COPERNICUS
MARINE ENVIRONMENT MONITORING SERVICE

European Union Programme

EGU session OS4.7
High-resolution model Verification Evaluation (HiVE). Part 1: Using neighbourhood techniques for the assessment of ocean model forecast skill

Jan Maksymczuk, Ric Crocker,
Marion Mittermaier and Christine Pequignet

Objectives of the HiVE project

The **High-resolution model Verification Evaluation (HiVE) project** considered CMEMS forecast products applicable to regional domains and aimed to demonstrate, for the first time, the utility of spatial verification methods (originally developed to evaluate km-scale forecasts from atmospheric models), for verifying km-scale ocean model forecasts. It was undertaken to address the need for new metrics adapted to the increased resolution in both observations and models, as identified in the CMEMS Service Evolution Strategy.

The project had two key objectives relating to the ongoing assessment protocols for ocean forecast models, and how they could be evolved to cope with future modelling systems.

1. *To understand the accuracy of CMEMS products at specific observing locations using neighbourhood methods and ensemble techniques*

2. *To understand the skill of CMEMS products in forecasting events or features of interest in space and time - see* **High-resolution model Verification Evaluation (HiVE). Part 2: Using object-based methods for the evaluation of chlorophyll blooms**

Context

Understand the accuracy of CMEMS products at specific observing locations using neighbourhood methods and ensemble techniques

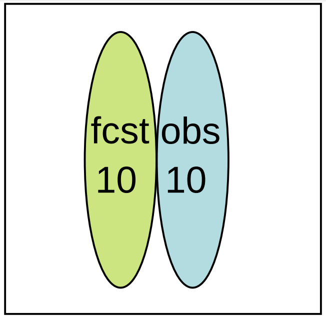
- Several CMEMS products covering the same locations, from both global and regional TACs, are available to users but are not necessarily assessed in a similar way.
- Trial a spatial verification methodology (HiRA - High Resolution Assessment framework) which uses ensemble and probabilistic scores to equitably compare these models with regards to their accuracy and predictive skill.
- Help inform CMEMS model developers and users on the basic accuracy and skill of ocean forecasts at a given location when comparing lower resolution models with their next generation evolutions (for example, the European NWS AMM7 and AMM15).

Why spatial verification methods?

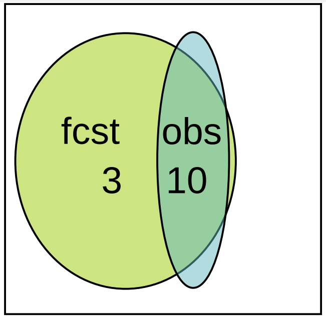
Traditional verification requires an exact match between forecasts and observations at every grid point or observing location.

- This means that in a categorical sense (threshold) the forecast and observed region of interest must overlap.
- In high-res forecasts the detail looks realistic but may not be in the right place at the right time.
- This leads to the so-called "**double penalty**" - event predicted where it did not occur, event not predicted where it did occur
- Traditional scores do not say very much about the source or nature of the errors

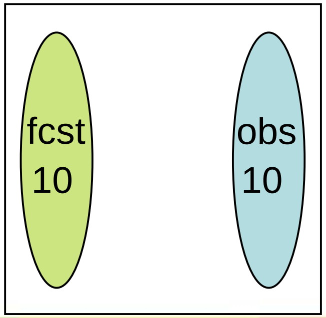
(Probability of Detection (POD); False Alarm Rate (FAR); Threat Score (TS))



Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0



Low res forecast
RMS ~ 2.7
POD~1, FAR~0.7
TS~0.3



In this example the low-res fcst has the smallest error and is the only fcst with non-zero POD

High Resolution Assessment

The **High Resolution Assessment (HiRA) framework** (Mittermaier 2014, Mittermaier and Csima 2017) was designed to overcome difficulties encountered using traditional verification in proving model skill when assessing high-resolution models, or when comparing deterministic and probabilistic models.

