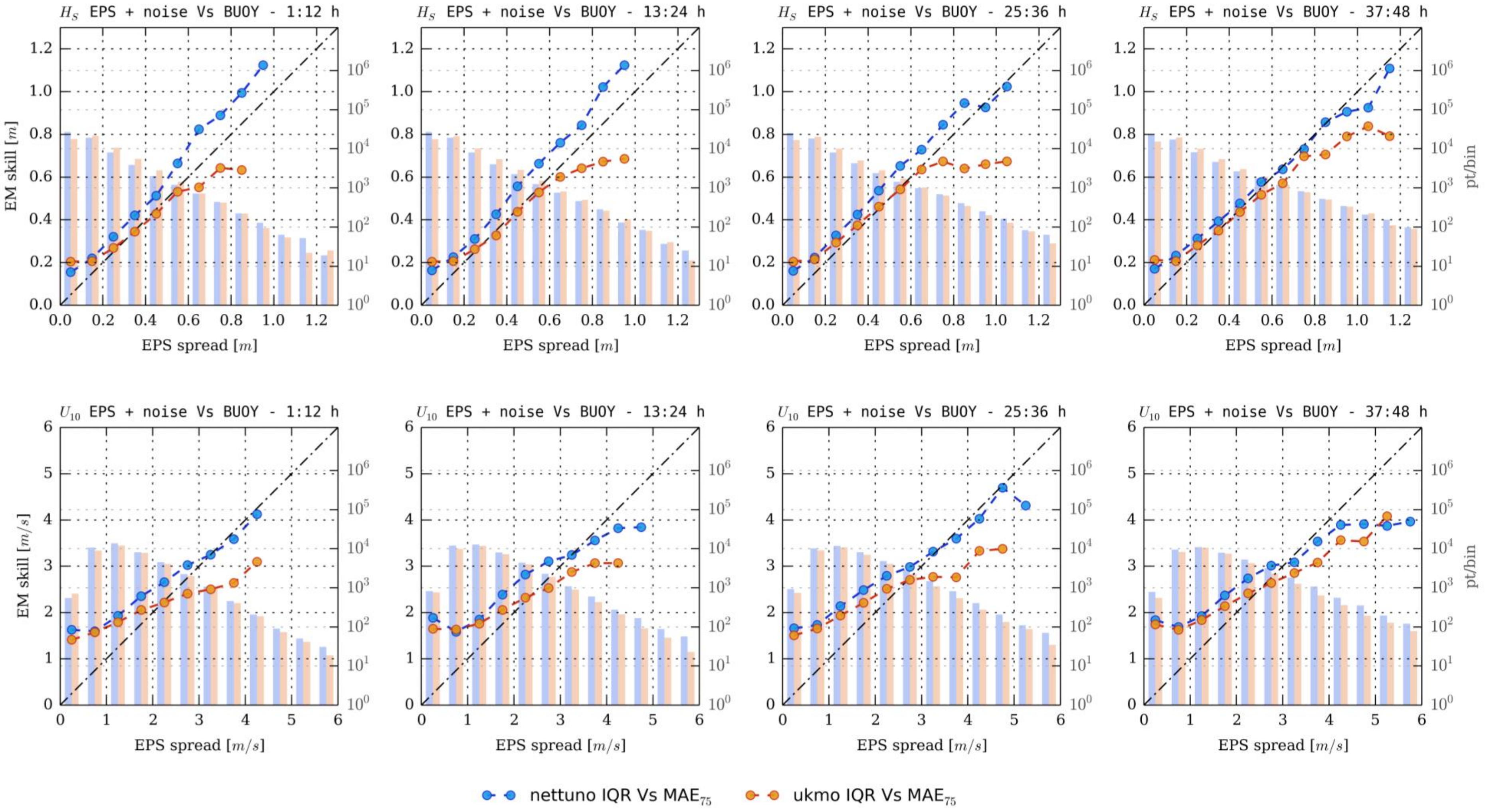


Figure 25: as Figure 23 but for buoy observations.

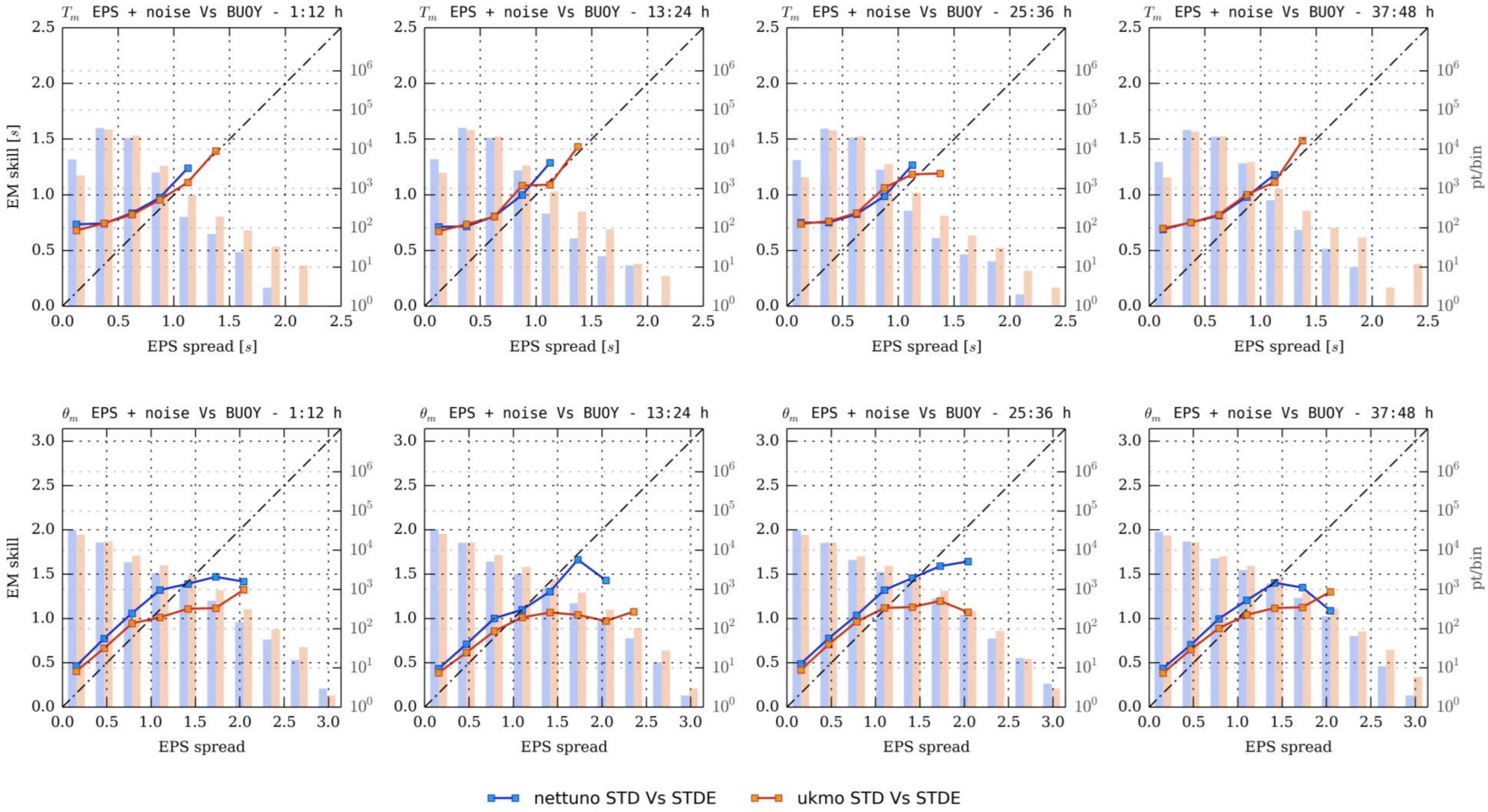


Figure 26. same as Figure 22 mean wave period (top) and mean wave direction (bottom) at buoy locations.

