Supplementary Material. Authors provide the following supplementary material with the publication. The plots provided below are complimentary to the plots presented in the main body of the publication.

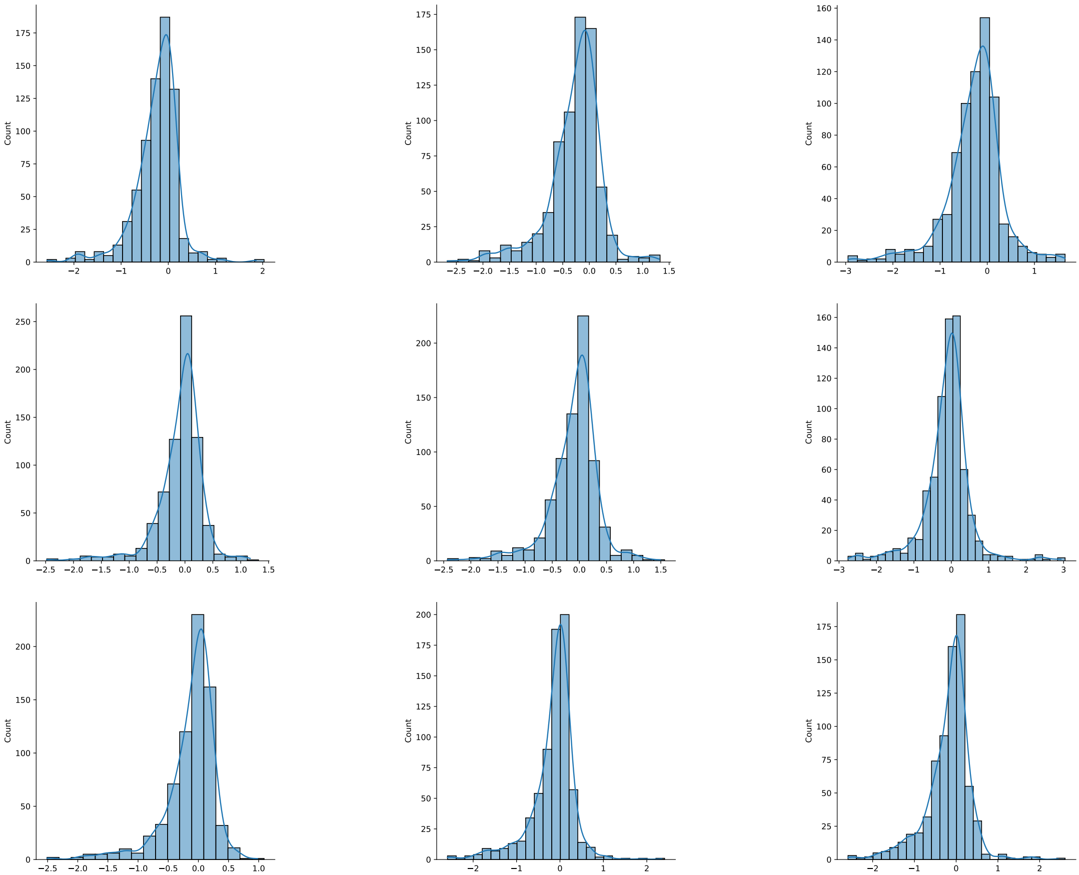

Figure 23. *Histograms of the difference between the original forecast for Wanderer I mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

S1 Forecasts versus buoy observations

In the main body of the article, we investigated the normality assumption by comparing the observed significant wave height to the bias-adjusted forecasts using histograms and QQ-plots, and presenting the results for the adjusted M6 winter period. Below we present all other missions - Explorer I, Wanderer I, Wanderer II, M6 summer, and M6 winter unadjusted.

In the main body of the article, authors present the comparison between the H_s recorded by the Wanderer buoy and the three model forecasts for the period between 15 June 2020 and 15 September 2022. In the same section 3, Results, authors make a reference to results to the second Wanderer and first Explorer missions. And the MAE for these two additional missions is presented in table 1. Actual 'brute force' comparison with the Wanderer and Explorer data and model forecasts is plotted in figures 33 and 34.

During Explorer's first mission, it went adrift on 5 September 2020. In the results presented, authors only concentrate on the period where the buoy stayed anchored in one geographical spot. But it is still beneficial to look at the drifting period.

While adrift, Explorer was approaching the shoreline. Closer to the shore the water depth is decreasing, and we see that the values of the H_s are lower than the predicted values. In general, it is expected to see larger waves closer to shore. This is well explained by the wave behaviour when waves transition from deep water to shallow water. This can be understood by looking at wave speeds. Once in deep water, the wave speed depends on the wavelength. However, once the waves enter shallow water, the wave behaviour is affected by the bathymetry. As waves feel the bottom, they slow down, and following waves 'catch up' with waves in front, so the wavelength is decreased. In this process, the energy contained in the wave is constant, hence the wave grows, and one is supposed to see higher waves. However, waves cannot grow infinitely: once a certain threshold is reached