The assessment does not rely on direct point-to-point comparison, but rather uses increasing sized neighbourhoods to the forecast field to generate a pseudo ensembles or a probabilistic value (based on the spatial fractional exceedance of a specified threshold) which can then be compared to an observed value. Different statistics, parameters, thresholds and neighbourhood sizes can be selected in order to assess two (or more) models.

The technique used is one example of a *single observation - neighbourhood forecast* verification approach (SO-NF). This assumes that the observation is a true value and not only representative of the exact location at which it observes, but also has the characteristics of a surrounding area as well. This serves to mitigate the 'double-penalty' effect which can occur when comparing at the grid scale.

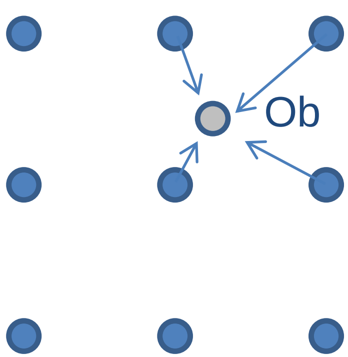
The user can look at different sized areas to see how the skill of models varies within the area around the model. If the area becomes too big the representativity of the point observation may be lost, depending on the variable. This imposes a maximum neighbourhood size on the evaluation framework, since any skill at a given location could be deemed entirely random.

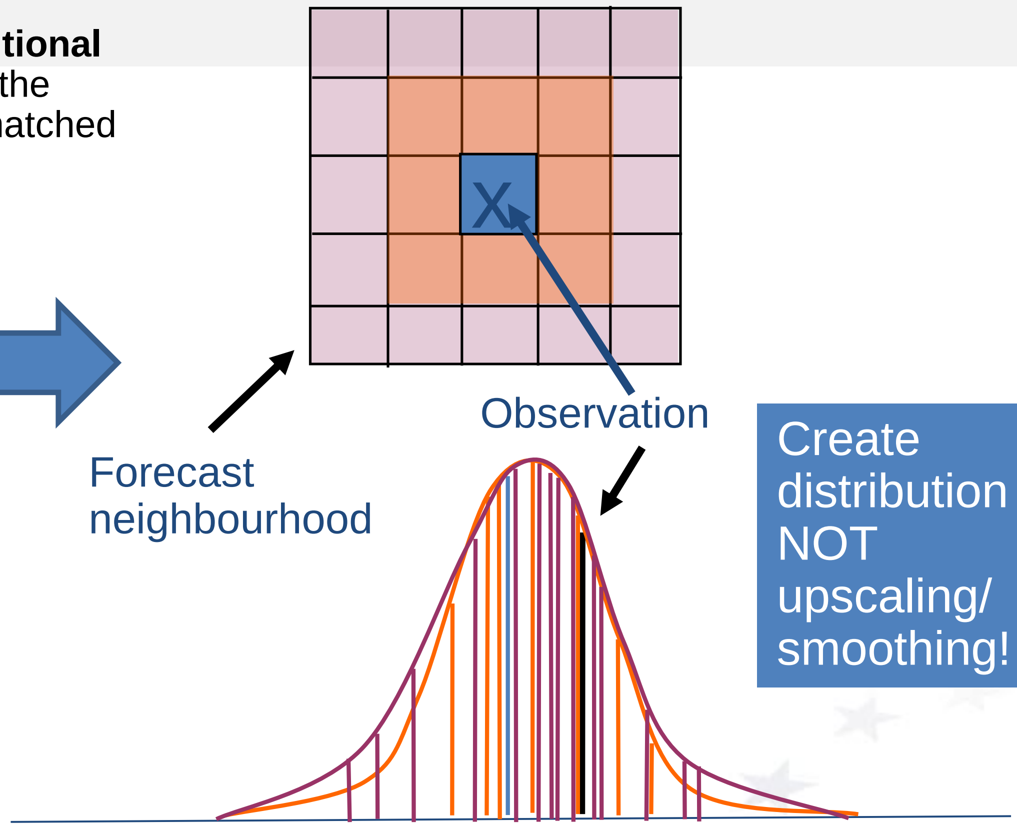
Spatial sampling

Make use of spatial verification methods which compare **single observations** to a **forecast neighbourhood around the observation location** → ***probabilistic framework***

Represents a **fundamental departure from traditional verification** where the emphasis is on extracting the nearest grid point or bilinear interpolation to get matched forecast-ob pair.

