Evaluation of the coupled system

Evaluation of the marine heatwave representation by the coupled hindcast

An important point that this article makes is that the marine heatwave started near the surface, became very intense in September, and then gradually penetrated through depth in the second half of September and first half of October. In this section, we compare the model with glider profiles, which are the main source of depth observations in this region. Unfortunately, there was no glider operating in the region of interest for this study, so we evaluated the model in the Northern North sea against glider profiles from the MOGLI deployments East of Orkney. The glider data are separated into upward and downward casts and only the downward ones are used here. Hourly instantaneous temperatures and salinities of the hindcast simulation are matched up, in time, location and depth, with the observations using the COAsT python package (Byrne et al, 2023). The glider track goes back and forth along a zonal section between depths of 100 and 150 m across the isobaths (Fig. S1). The depth profiles (Fig. S2) until 15 September shows nearly periodic variability associated with the alternating direction of the zonal transect during a stable period. From 26 September and the end of the deployment, the track changes from a meridional westward section to a zonal Northward section.

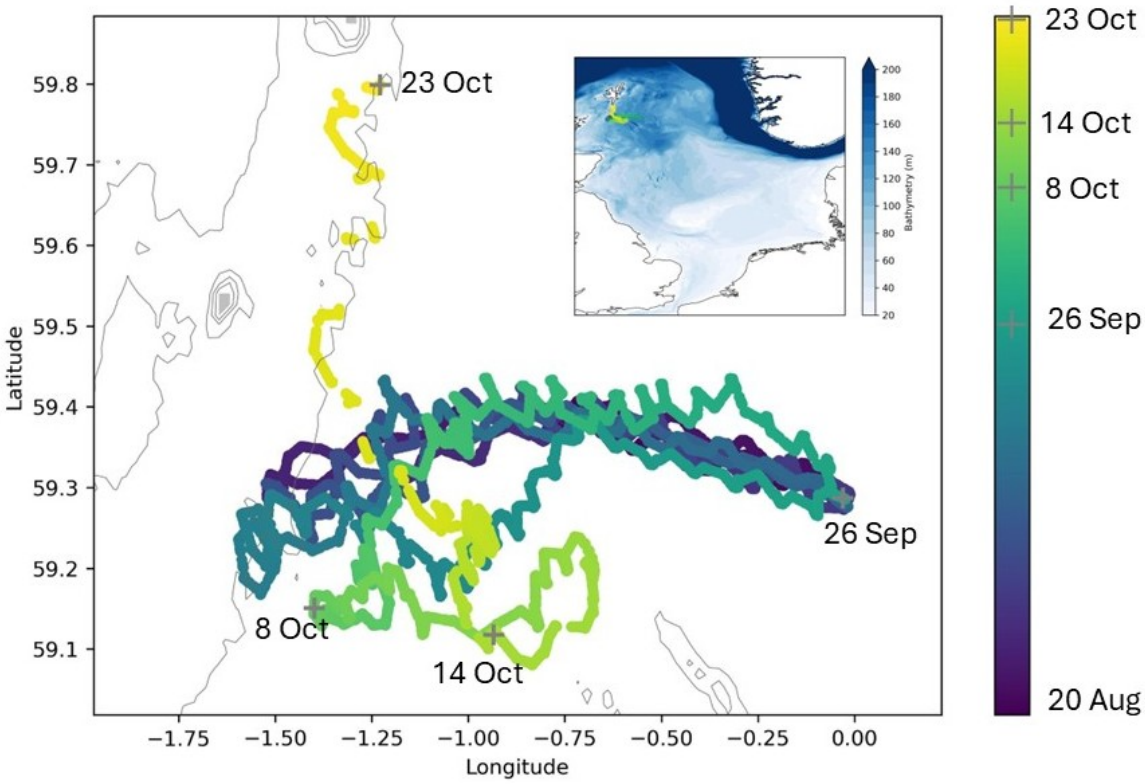

Figure S1: Glider track over the coupled hindcast run of the marine heatwave

The model shows good behaviour in terms of spatio-temporal variability along the glider track. The main bias is a deeper mixed layer by about 10 m, but the temperature in the mixed layer shows a bias lower than 1 K for most of the time: this means the coupled model is able to represent the vertical structure of the marine heatwave, and provides trust in its use to understand the origin of marine heatwaves. In particular, the episodes of mixed layer deepening between October 5 and 10 are well represented, and the observations confirm that the ocean is weakly stratified on October 15, before the start of storm Babet. The model is showing a similar lack of stratification, as previous mixing events have reduced the stratification. This supports the model findings that the cooling during storm Babet happened mostly because of intense latent and sensible heat fluxes.