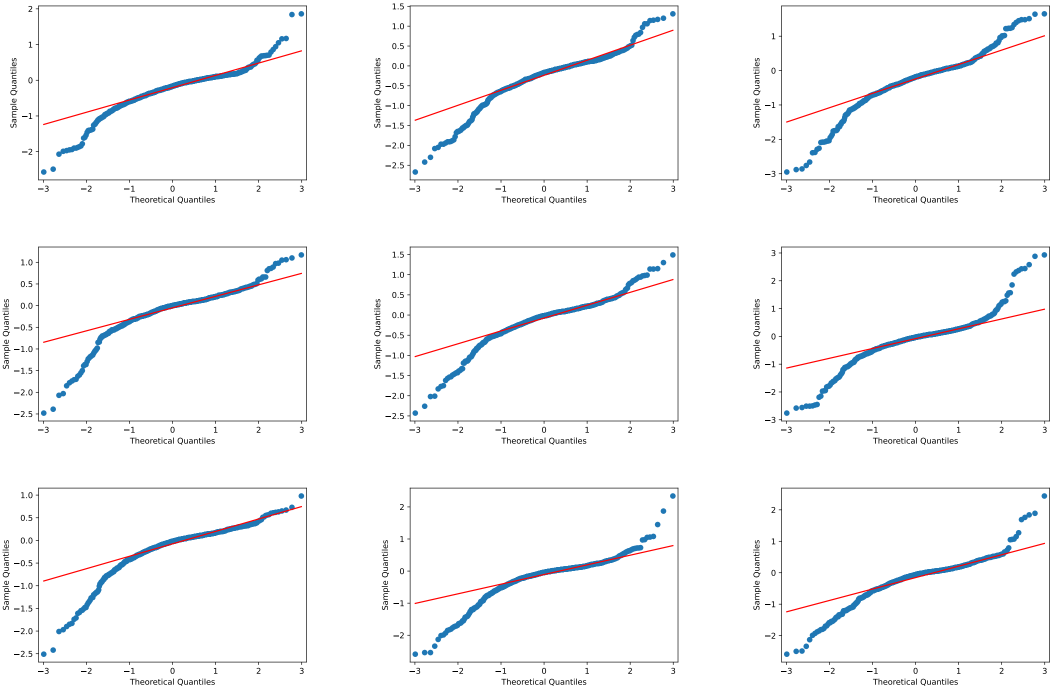

Figure 24. *Q-Q plots of the difference between the original forecast for Wanderer I mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

waves break (see for example **Tian Z., Perlin M. G., Choi W.** [2010], **Tian Z., Perlin M. G., Choi W.** [2012], **Ducrozet G., Gouin M.** [2017], **Derakhti M., Banner M. L., Kirby J. T.** [2018], **Barthelemy X., Banner M. L., Peirson W. L., Fedele F., Allis M., Dias F.** [2018], and the book by **Babanin A.** [2011]). Here we leave the process of breaking aside, but it is important to understand that shoaling waves grow and can eventually break. This is exactly what we see in figure 33, after the vertical black line that denotes drifting. The observed value of H_s is lower than the one predicted for the area with larger depth, but the shape of the evolution of the observed H_s compares very well to the shape of the forecast.

S2 Training window selection for the Wanderer II and Explorer I missions

In section 3. Results, authors discuss the selection of the sliding training window for the ensemble forecasts. The Wanderer I mission is used as an example, and the plots are presented in the main body of the article. Here the authors present similar plots, but for the Explorer I and Wanderer II missions. As previously discussed, longer training period is not always necessary. In the Wanderer II case, we see a plateau after certain number of training days is reached. Both MAE and CRPS for Wanderer II mission are presented in figure 35 top panels.

We see a different picture for Explorer, see figure 35 (bottom panels) - we see how both the MAE and CRPS are increasing with the number of training days, hence the shortest training period was selected for Explorer. To understand this error increasing behaviour in Explorer's training, we looked separately at training 24, 48, and 72 hour forecasts, the information regarding these separate training periods is presented in figure 36.

The 24H forecast displays the expected behaviour: decreasing values of MAE and CRPS with an increasing number of training days, with a plateau at some point. However, 48H and 72H forecasts appear to increase in MAE and CRPS values. The authors do not have a concrete explanation for this behaviour of 48H and 72H forecasts. However, one of the possible explanations is that as we know that 48H and 72H carry more inaccuracies than 24H ones, and with increasing training window, we keep accumulating the errors, thus increasing the values for MAE and CRPS. Accordingly, selecting a range of optimal training days adds to the training process, as it allows to assess the robustness of selecting a short over a long (or medium) training window.

Taking all of the above into the consideration, a 5 day training window was selected for Explorer, and a 20 day window for second Wanderer mission.