Traditional interpolate/nearest



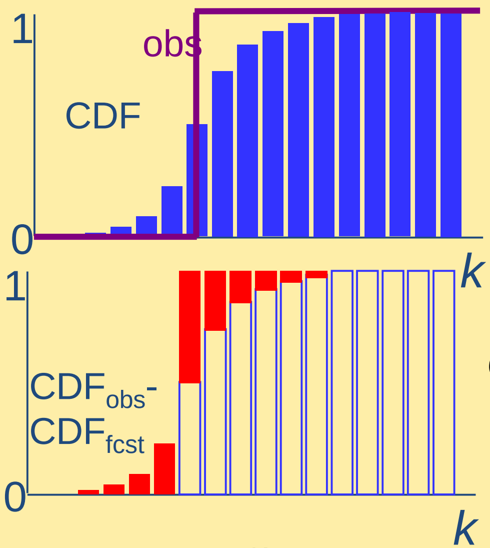


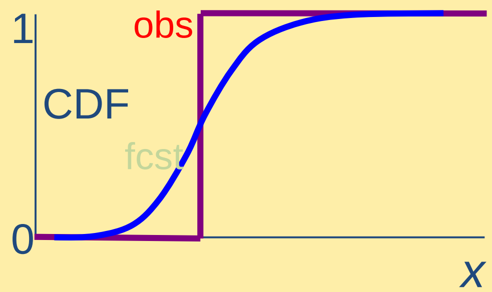
Create distribution NOT upscaling/ smoothing!

Model-specific v user-relevant

Probabilistic Forecast / Binary Observation

Modeller interested overall performance or pdf





$$CRPS = \int_{-\infty}^{\infty} \left(P_{fcst}(x) - P_{obs}(x)\right)^2 dx$$

Perfect score = 0

$$RPS = \frac{1}{K-1} \sum_{k=1}^{K} (CDF_{fcst,k} - CDF_{obs,k})^2$$

Perfect score = 0   K = discrete threshold categories

Users interested in specific thresholds: e.g. defining a decision point or hazard