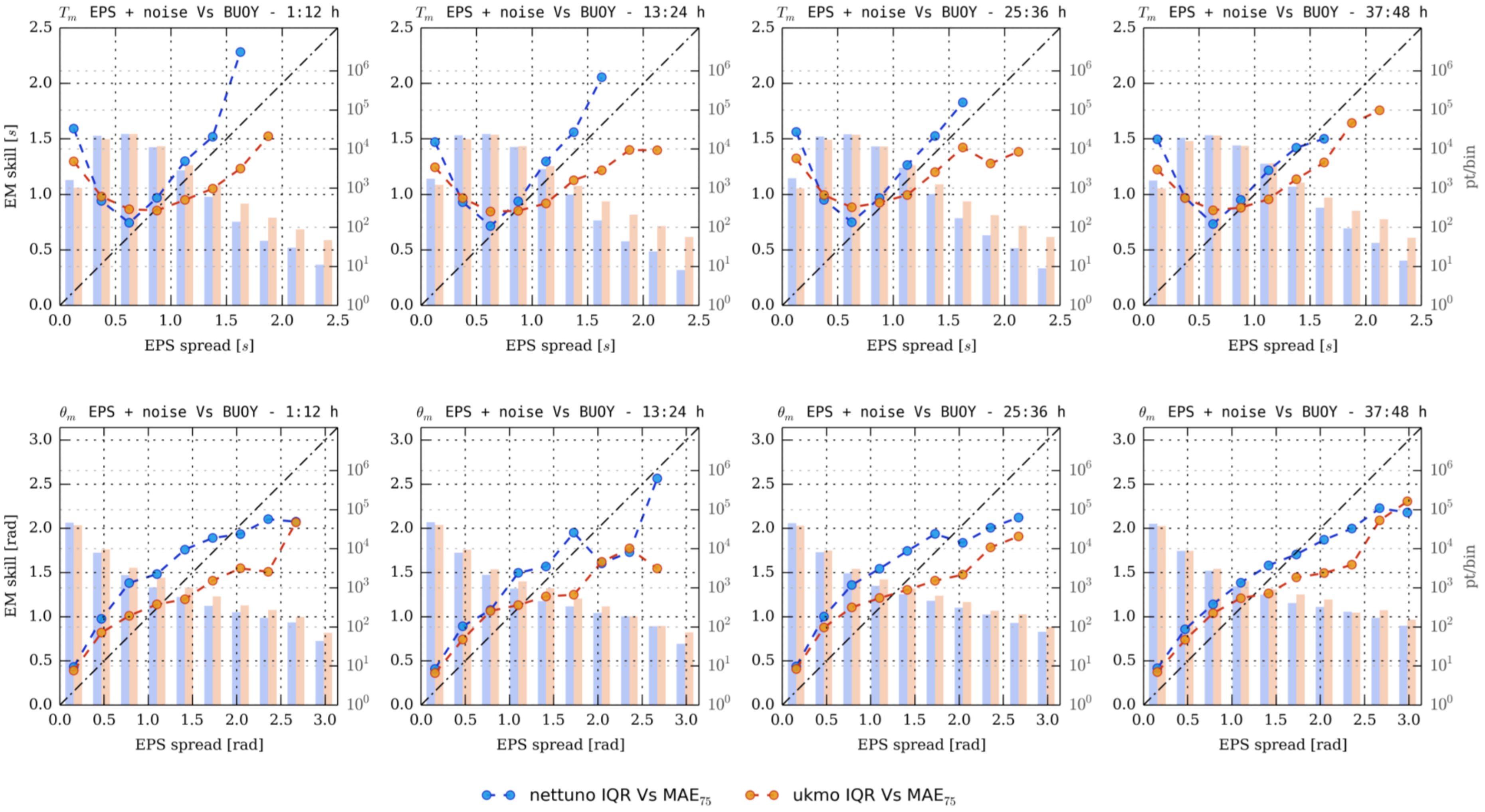


Figure 27: Same as Figure 26 but circles: 75th percentiles of absolute error in EPS_{IQR} bin; histograms: number of points per bin..

Table 3: Spread-skill linear relationships (weighted least squares fit) for significant wave height at altimeters location, according to (9). STD: EPS STD Vs STDE; SB: EPS$_{IQR}$ Vs 75th EM$_{AE}$ percentile.

	STD				**SB**			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	0.16	0.84	0.19	0.51	0.16	0.84	0.21	0.48
24 h	0.15	0.91	0.19	0.61	0.18	0.78	0.22	0.48
36 h	0.15	0.92	0.17	0.72	0.17	0.83	0.20	0.61
48 h	0.18	0.78	0.22	0.56	0.25	0.55	0.27	0.43

Table 4: As Table 3 but for wind speed.

	STD				**SB**			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	0.27	0.94	0.77	0.52	0.48	0.66	0.96	0.34
24 h	0.41	0.91	0.82	0.60	0.68	0.65	1.17	0.33
36 h	0.45	0.92	0.76	0.71	0.69	0.66	1.15	0.39
48 h	0.59	0.80	0.92	0.59	0.89	0.54	1.39	0.31

Table 5: As Table 3 but for buoys.

	STD				**SB**			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	0.06	1.14	0.09	0.94	0.08	1.06	0.12	0.69
24 h	0.08	1.10	0.09	0.94	0.08	1.06	0.13	0.68
36 h	0.08	1.05	0.10	0.95	0.12	0.87	0.17	0.59
48 h	0.09	0.95	0.09	0.94	0.14	0.73	0.12	0.74

Table 6: As Table 4 but for wind speed.

	STD				**SB**			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	1.03	0.73	1.08	0.55	1.25	0.60	1.33	0.38
24 h	0.96	0.84	0.95	0.70	1.30	0.62	1.29	0.47
36 h	1.15	0.72	1.10	0.64	1.34	0.63	1.36	0.46
48 h	1.10	0.72	0.92	0.73	1.39	0.56	1.25	0.54

Table 7: As Table 4 but for mean wave period.

	STD				**SB**			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	0.59	0.36	0.56	0.44	0.54	0.74	0.72	0.35
24 h	0.54	0.42	0.50	0.56	0.61	0.60	1.03	0.04
36 h	0.58	0.40	0.54	0.53	0.58	0.68	1.18	-0.02
48 h	0.55	0.44	0.49	0.59	0.69	0.47	0.98	0.14

Table 8: As Table 4 but for mean wave direction.

	STD				SB			
	Nettuno		**UKMO**		**Nettuno**		**UKMO**	
Lead Time	a	b	a	b	a	b	a	b
12 h	0.30	0.90	0.32	0.62	0.45	0.81	0.42	0.57
24 h	0.26	0.86	0.29	0.60	0.37	0.82	0.36	0.61
36 h	0.32	0.89	0.34	0.64	0.49	0.79	0.47	0.56
48 h	0.28	0.83	0.30	0.64	0.41	0.74	0.39	0.60

IV.5 Summary

In this section we have investigated and compared the ensemble spread characteristics of Nettuno and UKMO systems.

- We observe that UKMO spread is generally higher than Nettuno one, but the latter tends to develop faster than the former and, at least for wind speed, to reach it after 48 forecast hours.

- The spread information takes time to be transmitted from input (wind) to output (wave magnitudes), therefore the Nettuno waves spread remains lower than the UKMO one for the first 48 forecast hours. It is expected that the two ES would reach each other between 72 and 80 forecast hours.

- Even if substantial difference exists in the EPS width, both models capture the same number of observations within their extreme members. Therefore both systems output (waves) result under-spread.

- A detailed analysis show that both systems are under-spread where they should mimic low forecast uncertainty (with UKMO slightly more under-spread than Nettuno), and generally over-spread where the uncertainty is higher. In this error region UKMO is slightly more over-spread than Nettuno, and the behaviour of the latter is sometimes in line with the deterministic error distribution.

V EPS AS PROBABILISTIC FORECAST

This section investigates the EPS capacity in predicting events probability. We present the comparison between Nettuno and UKMO EPS systems in terms of detailed view of forecast-observations joint distributions (reliability diagrams §V.1). In §V.2 we compare the integral probability skills expressed as continuous ranked probability score breakdown. Paragraph V.3 shows information on the resolution of both systems derived forecasts: dressed control member, naked ensemble, and dressed ensemble. Finally (§V.4) we give estimates of the potential economic value of the two ensemble systems.

V.1 Reliability Diagrams

Reliability diagrams (RD) are graphical tools for the estimation of both reliability and resolution for probability forecasts of a dichotomous predictand, *e.g.* in Figure 28 (top left) we investigate the skill in predicting H_S>0.9m. A RD depicts the joint distribution of forecasts and observations. More specifically, it represents the so called *calibration function* related to the predictand, that is the probability of the observed frequency, *z* , given the forecast probability *p(z|y)*, as a function of the forecast probability *y* (see Figure 28). The plot is enriched with an histogram representing the *refinement distribution*, *p(y)*, *i.e.* the frequency of use of the possible forecasts.

Perfect skill is given by calibration function lying on the diagonal, together with u-shaped refinement distribution. The shape of the refinement distribution (histogram) is a measure of confidence: the wider the distribution, the higher the confidence on the calibration function. Points lying above (below) the diagonal are symptoms of under (over) forecasting.

In Figure 28 we present four couples of reliability diagrams. Each couple is made with one diagram for Nettuno (left), and one for UKMO (right). Top row accounts for buoy measurements, and bottom row for altimeter observations. Left (right) couples depict the reliability for the forecast day 1 (day 2). The prediction considered in Figure 28 is related to the forecasting probability of significant wave height being higher than the median of the measurement distributions, while Figure 29 we set this limit to the upper quartiles. The same exercise is made for wind speed in Figure 30 and Figure 31, for U_{10} higher than the 50th and 75th percentiles.

The common shape which is visible in (almost) all the given plots states that both EPS are generally overconfident, and lack in resolution. UKMO tends to under-forecast also in the upper tail, especially when comparing forecast with buoy observed H_S distributions (Figure 28 and Figure 29). With respect to the probabilistic prediction of wind speed, the skill of the two systems are very similar (see Figure 30). On the contrary, Nettuno lacks of reliability in some cases when predicting significant wave heights probabilities.