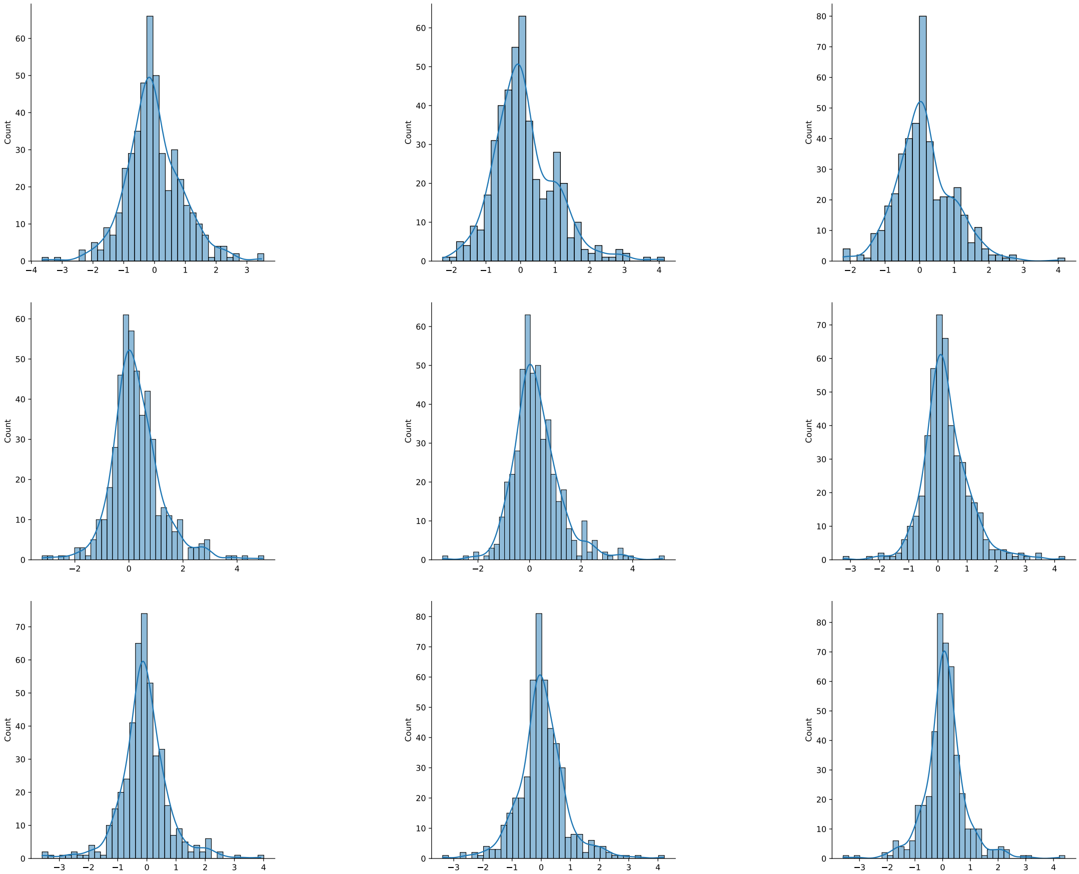

Figure 25. *Histograms of the difference between the original forecast for Wanderer II mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

S3 Number of effective forecasts for Wanderer II and Explorer I missions

In figure 38 authors present the number of effective forecasts for Explorer I and Wanderer II missions, similar to section 3.2. For the two missions same trend is visible - there are instances where only one forecast is an effective forecast in the overall ensemble. However, in the case of Explorer I mission two forecasts are leading number (50%) and three forecasts are effective in 25% of cases. During Wanderer II mission we see that it s between four and five forecast models are used mostly in the ensemble (40.7% and 28.8% of cases respectively). This figure 38 and the figure in the main body of the text support the authors hypothesis that the ensemble forecast made from more than one model is more effective in predicting short-term significant wave height than just one particular model.

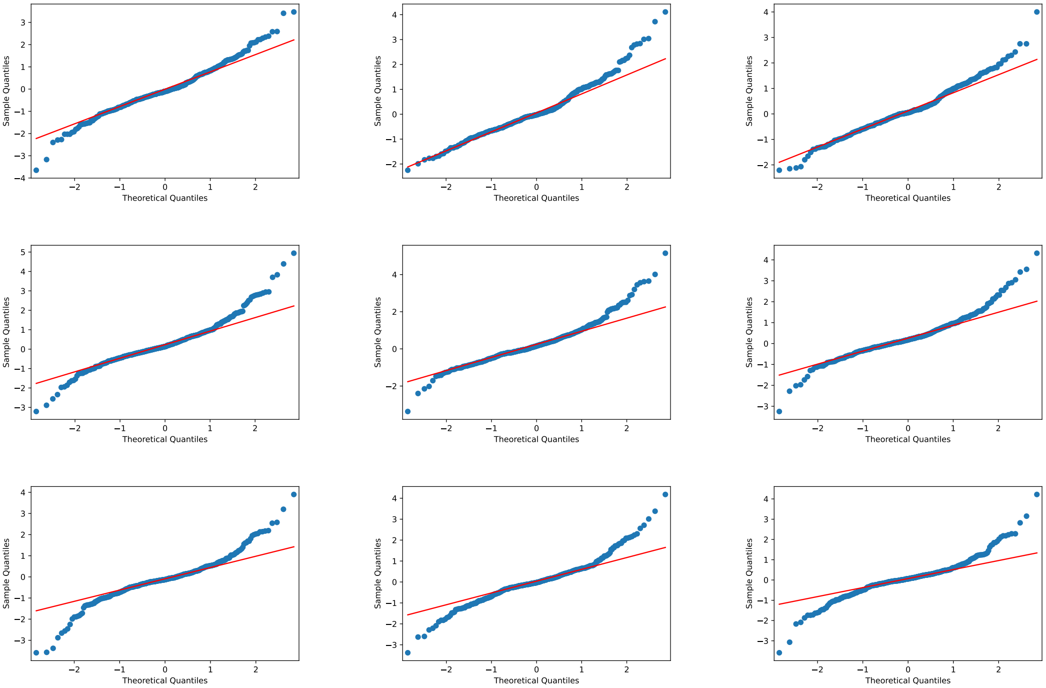

Figure 26. *Q-Q plots of the difference between the original forecast for Wanderer II mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