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

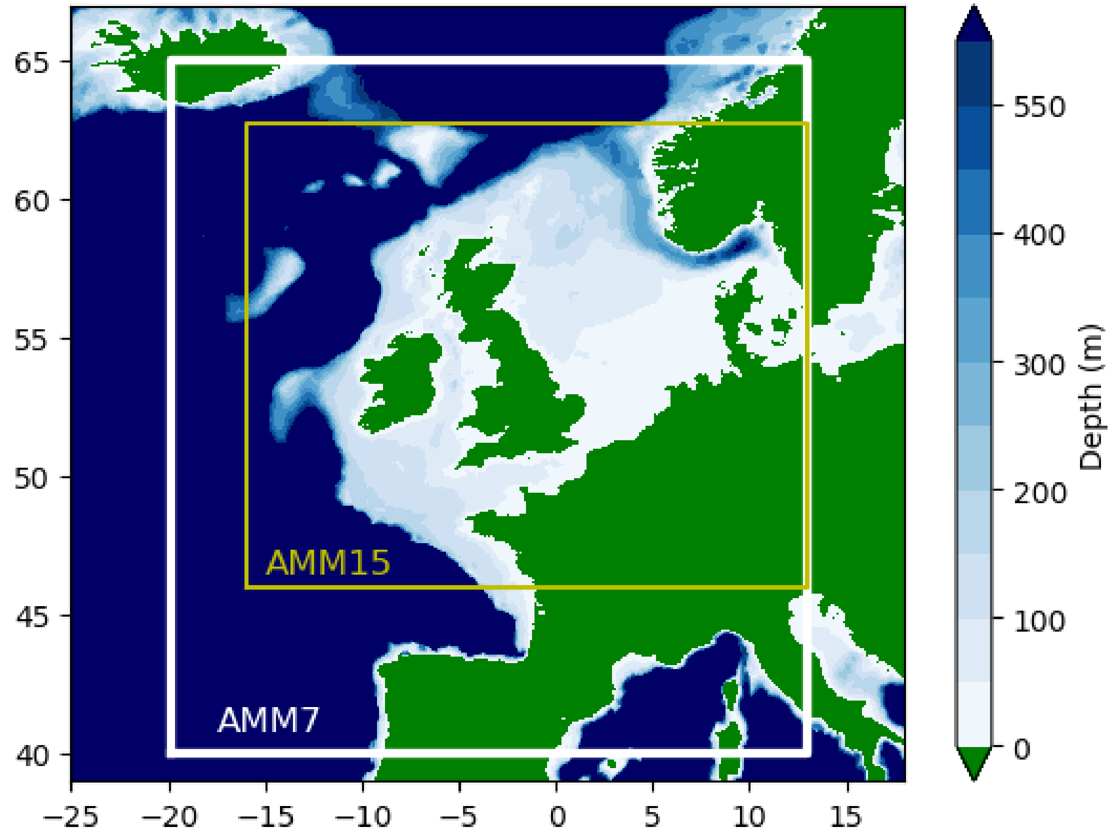
p_i = forecast probability
o_i = observed occurrence (0 or 1)

Perfect score = 0

Model Domains

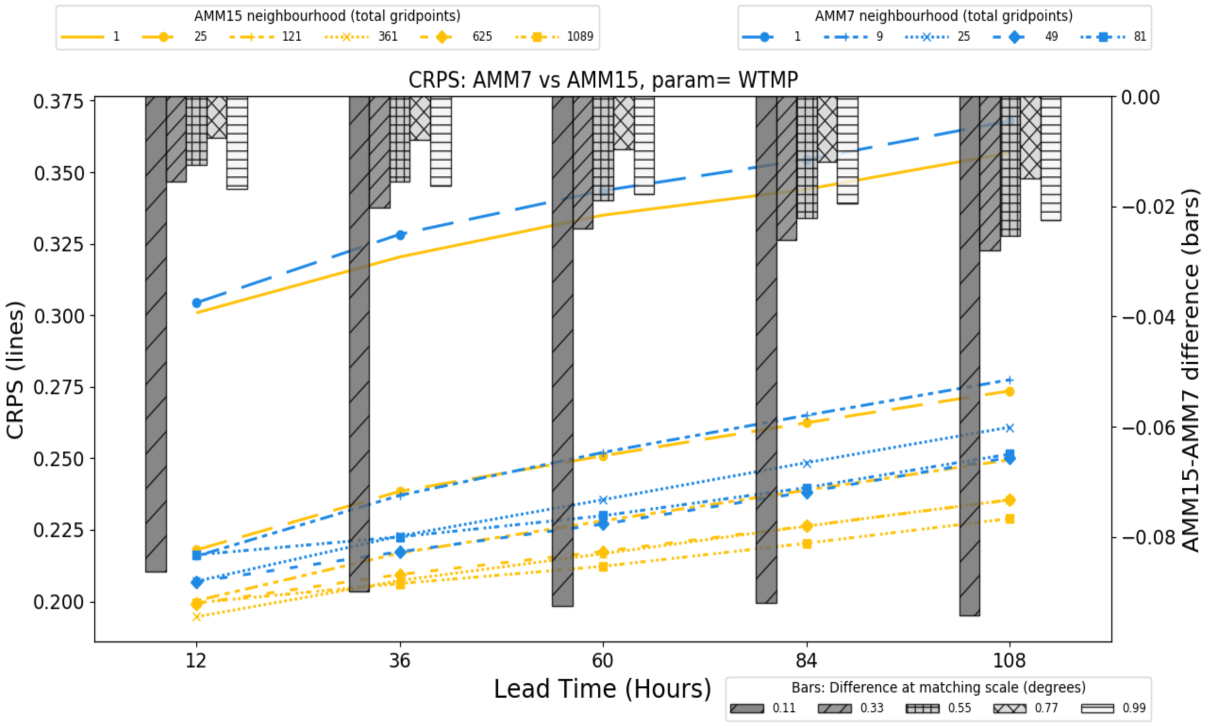
Assessments were made using forecast products from the CMEMS catalogue