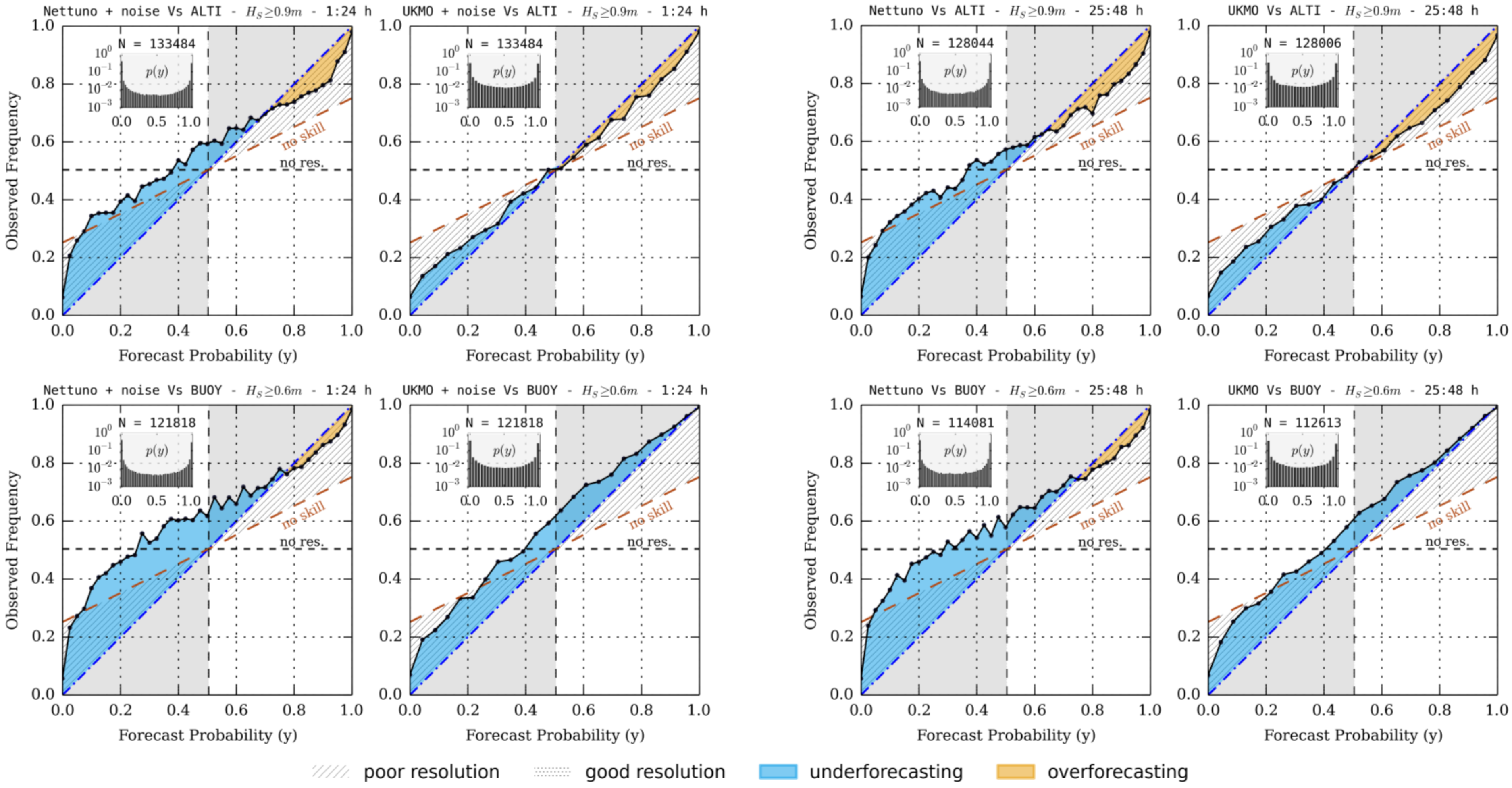


Figure 28: Reliability diagrams for mean wave period at altimeter (top) and buoy (bottom) locations for forecast day 1 (left) and day 2 (right). Thresholds: 50th percentile of underlying observation distributions.

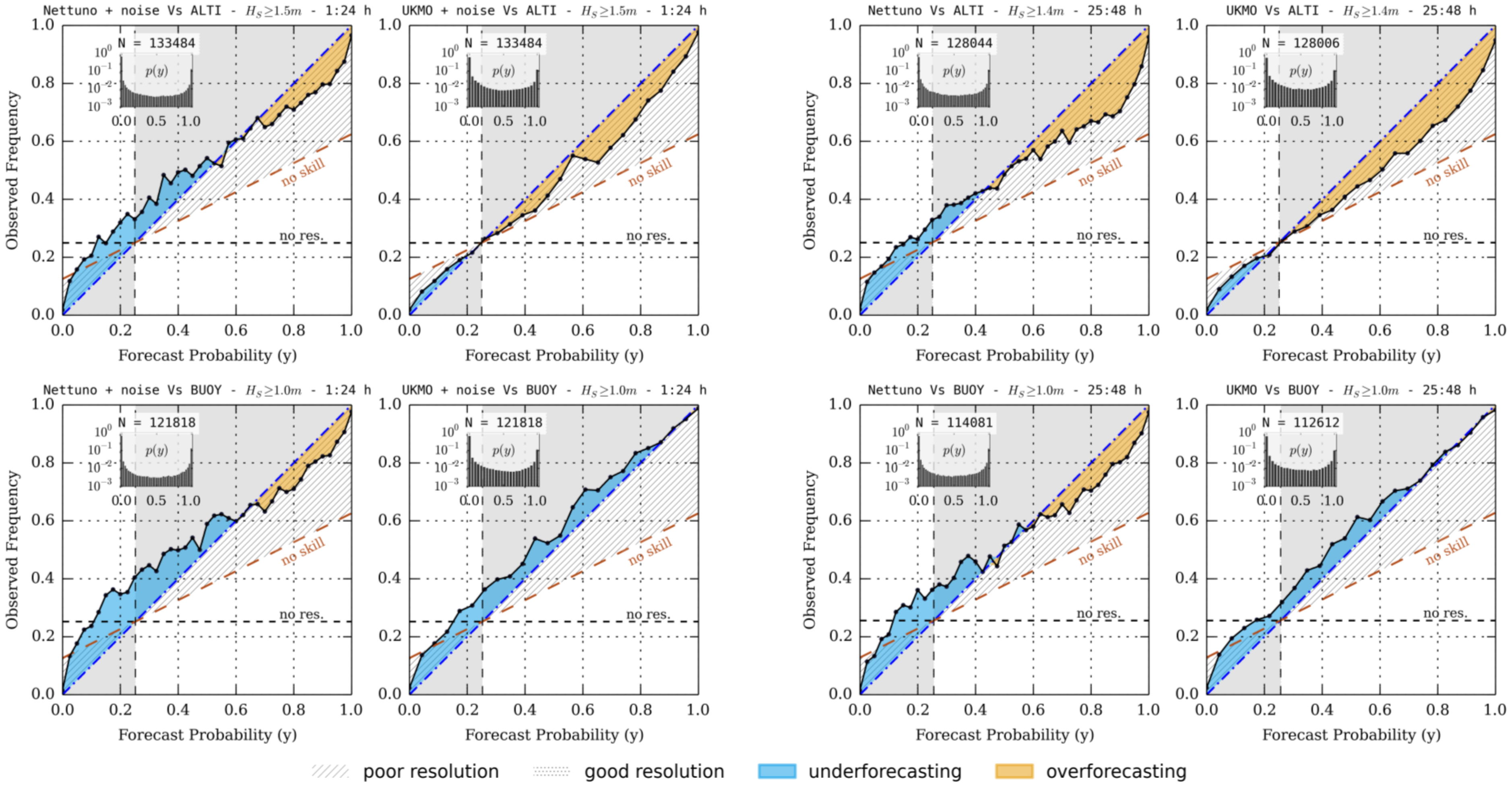


Figure 29: same as Figure 28 but for 75th percentile of underlying observations.

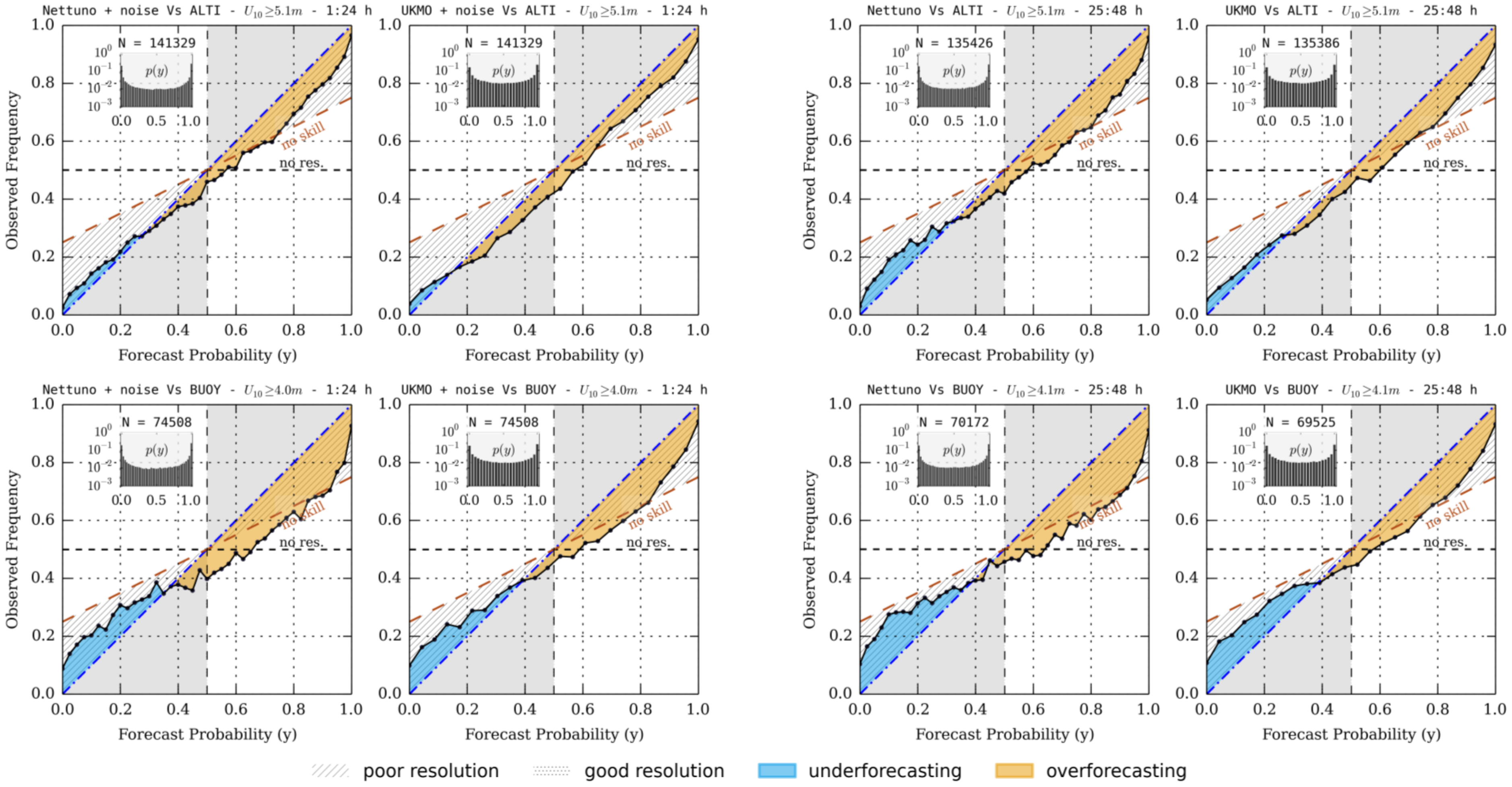


Figure 30: same as Figure 28 but for wind speed.