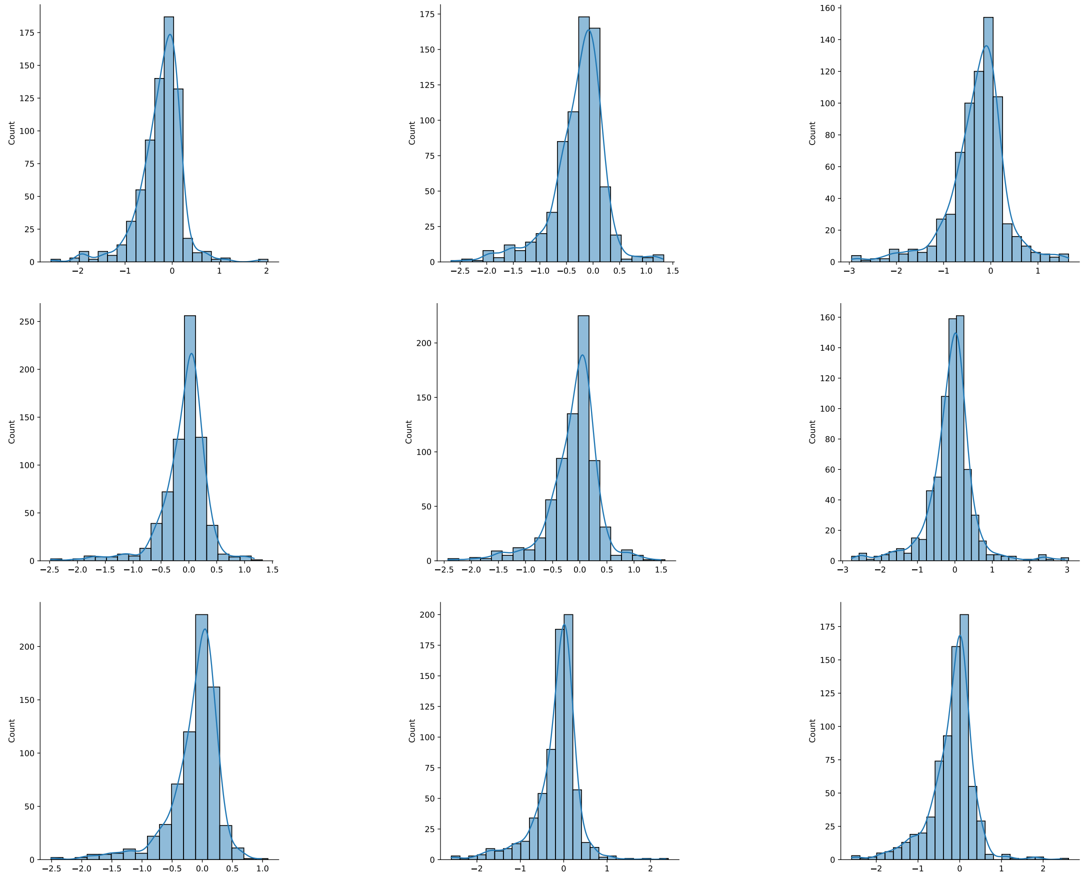

Figure 27. *Histograms of the difference between the original forecast for Explorer I mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

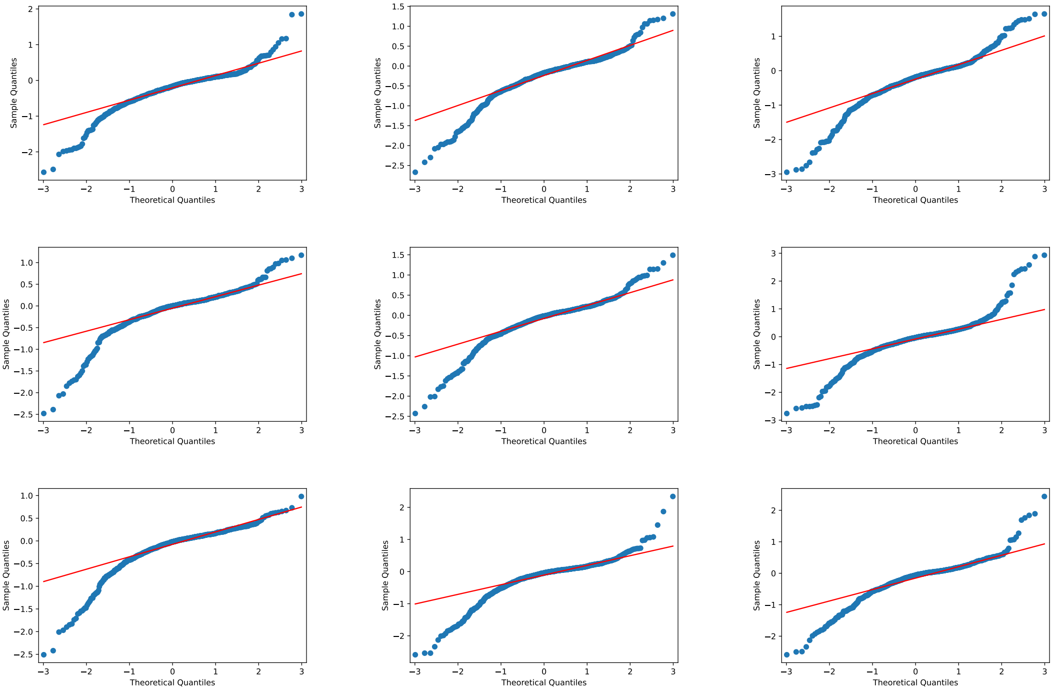

Figure 28. *Q-Q plots of the difference between the original forecast for Explorer I mission versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