Evaluation of the coupled ensemble forecast

Storm Babet was an intense event for wind, waves, precipitation and river discharge. We used the coupled system to assess the impact that the marine heatwave had on these variables. In this section, we evaluate the coupled model. We make use of in-situ buoy data, available from WaveNet and MetDB databases, we used the data points of best quality and closest to the backward trajectories shown in **Figure 4** and **Figure 5**. Figure S5 shows that the model has a good time evolution of the winds for the first four days of the runs, with winds

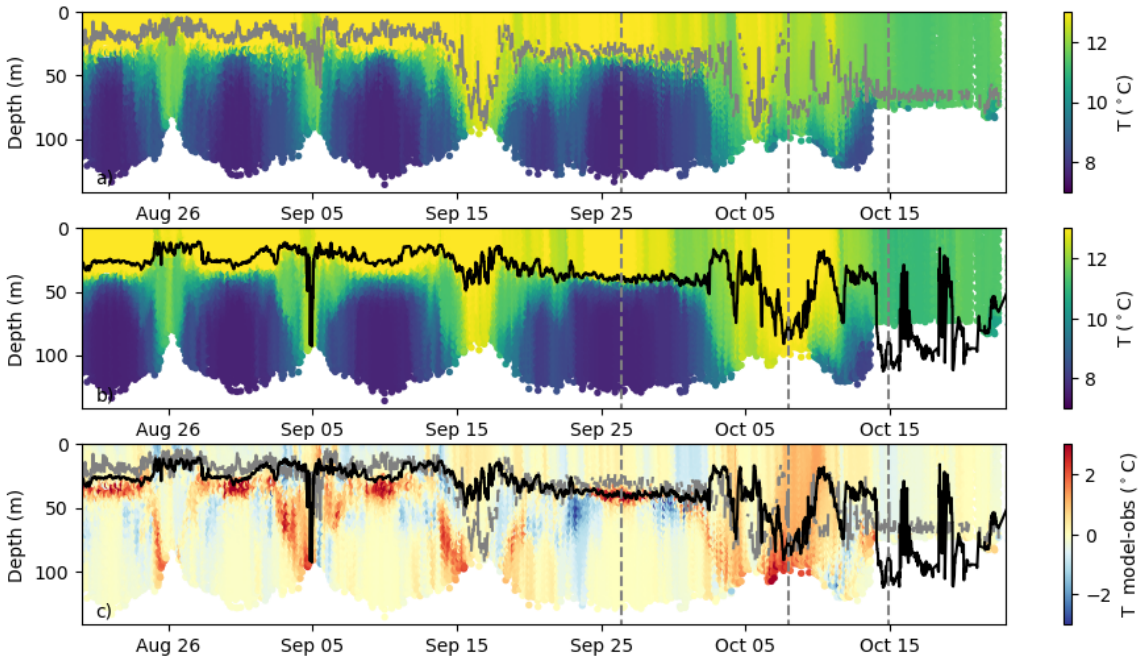

Figure S2: Temperature profiles sampled by the glider (top), coupled hindcast (middle) and difference between the two (bottom).