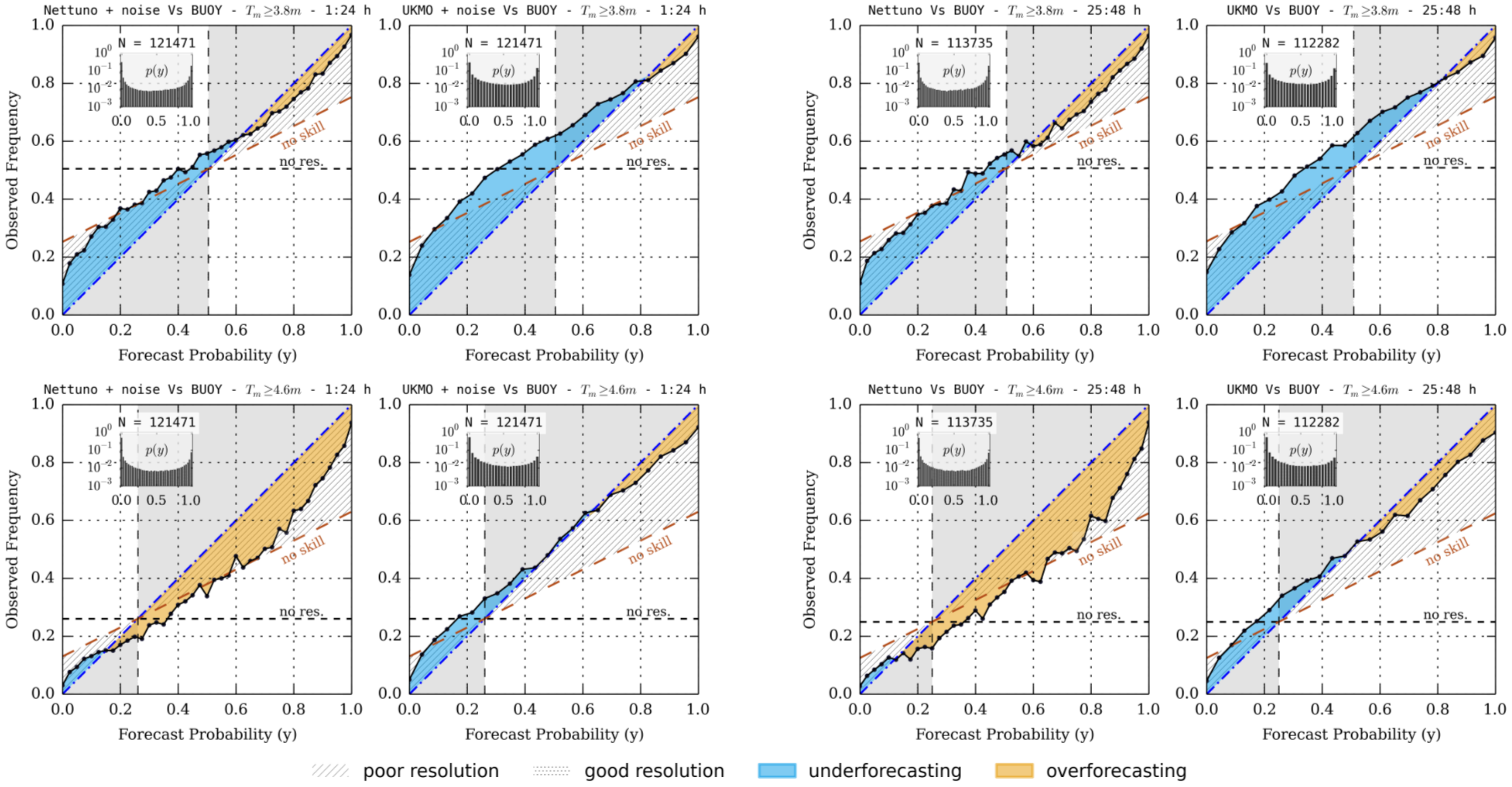


Figure 31: same as Figure 29 but for wind speed..

V.2 Continuous Ranked Probability Score

Reliability diagrams, and their integration (Brier Score) are case dependent. In other words they are inspection of the forecast behaviour on a single dichotomous predictand. If there's the need to measure the probability prediction skill with respect to a number of predictands, one would evaluate a ranked probability score, summing up over a number of classes. This procedure is case dependent too. One natural approach for the generalization to a universal skill is to count for an infinite number of infinitesimal classes, *i.e.* to extend the ranked probability score to the continuous case (Wilks, 2011). The result is the continuous ranked probability score, defined by the following integral:

$$CRPS = \int_{\Re} \left[F(x) - H(x-o) \right]^2 dx \tag{10}$$

where F(x) is the cumulative distribution predicted by the EPS. H(·) is the Heaviside function which maps negative argument to 0, and the rest to 1. CRPS values range from 0 (ideal skill) up to infinity (there's no upper bound). The lower the CRPS, the better the skill. Ingredients for a good skill are the narrowness of the predicted F, together with its centring with respect to the observation *o*. Results are usually given as an average over a number of observation-forecast couples. Therefore, in the following with *CRPS* we indicate the arithmetic average over all the scores computed in a homogeneous sub-sample.

Hersbach (2000) shows that average CRPS has an algebraic decomposition into reliability and potential CRPS (which, in turn, can be decomposes into resolution and uncertainty). Potential CRPS corresponds the score of the forecast if it were perfectly reliable, *i.e.* if the reliability were equal to zero. Figure 32 shows this breakdown for the CRPS of Nettuno (blue) and UKMO (red) predicted wind speed probabilities, distinguishing the results related to buoy and altimeter observations. The scores related to wind significant wave height is depicted in Figure 33.

In the comparison against altimeter observations, the reliability scores of Nettuno and UKMO predicted U_{10} probability are very close to each other, as it was somehow appreciated (but for a single case) in Figure 30. In the comparison against buoy measurements, UKMO average reliability score beats the Nettuno one more clearly (Figure 32). Compared to the ones obtained for wind speed, the scores related to significant wave height are one order of magnitude lower (Figure 33). In this case, the Nettuno potential (average) CRPS is always slightly better than the UKMO one (right panel). The differences in average reliability of forecast H_S probability are such that the potential score is compensated at altimeter locations (as it occurs for wind speed). At buoy locations, Nettuno's reliability score is lower (better) than its counterpart one.

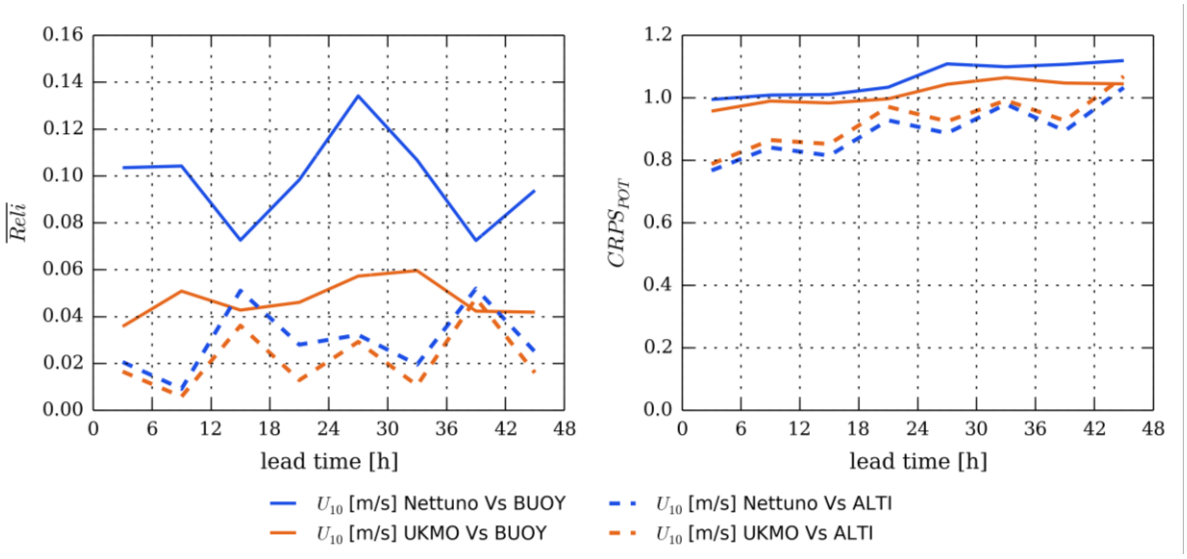


Figure 32: Breakdown of CRPS for wind speed forecast. Left: reliability; right: potential CRPS. Dashed lines: altimeter; continuous lines: buoy. Time windows: 1/12/2013 to 31/12 2013. Coverage: Mediterranean Sea.

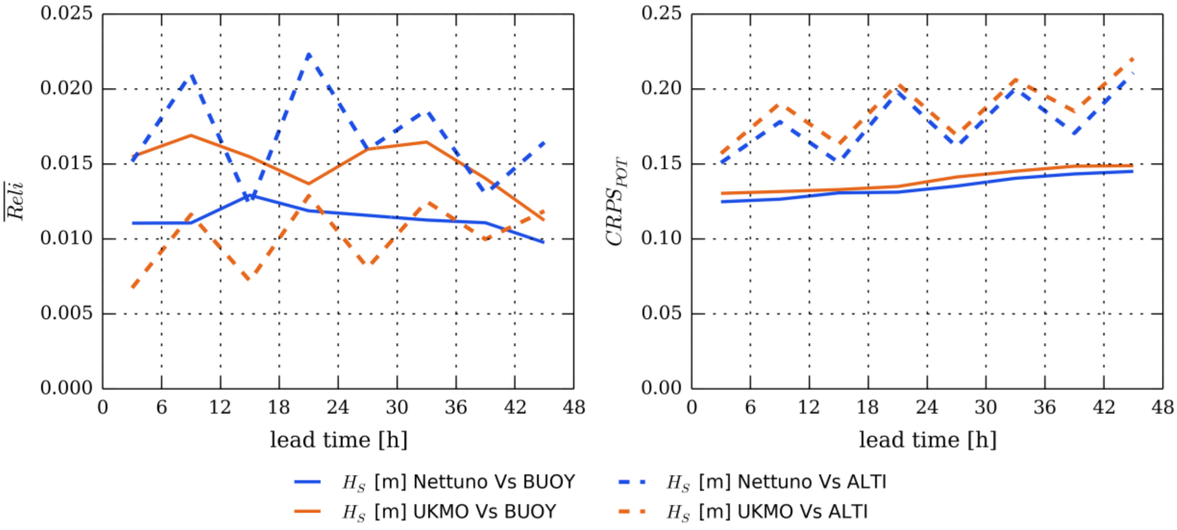


Figure 33: Same as Figure 32 but for significant wave height.