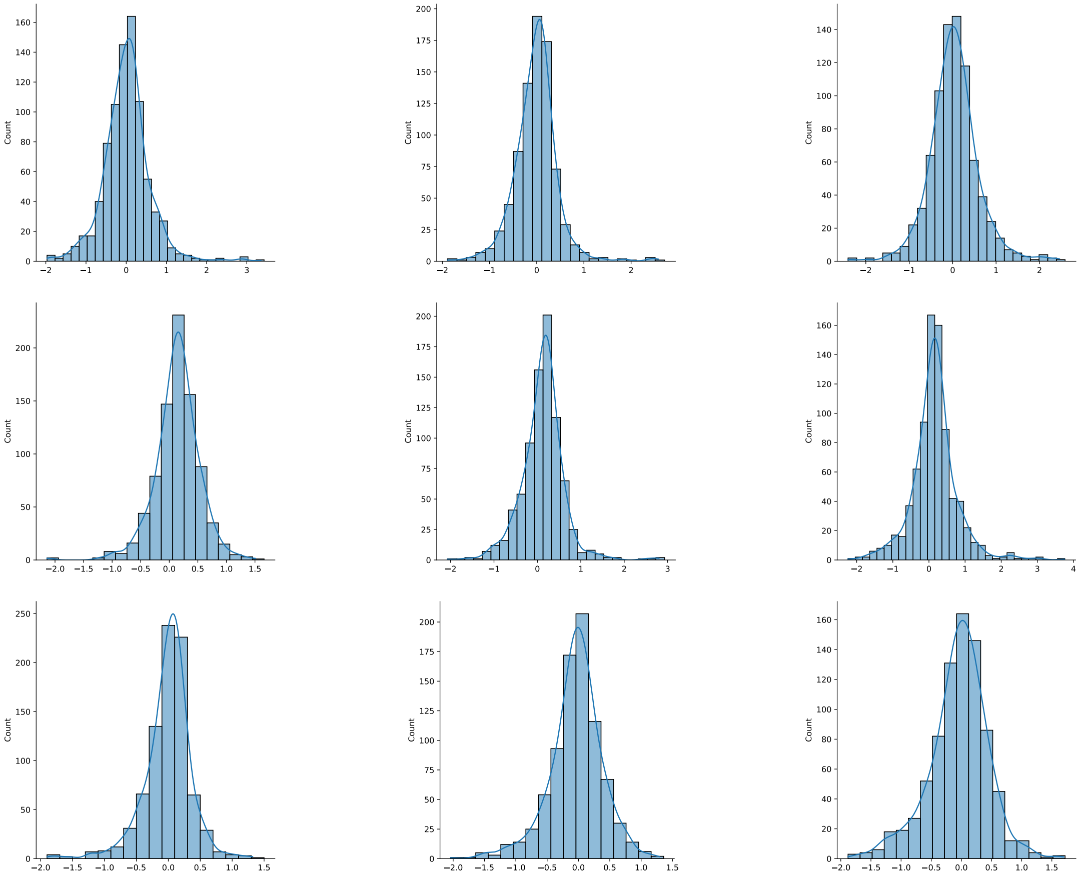

Figure 29. *Histograms of the difference between the original forecast for M6 summer period versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

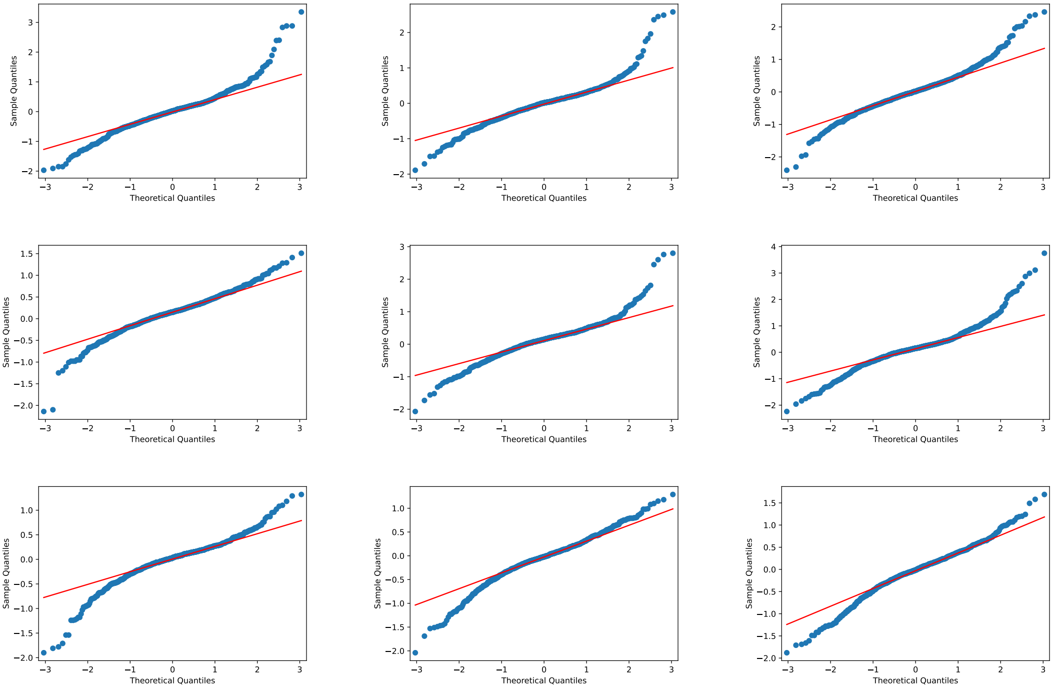

Figure 30. *Q-Q plots of the difference between the original forecast for M6 summer period versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