- NWS - AMM7 (1/10°)
- NWS - AMM15 (1.5km)



High Resolution Assessment

AMM15 vs AMM7



Results for January - December 2019.

For the CRPS the smaller the score the better.

Daily mean SST

Continuous Ranked Probability Score (CRPS) (lines) as a function of lead time (x-axis) for a range of neighbourhood sizes in grid-squares: 1 x 1(1), 3 x 3 (9), 5 x 5 (25), etc

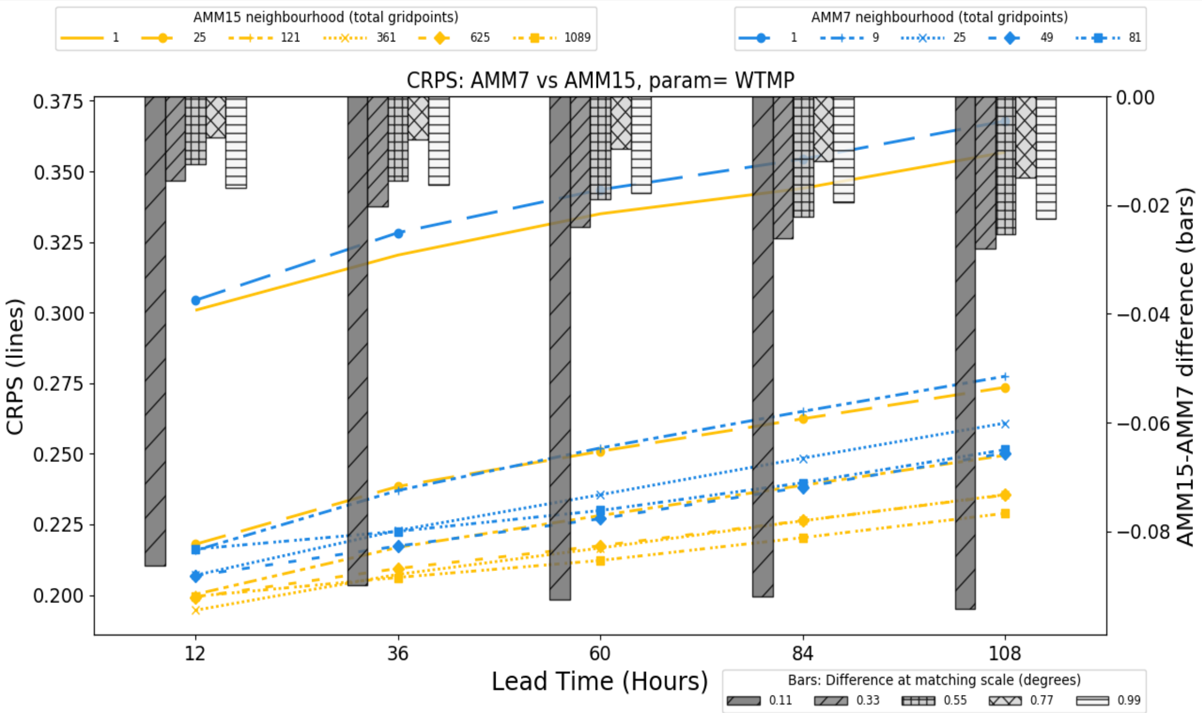
This equates to 9, 25, 49, 81, …pseudo “ensemble” members.