V.3 Receiver Operating Characteristic

Receiver Operating Characteristic diagrams are a useful tool for both forecasters and decision maker users (see e.g. Wilks, 2011). A point in the diagram is given by the couple hit rate (H, ordinate) versus false alarm rate (F, abscissa). H is evaluated as the percentage of correct forecasts with respect to the number of times the event has been forecast. F is the percentage of wrong forecasts with respect to the observations sub-sample in which the event was not observed. The best that one expects from a forecast is to have H = 1 and zero false alarms, and a forecast based on purely random choice (in the climatology distribution) would have same false alarm and hit rates. On the basis of a probability forecast for a predictand, the decision maker has to make a choice between an action being preferred if the event occur, and another action being preferable if the event does occur. From the forecaster point of view the ROC is a measure of the forecast resolution, *i.e.* the ability to discriminate between frequencies of occurring and non-occurring events.

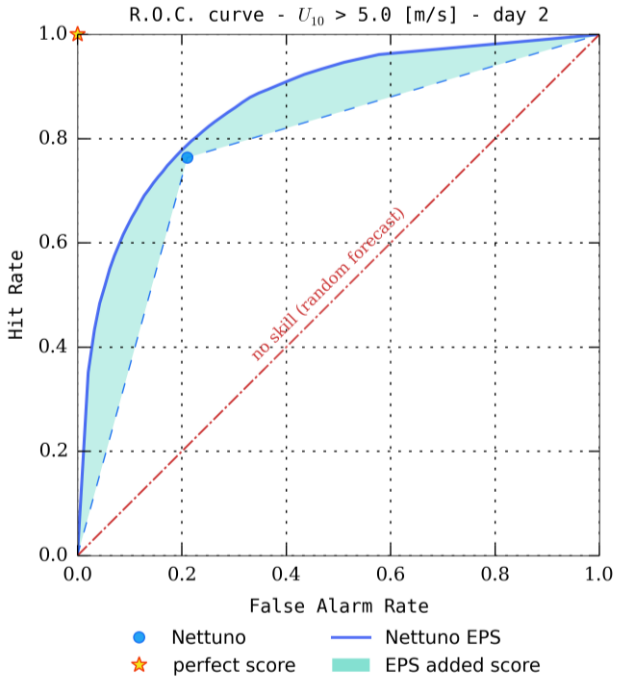


Figure 34: Example of Receiver Operating Characteristic diagram.

Consider the figure on the left as an example. The blue circle (F=0.2, H=0.75) denotes the couple H Vs F for a *deterministic* forecast (Nettuno in this case) skill in predicting U_{10}>5m/s. If the forecast is made up with an ensemble system, one can plot N points, one for each ensemble member, on the F, H plane. These point define a ROC curve for a single predictand (continuous blue line) which can be completed by adding (0,0) and (1,1) without altering the score. In the same way the *deterministic* single point can be connected to the two extremes to define a “curve”. The shaded area, In the figure on the left, denotes the ROC skill improvement, at each false alarm rate. This is, for the decision maker, the added feature if he makes a choice based on the ensemble forecast, instead of the deterministic one.

A natural choice for the comparison between the two curves is the integration (sum) over false alarm rate. The area under the R.O.C. curve (A_{ROC}) is, in this sense, a valuable skill indicator. In fact, the ROC area for a perfect forecast is equal to 1, the worst purely random forecast has A_{ROC}=½. All other prediction would take ROC values in the upper left triangle, thus intermediate area values between 0.5 and 1. Nevertheless, keeping random forecast as reference, the common practice ROC standardized skill score is:

$$SS_{ROC} = 2A_{ROC} - 1 \qquad (11)$$

which is linear with respect to the score, but it has the advantage that it is bounded between 0 (worst case) and 1 (perfect forecast).

Figure 35 shows the SS_{ROC} for Nettuno (top) and UKMO (bottom) ensemble forecast systems (central columns) relative to the prediction of significant wave heights exceeding a certain threshold at buoy locations. Right and left columns allow comparison between the naked EPS and a dressed version of control member (left) and the ensemble itself (right). Each one of the plots depict (in colour scale) the SS_{ROC} at various forecast lead time (ordinate) and thresholds (abscissa). In Figure 36 we give the same results but for altimeter locations, while wind speed skills are shown in Figure 37 and Figure 38. The comparison with the dressed form of the control member, and the dressed ensemble is very useful. In fact the number of members in the UKMO system is much lower than Nettuno, and the ROC is very sensitive to the ensemble size. In general, Nettuno SS is everywhere higher than UKMO. What is worth to underline is that the three UKMO forecasts show a drastic drop of ability in discriminating H_S higher than 3m at buoy locations (Figure 35).

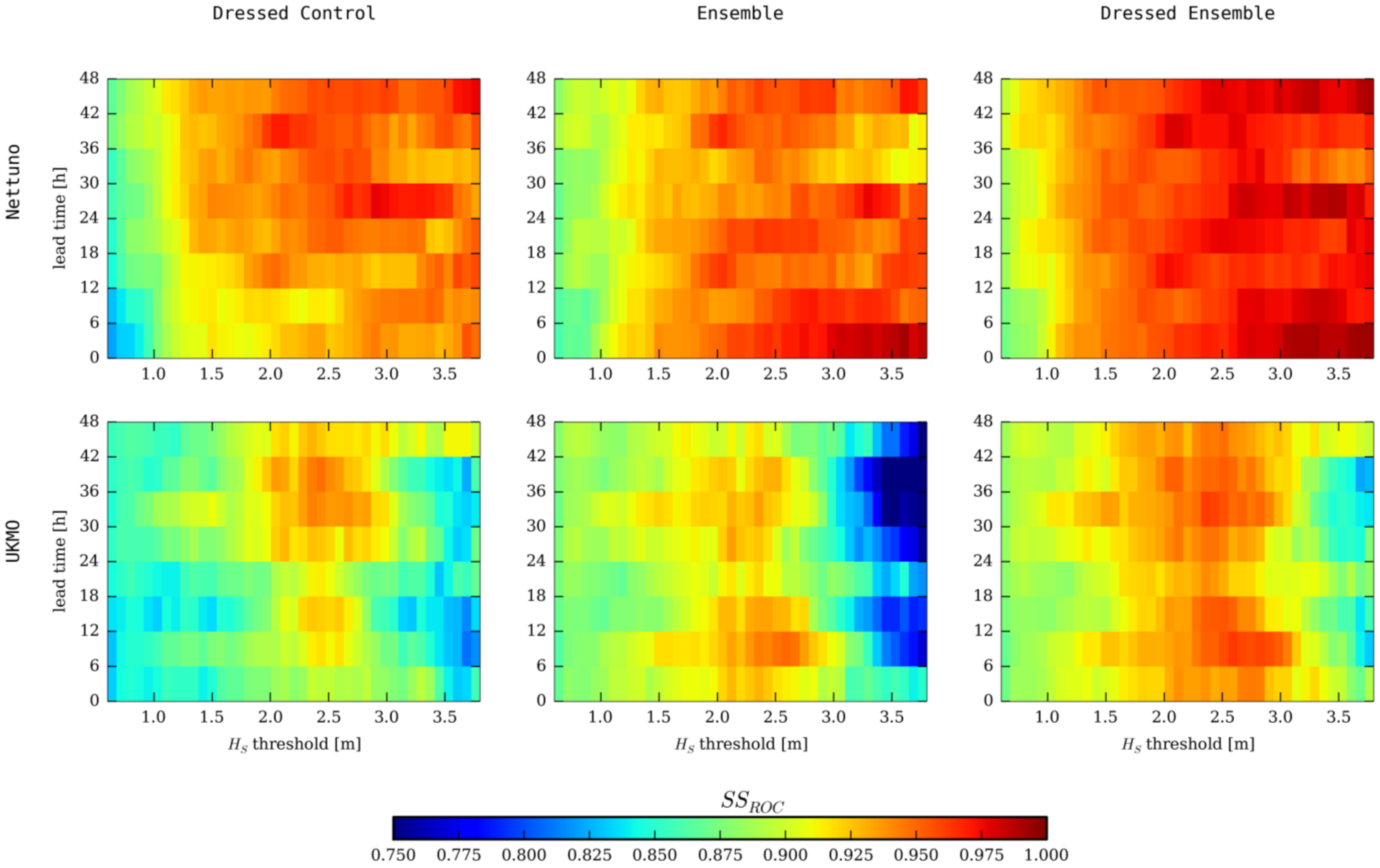


Figure 35: ROC Skill Score for various significant wave height thresholds (abscissa) and forecast lead time bins (ordinate). Forecasts are compared with buoy measurements.