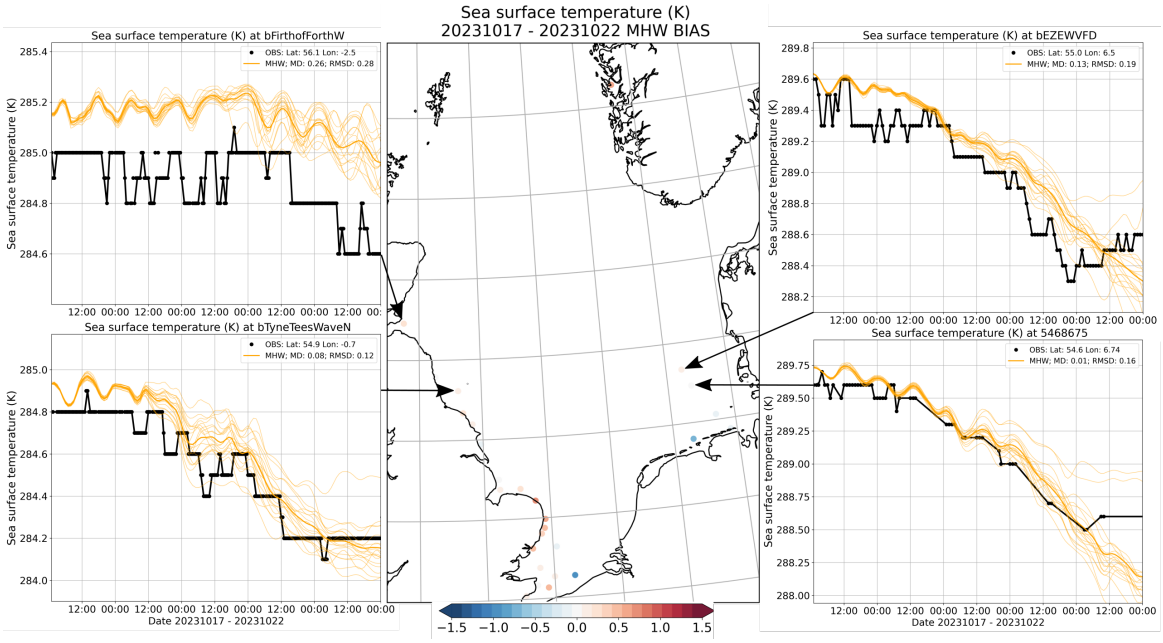

Figure S3: Map: average ocean surface temperature bias of UKC4 ensemble mean over the 5 days of the forecast. Point location time series: black: observations, yellow: thin lines show individual model members, thick line shows ensemble mean.

increasing on June 18. However, the ensemble average underestimates winds by $5\,\mathrm{m\,s^{-1}}$ in the southeastern part of the North Sea, and the strongest members by 2-$4\,\mathrm{m\,s^{-1}}$. The model's winds remain strong on June 21, whereas the observations show a decrease: the ensemble becomes unreliable for day 5 of the forecast. In the central part of the North Sea, the underestimation is smaller ($2\,\mathrm{m\,s^{-1}}$ against one anemometer, and no bias against another anemometer 10 km apart, indicating that the observations do have $2\,\mathrm{m\,s^{-1}}$ uncertainty on these oil rigs).

In terms of significant wave height (Fig. S6, the timing of wave increase is well captured by the ensemble on June 17 and 18, but the model fails to capture the very fast increase in significant wave height on June 19, and the maximum is then underestimated: the observations show 54 h of waves greater than 6 m on Scottish coastlines, with peaks at 8 m (and a single value reaching 10.3 m), whereas the ensemble mean shows 54 h