Note that you can equalize on “ensemble” members or on area. When comparing models at very different resolutions it is better to compare scores for equivalent areas.

Here, the matching line styles represent the equivalent neighbourhood sizes that should be compared

High Resolution Assessment

AMM15 vs AMM7



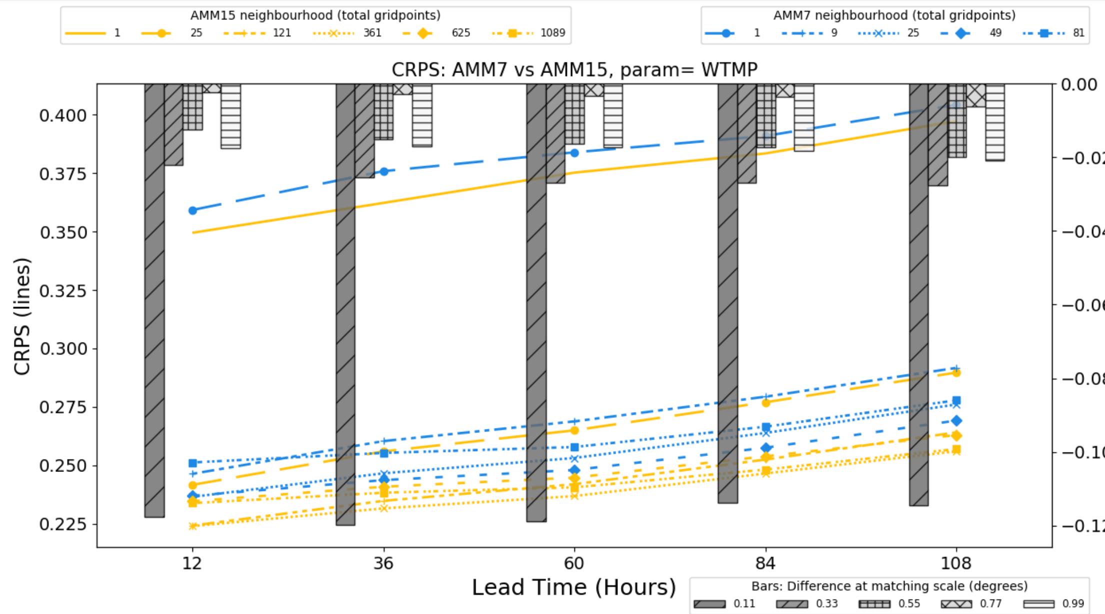
For the CRPS the smaller the score the better.

- CRPS decreases with increasing neighbourhood size, suggesting that some spatial mismatches exist.
- Overall, the higher resolution AMM15 consistently has lower errors when equivalent neighbourhood extents were compared.
- The bars are included here to illustrate the differences in the CRPS between the two configurations, where a positive bar implies AMM7 has better (lower) CRPS, and a negative bar implies AMM15 is better.

High Resolution Assessment

AMM15 vs AMM7 – On-shelf

- Sub-setting the data demonstrates that different signals can be obtained by isolating different components of the domain.
- On-shelf the application of neighbourhoods shows a large decrease in error when moving from grid scale to neighbourhood assessment as small-scale spatial variability becomes encompassed in the scores.
- AMM15 has consistently smaller errors.
- Both models exhibit a worsening of CRPS as the forecast lead time increases