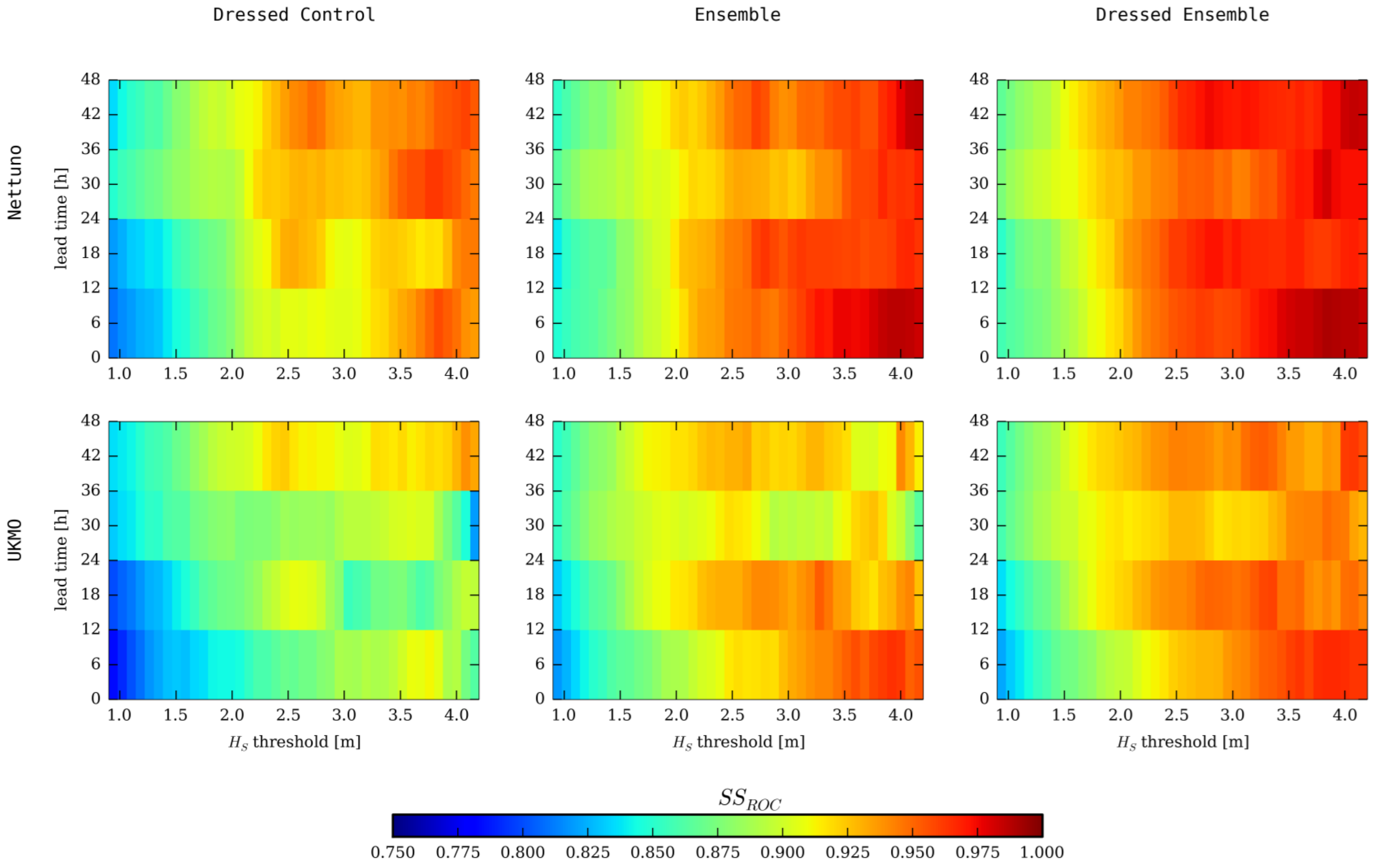


Figure 36: same as Figure 35 but for altimeter observations.

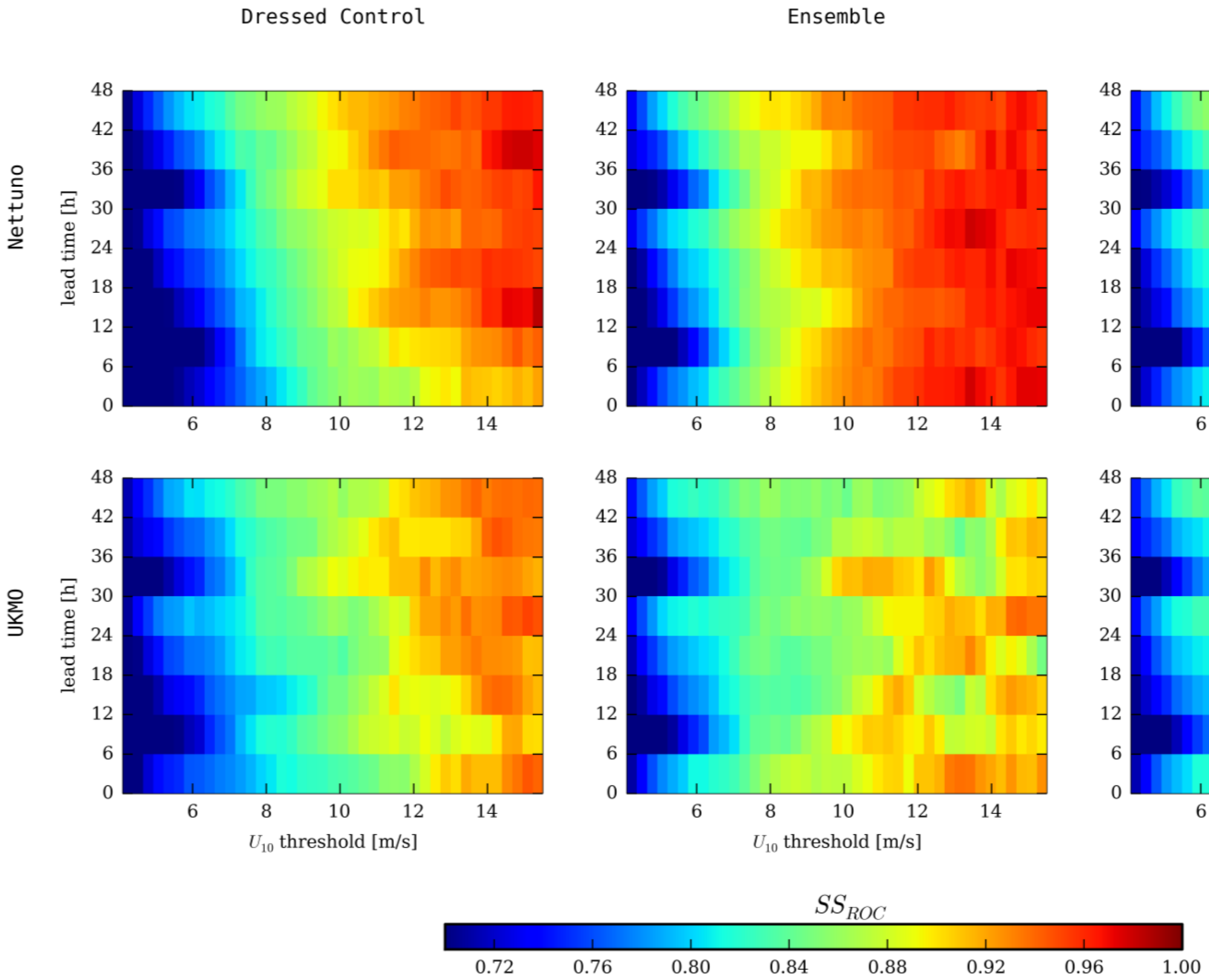


Figure 37: same as Figure 35 but for wind speed.

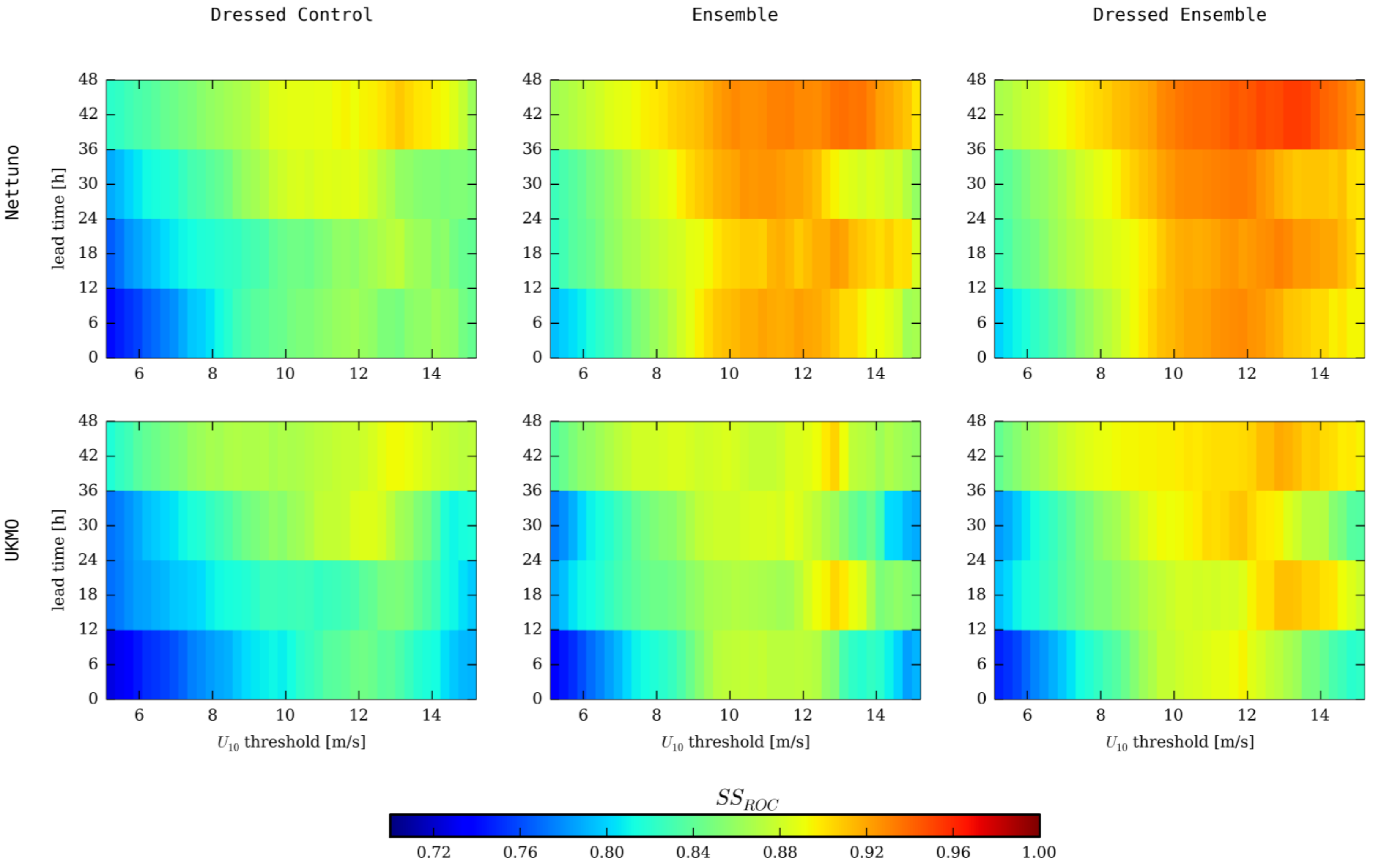
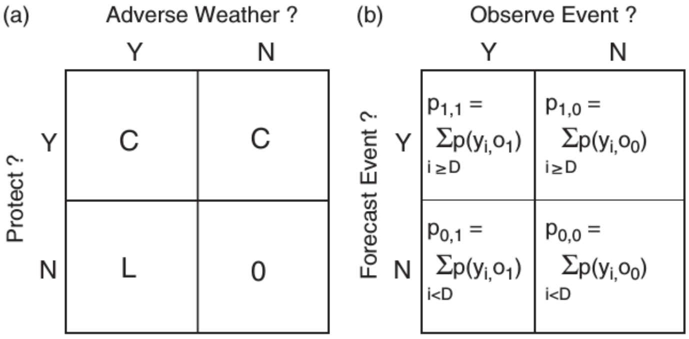


Figure 38: same as Figure 37 but for altimeter measurements.