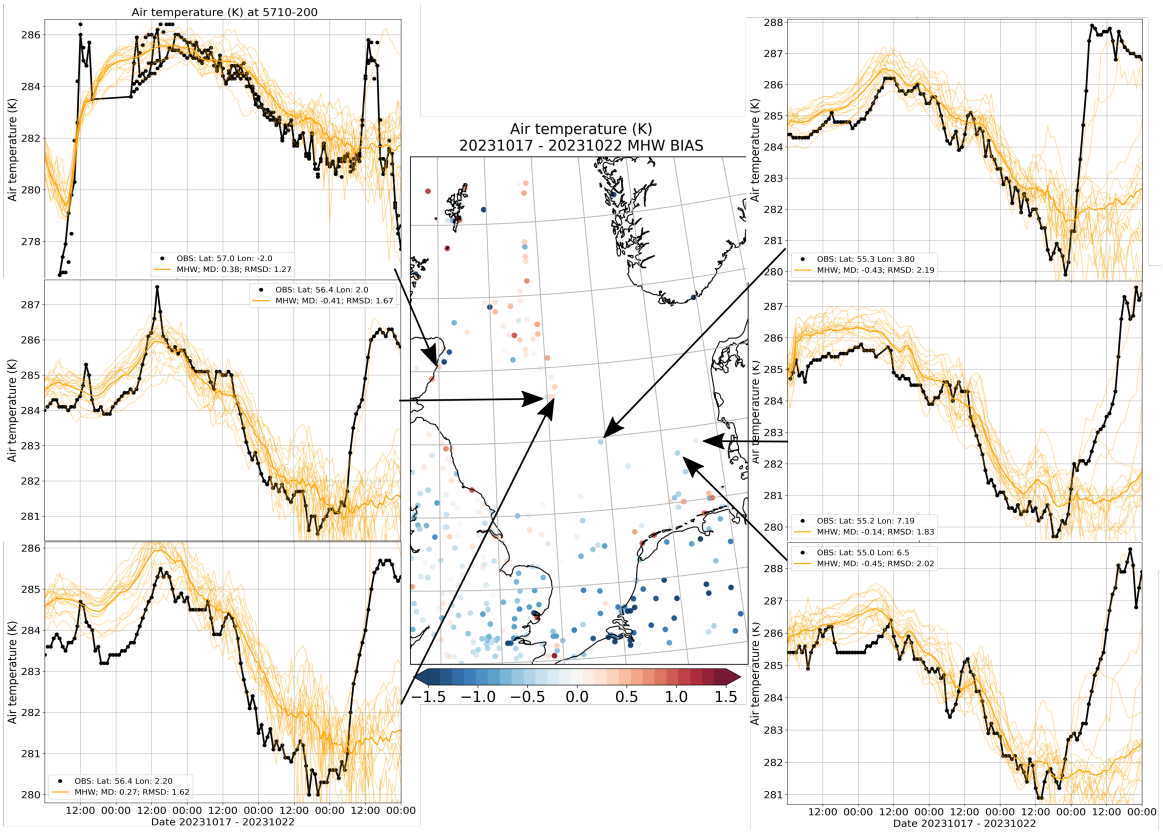

Figure S4: Map: average 1.5 m air temperature bias of UKC4 ensemble mean over the 5 days of the forecast. Point location time series: black: observations, yellow: thin lines show individual model members, thick line shows ensemble mean.

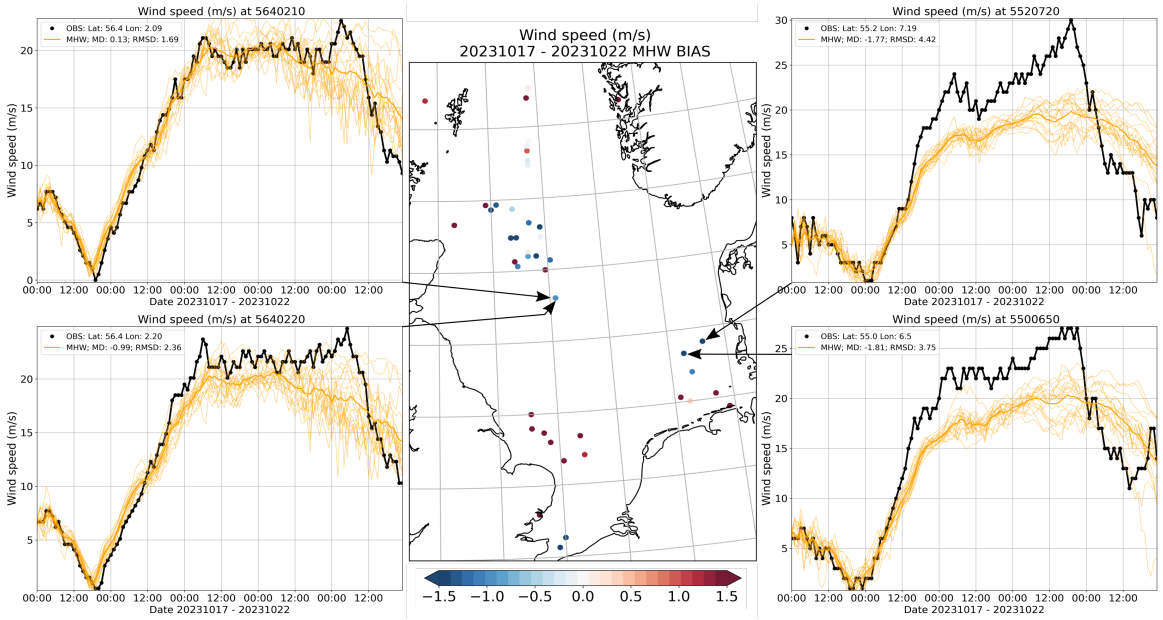

Figure S5: Map: average wind bias of UKC4 ensemble mean over the 5 days of the forecast. Point location time series: black: observations, yellow: thin lines show individual model members, thick line shows ensemble mean.

of waves above 4.5 m, with 6 m peak. Nevertheless, some ensemble members reach wave maxima of 7 m. In the oil rigs, wave maxima of 9 m are reached, with 66 h of significant wave height above 6 m. The ensemble mean reaches 5 m for this duration, with peaks at 7 m, and some ensemble members show peaks at 8 m. The underestimation is largest on June 21, which is consistent with an underestimation of the June 20 18:00 UTC wind peak in southeastern North and June 21 06:00 UTC in Central North Sea seen in Fig. S5. In conclusion, the model captures well the overall wave growth and long-lasting intense wave height, albeit with a 1-1.5 m