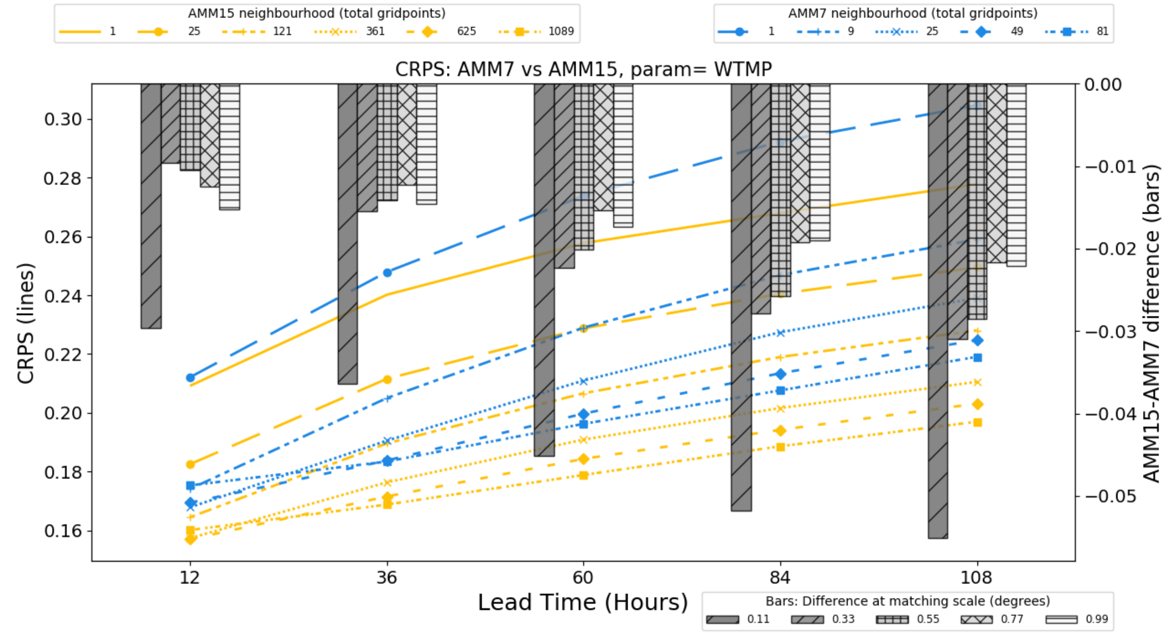


On-shelf

High Resolution Assessment

AMM15 vs AMM7 – Off-shelf

- CRPS scores are better (lower) for neighbourhoods in the off-shelf sub-domain, where temperatures are spatially smoother.
- AMM15 has consistently smaller errors.
- Both models exhibit a worsening of CRPS as the forecast lead time increases, which is more rapid for the off-shelf assessment.



Off-shelf

Conclusions

High Resolution Assessment:

- It has been demonstrated statistically that higher-resolution models can be assessed equitably against lower resolution configurations
- Spatial methods counteract the impact of the double-penalty effect, which acts to destroy the positive signal obtained from the qualitative benefits of km-scale models
- Assessments have also been undertaken for the overlapping domains of **AMM7 vs IBI** and **AMM15 vs IBI**, giving similar results
- Applications of HiRA could be limited in some cases due to sparsity of observation coverage, as the number of available observations decreases with increasing neighbourhood size (this is mostly an issue for sites closer to the coast)
- Future comparative assessments of ocean models with different resolutions would benefit from using HiRA as part of the evaluation process, as it gives a more equitable and appropriate reflection of model performance at higher resolutions

Questions and discussion

Further detail at:

- Ocean Science paper, submitted in February 2020 -

Crocker R., Maksymczuk J., Mittermaier M., Tonani M., Pequignet C.
An approach to the verification of high-resolution ocean models using spatial methods

This work has been carried out as part of the Copernicus Marine Environment Monitoring Service (CMEMS) HiVE project. CMEMS is implemented by Mercator Ocean International in the framework of a delegation agreement with the European Union.

Verification was performed using the Point-Stat tool, which is part of the Model Evaluation Tools (MET) verification package, that was developed by the National Center for Atmospheric Research (NCAR), and which can be configured to generate CRPS results using the HiRA framework. MET is free to download from GitHub at https://github.com/NCAR/MET.