V.4 Relative Economic Value

This last metric is designed more for the user (decision maker) than for the forecaster. Anyhow it is useful for both. Relative Economic Value of a forecast is based on a simple cost/loss model, the static cost-loss model. The cost/loss decision problem relates to a hypothetical user who has to make a decision between the choices of protecting or not against some kind of adverse weather, in terms of its economic effects. The scheme for the construction of the metric is depicted in Figure 39: it is a combination of a 2x2 contingency table for a dichotomous predictand, and the related 2x2 cost/loss decision table. The cost of protection is C, the losses for non protecting against an event that occurs are denoted with L, and α=C/L is the cost/loss ratio which takes values from 0 to 1.



(a)

Protect ?	Adverse Weather ? Y	Adverse Weather ? N
Y	C	C
N	L	0

(b)

Forecast Event ?	Observe Event ? Y	Observe Event ? N
Y	$p_{1,1} = \sum_{i \geq D} p(y_i, o_1)$	$p_{1,0} = \sum_{i \geq D} p(y_i, o_0)$
N	$p_{0,1} = \sum_{i < D} p(y_i, o_1)$	$p_{0,0} = \sum_{i < D} p(y_i, o_0)$

Figure 39: (a) Loss function for the 2 by 2 cost/loss ratio situation. (b) Corresponding 2 by 2 verification table resulting from probability forecasts characterized by the joint distribution p(y_i, o_j) being transformed to non-probabilistic forecasts according to a particular decision maker's cost/loss ratio. [Wilks, 2011]

The relative economic value of the protection/non-protection choice based on a forecast can be derived on the basis of the above hypotheses, and it is expressed by Richardson (2000) in terms of non-dimensional quantities as:

$$REV = \begin{cases} 1 - F - \left(\dfrac{\bar{o}}{1-\bar{o}}\right)\left(\dfrac{1-\alpha}{\alpha}\right); & \alpha < o \\ H - \left(\dfrac{1-\bar{o}}{\bar{o}}\right)\left(\dfrac{\alpha}{1-\alpha}\right); & \alpha > o \end{cases} \qquad (12)$$

with ō the observed relative frequency of the event for which a decision has to be made, F and H are false alarm and hit rates, and α is the user dependent cost/loss ratio.

The black dashed line in Figure 40 depicts REV for a user making a decision of protecting against H_S higher than 1 m, based on the actual operational Nettuno forecast (control member). If the Nettuno EPS is used, than each ensemble member has an independent REV, and the best choice for the user is to choose the outcome for the member with the maximum REV. The optimal choice is described by the continuous green line (Figure 40). Each EPS member has an associated threshold probability (grey stairs in the background), *p*, which is the forecast probability associated to the optimal protection choice. The green shaded area in Figure 40, denotes the improvement in considering the EPS forecast (instead of the *deterministic* one) as the source for decision making. Please note that the REV of a perfect ensemble does not equal unity.

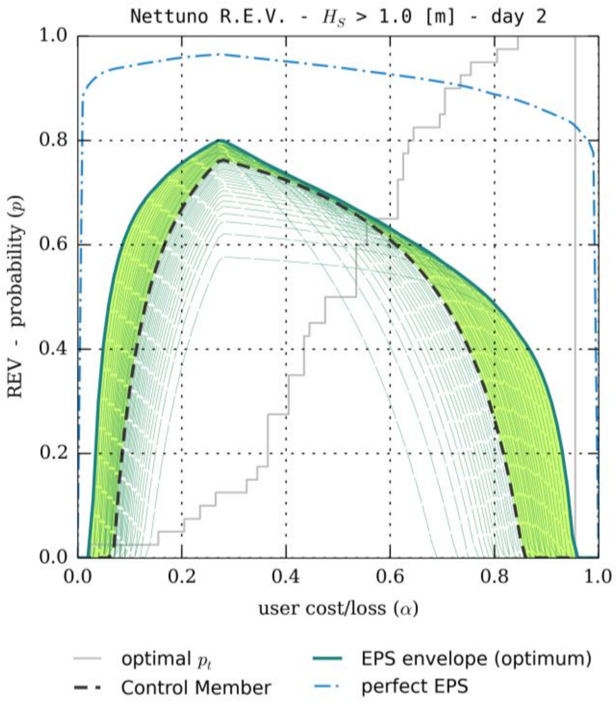


Figure 40: Example of Relative Economic Value for an EPS.

Figure 41 compares Relative Economic Value diagrams of Nettuno and UKMO ensemble systems, for decision makers wishing to investigate the convenience of protecting against significant wave height exceeding 1, 2, or 3 m. The metric is given with reference to buoy observations only. Left plot in Figure 41 is designed for the user who must make a decision within 24 hours from forecast time, and the right plot is for the forecast day 2. Figure 42 replicates the same results at altimeter locations.

It turns out that, for high cost/loss ratio users who want to make decisions on low energetic events, the UKMO forecast is more convenient. Especially if the will is to make a decision in the first 24 forecast hours, for actions to be taken close to the buoys region (mainly close to the coast).

Nettuno eps forecast generally fits better to all other users, and particularly to a decision maker who must make a protection/non-protection choice with respect to high energetic sea states, and whose costs of protection are lower than potential losses.

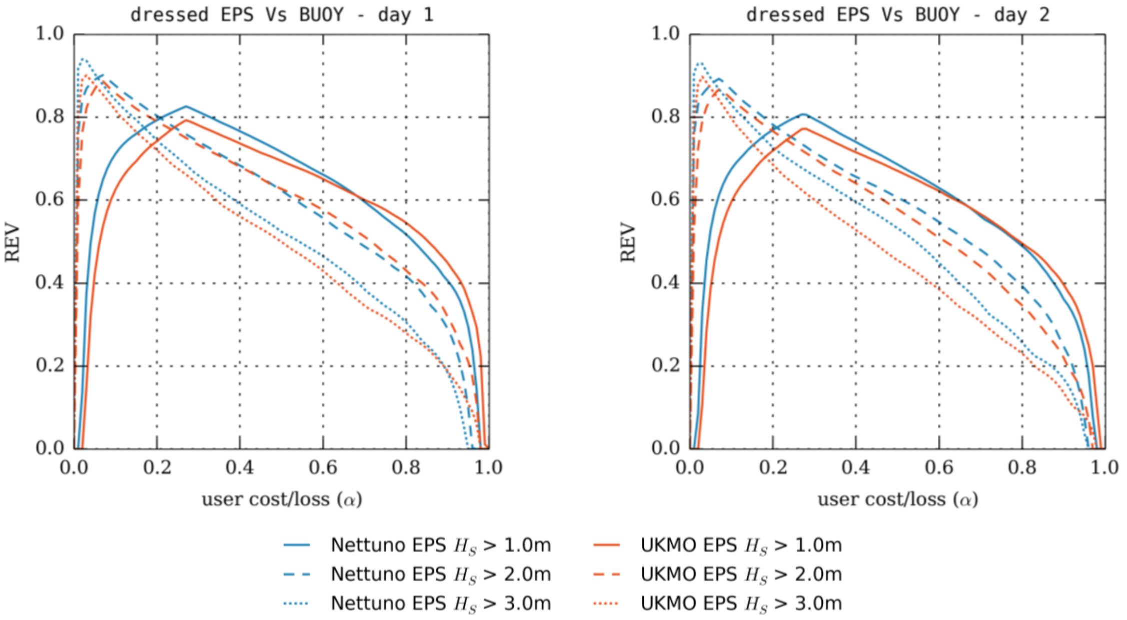


Figure 41: Relative Economic Value diagrams for Nettuno (blue) and UKMO (red) systems, associated to the forecast of significant wave height exceeding the thresholds in the legend at forecast day 1 (left) and day 2 (right), at buoy locations.

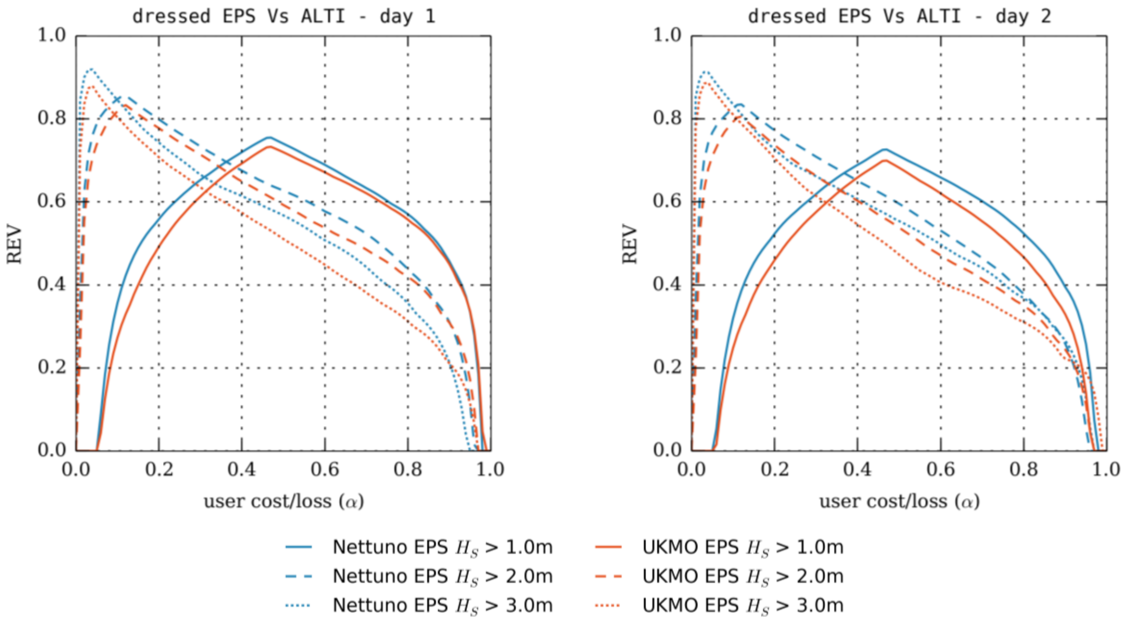


Figure 42: Same as Figure 41 but for altimeter observations.