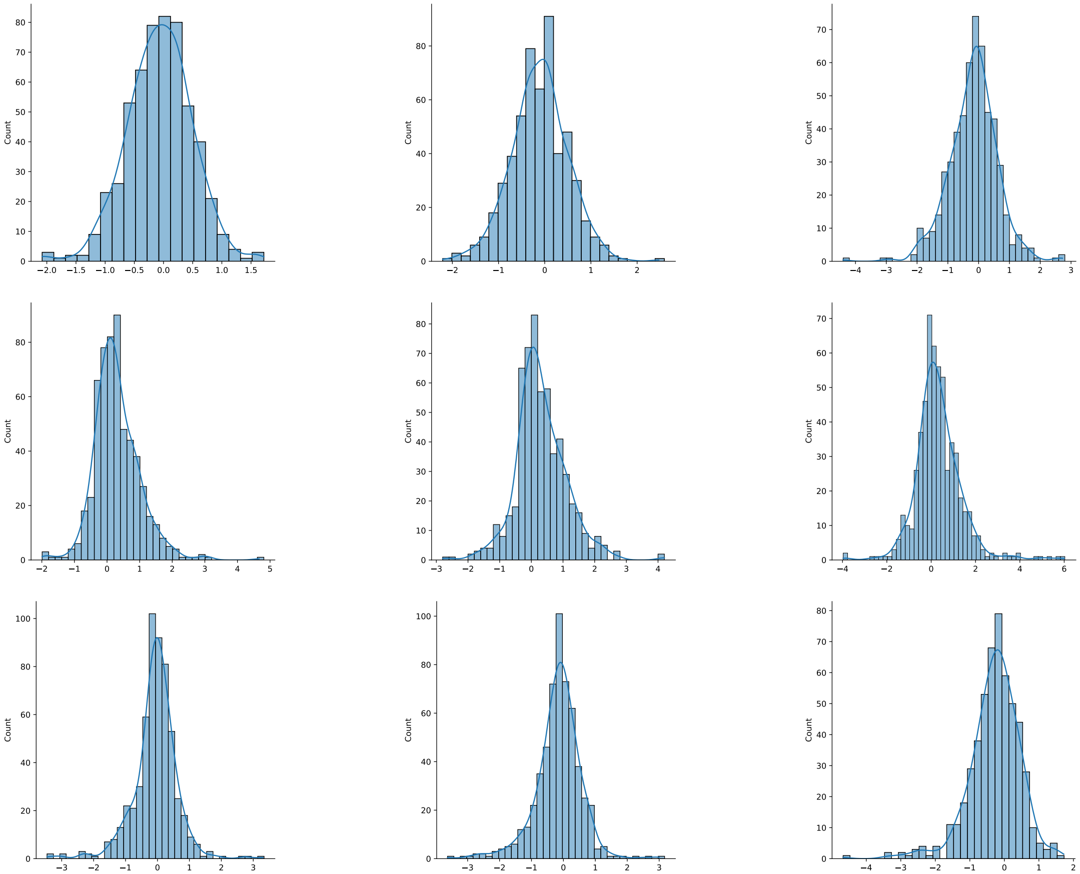

Figure 31. *Histograms of the difference between the original forecast for M6 winter period versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

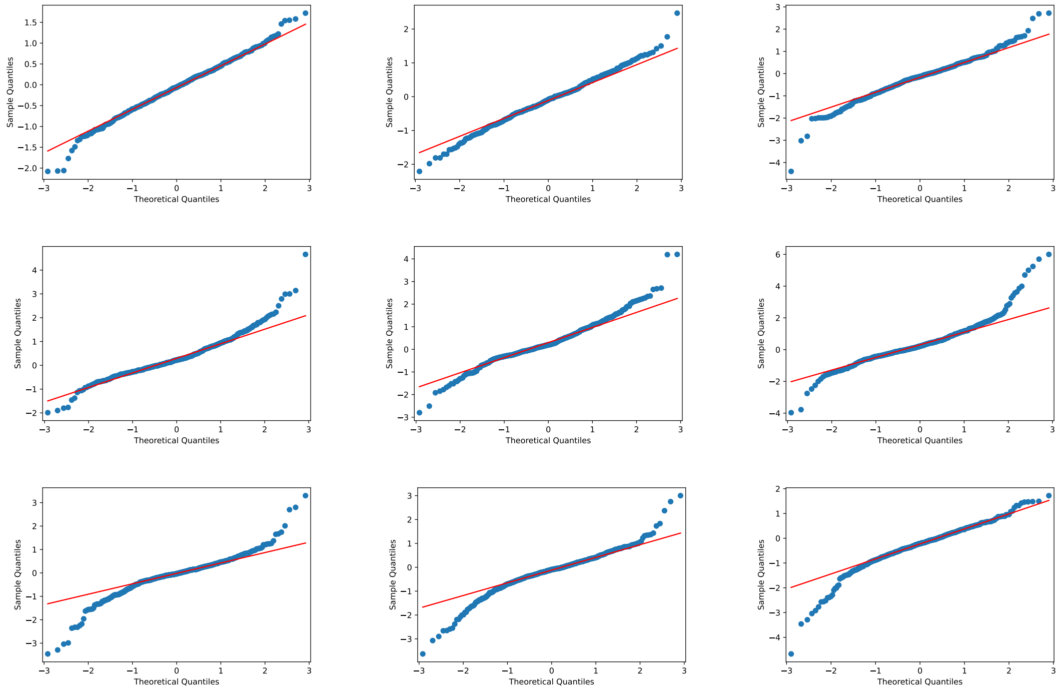

Figure 32. *Q-Q plots of the difference between the original forecast for M6 winter period versus the actual observed value. From top to bottom: DWD, WW3, MF; from left to right: 24H, 48H, 72H.*

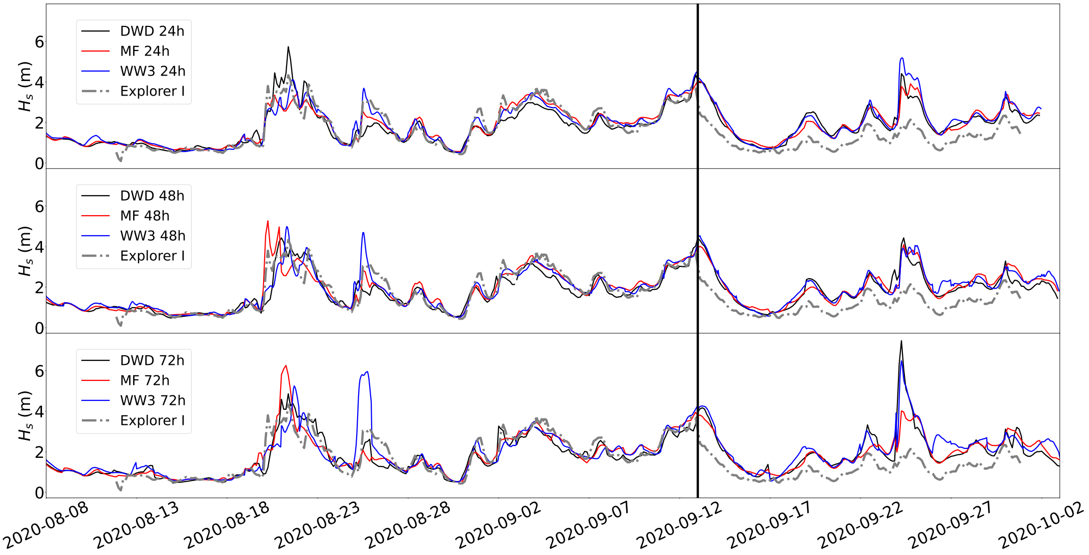

Figure 33. *Comparing actual H_s to the 24 hour (top), 48 hour (middle), and 72 hour (bottom) forecast H_s for the first Explorer mission. The vertical black line represents the time when Explorer started drifting. The discrepancy seen in the actual H_s recorded and the predicted H_s after that point is due to the Explorer drifting towards Connemara.*