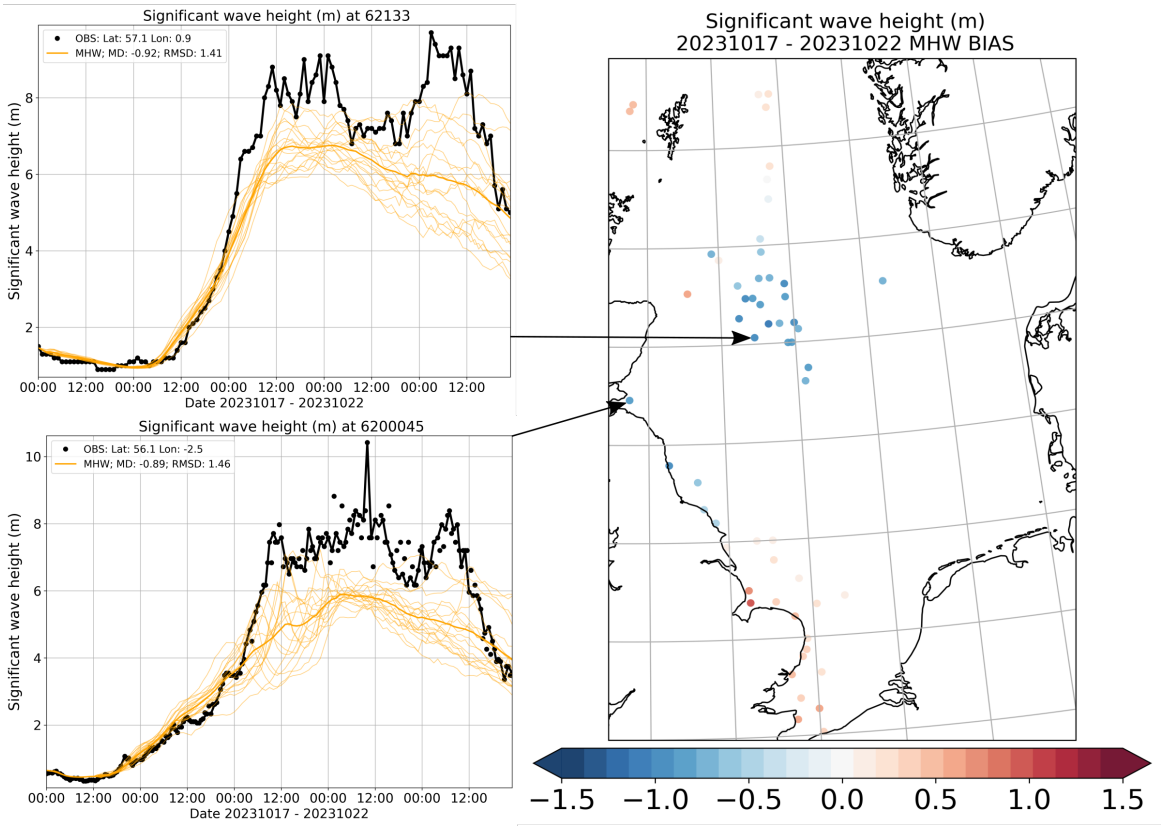

Figure S6: Map: average wave bias of UKC4 ensemble mean over the 5 days of the forecast. Point location time series: black: observations, yellow: thin lines show individual model members, thick line shows ensemble mean.

underestimation. This is a common problem for both uncoupled and coupled wave models: wind extremes and wave extremes are often underestimated, and may need a re-design of boundary layer/free troposphere interactions for high wind speeds.

Evaluating the 1.5 m air and sea surface temperature of the ensemble forecast are also key to understand if the ensemble is capturing the air-sea interactions with reasonable skills. Although satellite-derived SST datasets are generally used for model verification, their quality is reduced when cloud cover is high, which is the case in this weather situation. Therefore, we use in-situ observations, although we note their scarcity (Fig. S3 compared with air temperature observations (Fig. S6). Again, we show observation points closest to the air-parcel trajectories. The ensemble shows very good skills in terms of sea surface temperature decrease, capturing the intense 1 K cooling in 48 h in the southeastern part of the North Sea and 0.6 K offshore the Northern England coastline, and 0.2 K cooling on Scottish coastline (Firth of Forth). Regarding the air temperature, the variability is much larger, and all the observations in the North Sea support the evidence from the backward trajectories that the air flow origin changes during the storm, with a 4 K drop in 32 h for offshore sites situated closest to the trajectories. We note that the air temperature is always cooler than the sea temperature on the southeastern part of the North Sea, whereas it only becomes cooler than the sea temperature on June 20, explaining the weaker cooling closer to Scotland, despite similar wind strength. The model is capturing this overall evolution well, but is warmer by 0.5 to 1.5 K (except on the last day, where it fails to capture the intense warming, apart from a few members). The wind underestimation and temperature overestimation suggest the model fluxes may be underestimated compared to the observations, but the SST cooling by the coupled model is correct, suggesting that latent heat fluxes may be compensating sensible heat flux errors. However, observations of relative or specific humidity are not available to evaluate the model on this region.