This verification measure is very sensitive to the number of EPS members. Generally, the higher the number of members, the higher the economic value. Probably this is the main reason for the difference between the two models score. In order to get rid of this doubt, the REV diagram in Figure 42 (left) is reproduced for limited number (N, the same for both systems) of randomly chosen EPS members. As expected the score of both models decreases with N. However, the difference between Nettuno and UKMO persists for almost all cost/loss ratios, except in the low *α* region where the gap could be recovered if UKMO EPS had the same size of Nettuno EPS.

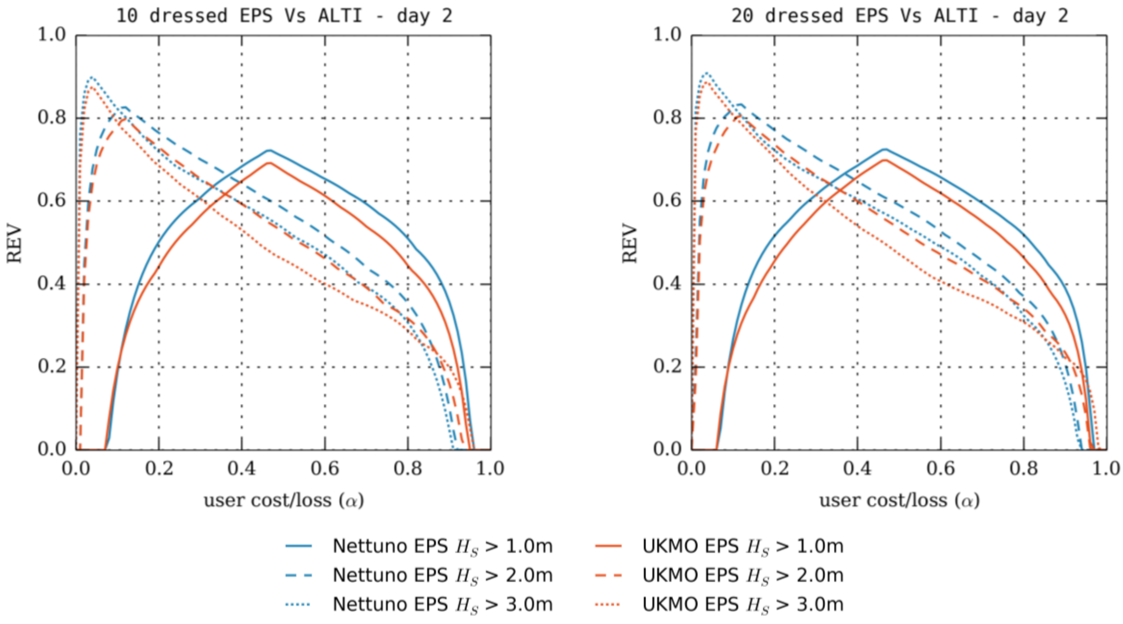


Figure 43: Effect of ensemble size on REV. Results for N randomly chosen ensemble members. N = 10 (left), N = 20 (right). N is the same for Nettuno and UKMO.

V.5 Summary

In this section we have investigated and compared Nettuno and UKMO EPS capacity in predicting events probability.

- Based on a number of selected cases we observe that UKMO outputs are slightly more reliable than Nettuno ones. The integrated scores are similar, with differences depending by the observations sub-sample. Anyhow, the average UKMO reliability is somehow lower (better) than Nettuno one.

- On the other hand, receiver operating characteristic skill scores (referred to common climatology) indicate that Nettuno resolution is superior to UKMO one. In particular, the latter lacks in ability to discriminate occurrence/non-occurrence of high energetic events, especially at buoy locations (mainly distributed close to coastal regions).

- Nettuno eps forecast generally fits better than UKMO to decision makers who must make a protection/non-protection choice with respect to high energetic sea states, and whose costs of protection are lower than potential losses. While UKMO is in some cases more suitable for high cost/loss ratio and low energetic, early occurring events. The ensemble size influence on the differences in REV is negligible.

VI SUMMARY AND CONCLUSIONS

In this report we have described the processes of verification and inter-comparison of Nettuno and UKMO ensemble prediction systems, the two EPS for the Mediterranean Sea that have been independently developed and implemented within the framework of MyWave project, work package n°3, by U.K. MetOffice and CNMCA (Italian Meteorological Service). The validation has been pursued with respect to a rich observations dataset coming from moored buoys, satellite altimeters and scatterometers.

We have found that, is the bias is removed, the ensemble mean is generally more skilled than a *deterministic* forecast, if the EPS is built with the same model grid resolution. However, the analysis of both wave model input and output bias shows that the EMs are more energetic than the control members, and this is partly due to the way in which the results are post-processed to get the EPS statistics. In *deterministic* sense, we have observed that UKMO has a better ability to mimic the overall climatology distribution, and Nettuno tends to overestimate the distribution upper tail. On the other hand, Nettuno is better correlated with the observations than UKMO, and this aspect compensates the skill gap due to the higher variability.

UKMO spread is generally higher than Nettuno one, but the latter tends to develop faster than the former and, at least for wind speed, to reach it after 48 forecast hours. The spread information takes time to be transmitted from input (wind) to output (wave magnitudes), therefore the Nettuno waves spread remains lower than the UKMO one for the first 48 forecast hours. It is expected that the two ES would reach each other between 72 and 80 forecast hours. Even if substantial difference exists in the EPS width, both systems predicted significant wave height distributions result under-spread. A detailed analysis shows that both systems are under-spread where they should mimic low forecast uncertainty (with UKMO slightly more under-spread than Nettuno), and generally over-spread where the uncertainty is higher. In the high uncertainty region UKMO is more over-spread than Nettuno, and the behaviour of the latter is sometimes in line with the deterministic error distribution.

Both the average UKMO reliability, and the one evaluated for a number of discrete cases, is generally lower (better) than Nettuno one. The differences mainly depend on the observations sub-sample. On the other hand, Nettuno resolution is superior to UKMO one. In particular, the latter lacks in ability to discriminate occurrence/non-occurrence of high energetic events, especially at buoy locations (mainly distributed close to coastal regions). Nettuno eps forecast generally fits better than UKMO to decision makers who must make a protection/non-protection choice with respect to high energetic sea states, and whose costs of protection are lower than potential losses. While UKMO is in some cases more suitable for high cost/loss ratio and low energetic, early occurring events. The ensemble size influence on the differences in REV is negligible.

VII REFERENCES

Anderson, J.L., 1996: *A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations*. Journal of Climate, **9**, 1518–1530.

Atger, F., 2004: *Estimation of the reliability of ensemble-based probabilistic forecasts*. Q.J.R. Meteorol. Soc., 130: 627–646. doi: 10.1256/qj.03.23

Buizza, R. and T.N. Palmer, 1998: *Impact of ensemble size on ensemble prediction*. Mon. Weather Rev., 126, 2503–2518

Bidlot, J.R. and M. Holt, 2006: *Verification of Operational Global and Regional Wave Forecasting Systems Against Measurements for Moored Buoys*. JCOMM Technical Report No. 30. ftp://ftp.wmo.int/Documents/PublicWeb/amp/mmop/documents/JCOMM-TR/J-TR-30/J-TR-30.pdf

Candille, G., C. Côté, P. L. Houtekamer, G. Pellerin, 2007: *Verification of an Ensemble Prediction System against Observations*. Mon. Wea. Rev., **135**, 2688–2699.

Candille, G., and O. Talagrand, 2004: *Impact of observational errors on the validation of ensemble prediction systems*. Ensembles Workshop, Exeter, United Kingdom.

Hamill, T., and S. J. Colucci, 1997: *Verification of Eta–RSM Short-Range Ensemble Forecasts*. Mon. Wea. Rev., **125**, 1312–1327.

Hersbach, Hans, 2000: *Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems*. Wea. Forecasting, **15**, 559–570.

Janssen, P.A.E.M., S. Abdalla, H. Hersbach and J.R. Bidlot, 2007: *Error Estimation of Buoy, Satellite, and Model Wave Height Data*. J. Atmos. Oc. Tech., **24**, 1665-1677. doi:10.1175/JTECH2069.1

Richardson, D.S., 2000: *Skill and Relative Economic Value of the ECMWF Ensemble Prediction System*. Quart. J. Royal Met. Soc., **126**, 649-667.

Saetra, Ø. and J.R. Bidlot, 2004: *Potential Benefits of Using Probabilistic Forecasts for Waves and Marine Winds Based on the ECMWF Ensemble Prediction System.* Weather and Forecasting, **19**, 673-689.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, D.S. Richardson, 2004: *Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability*. Mon. Wea. Rev., **132**, 1487–1501.

Scherrer, S.C., C. Appenzeller, P. Eckert and D. Cattani, 2004: *Analysis of the Spread–Skill Relations Using the ECMWF Ensemble Prediction System over Europe*. Weather and Forecasting, **19**, 552-565.

Talagrand, O., R. Vautard, and B. Strauss, 1997: *Evaluation of Probabilistic Prediction*

Systems. Proc. ECMWF Workshop on Predictability. Reading, United Kingdom, ECMWF, 1–25.

Taylor, K. E., 2001: *Summarizing Multiple Aspects of Model Performance in a Single Diagram.* J. Geophys. Res., **106**(D7), 7183–7192, doi:10.1029/2000JD900719.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences: an Introduction*. Academic Press, 467 pp.

VIII ANNEX A

Measures Definitions

We recall here a set of measures, together with their definitions, used in the document (see also DR 3 and DR 4). Being **f** a subsample of forecast realizations, co-located with the subsample of observations **o**, let define:

$$BIAS = \left\langle \mathbf{f} - \mathbf{o} \right\rangle \tag{13}$$

$$MAE = \left\langle \left| \mathbf{f} - \mathbf{o} \right| \right\rangle \tag{14}$$

$$MSE = \left\langle \left(\mathbf{f} - \mathbf{o} \right)^2 \right\rangle \tag{15}$$

$$RMSE = \sqrt{MSE} \tag{16}$$

$$SI = \frac{RMSE}{\sigma(\mathbf{o})} \tag{17}$$

$$RVAR = \frac{\sigma^2(\mathbf{f})}{\sigma^2(\mathbf{o})} \tag{18}$$

$$SIVAR = \frac{MSE}{\sigma^2(\mathbf{o})} \tag{19}$$