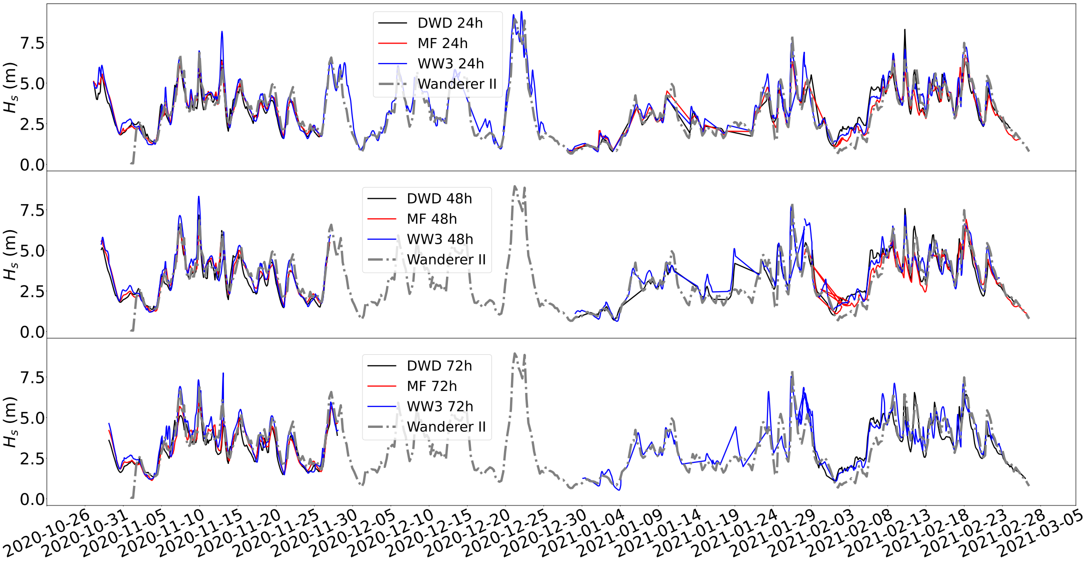

Figure 34. *Comparing actual H_s to the 24 hour (top), 48 hour (middle), and 72 hour (bottom) forecast H_s for the second Wanderer mission. Note that during December 2020 - January 2021 there were new developments to the data collecting algorithms, hence the authors did not account for that data in the post-processing.*

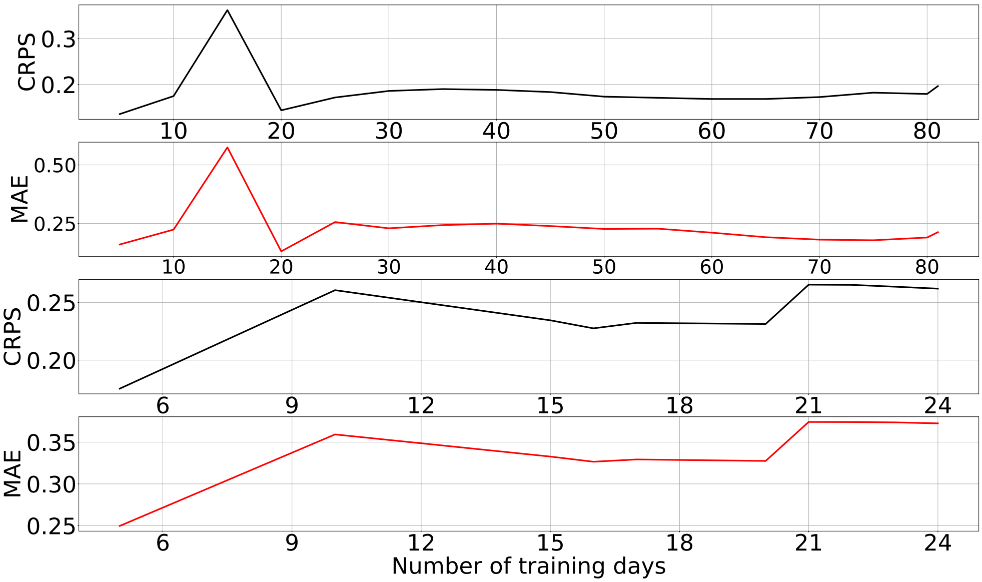

Figure 35. *Comparison of second Wanderer mission (top panels) training period lengths for H_S: MAE (top, in meters), CRPS (bottom), and Explorer first mission (bottom panels)..*

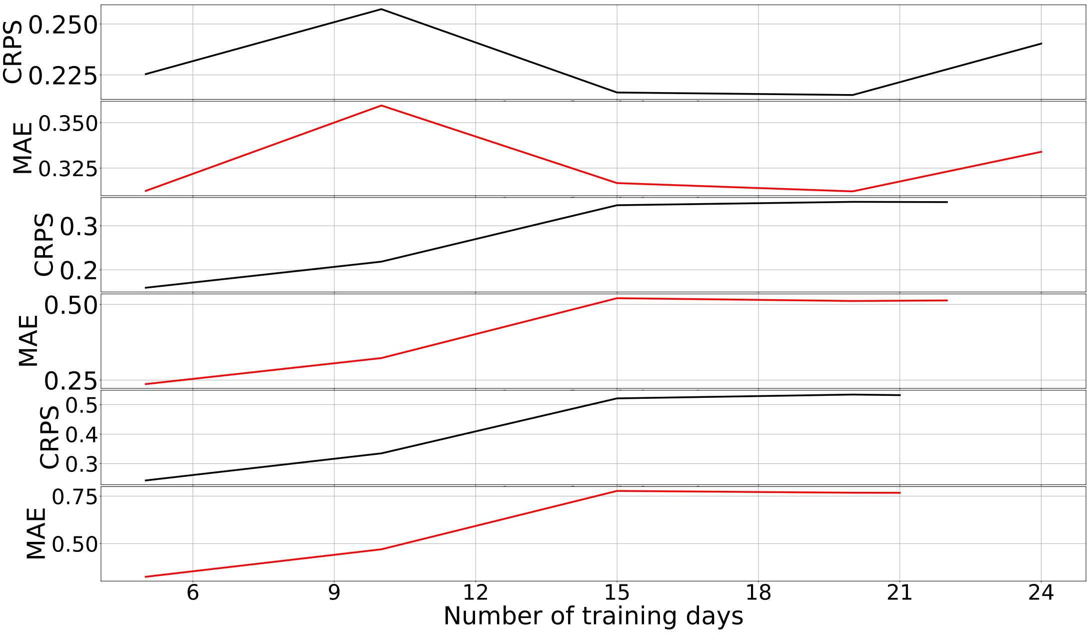

Figure 36. *Comparison of Explorer 24H (top), 48H (middle), 72H (bottom) forecasts training period lengths for H_s: MAE (top, in meters), CRPS (bottom).*

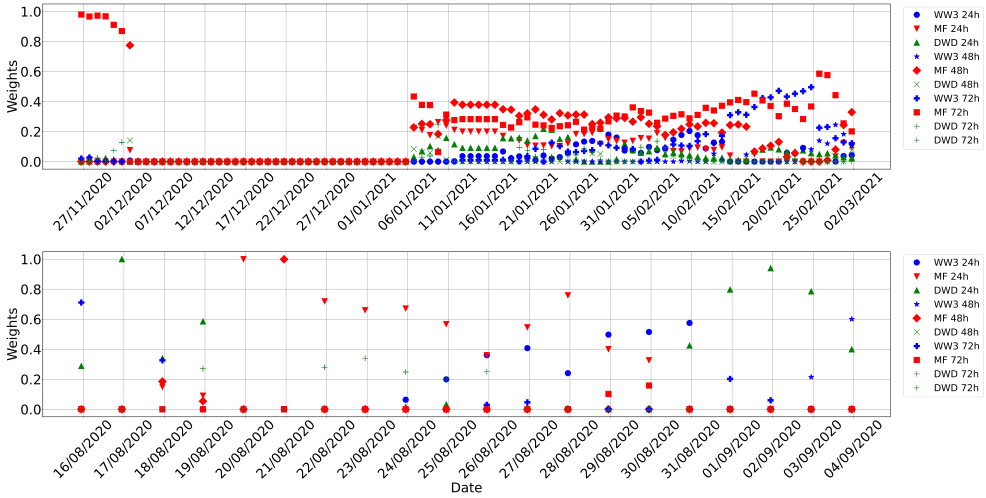

Figure 37. *Wanderer II weights of individual forecast models. It is clearly visible, how MF is dominating the weight count towards the higher contribution to the ensemble forecast (top panel). Explorer I individual weights of every forecast contributing to the overall trained prediction. We see how with time, the higher contributor changes. For example, we see MF dominating between 20 August and 28 August, and then WW3 taking over for three days, followed by DWD (bottom panel). .*

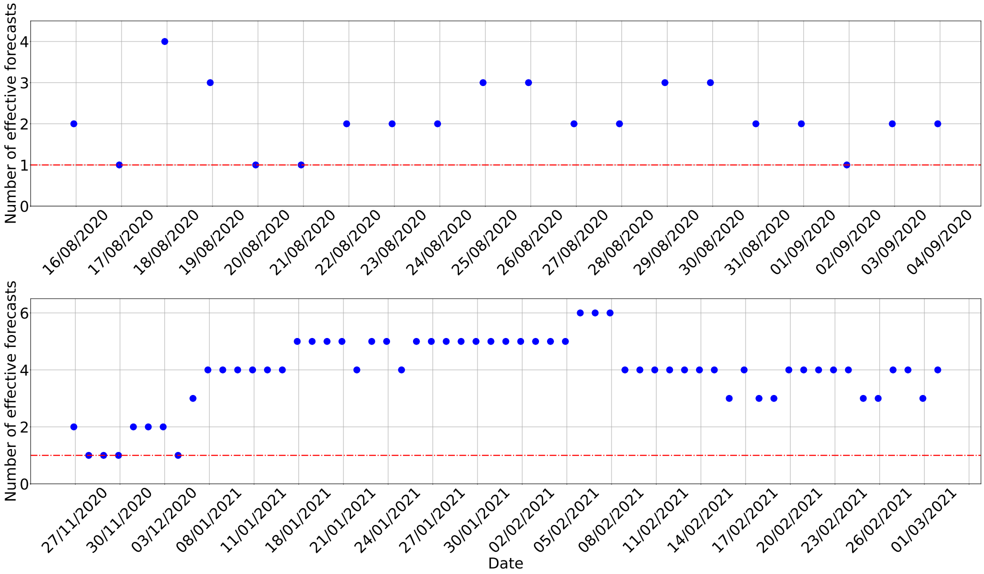

Figure 38. *Explorer I number of effective forecasts (top panel). Wanderer II number of effective forecasts (bottom panel)..*