Supplement of Biogeosciences, 21, 2159–2176, 2024
https://doi.org/10.5194/bg-21-2159-2024-supplement


Biogeosciences Open Access EGU

*Supplement of*

# Assessing improvements in global ocean $pCO_2$ machine learning reconstructions with Southern Ocean autonomous sampling

**Thea H. Heimdal et al.**

*Correspondence to:* Thea H. Heimdal (theimdal@ldeo.columbia.edu)

## S1. Supplementary Text

The hyperparameters for the XGB algorithm used in this study were fixed for all experiments. As we are comparing how sampling impacts the reconstruction, changing the decision trees and depth levels for each experiment would make it difficult to assess whether or not potential changes in bias and RMSE are due to the different sampling strategies or the optimization process. However, **Figure S6** demonstrates a statistically significant difference between train and test set error for the 'SOCAT-baseline' experiment, which may indicate overfitting in our ML model. This suggests that further tuning of the hyperparameters of our ML algorithm might increase generalization skill, and thus reduce test and 'unseen' reconstruction errors.

However, further tuning of the algorithm is not the purpose of this study nor is it necessary for the evaluation performed here. As the only factor we change between the experiments is additional Southern Ocean sampling (e.g., the SOCAT mask, algorithmic approach and hyperparameters are the same), we can compare the experiments and understand how different sampling patterns and strategies would change skill in $pCO_2$ reconstructions compared to SOCAT-sampling only.

**Figure 3** (in main text) shows that errors are higher in locations where SOCAT observations are scarce, such as in the Southern Ocean. The improvements in bias and RMSE when USV observations are combined with SOCAT generally occur at times/locations where errors were originally high, and in the Southern Ocean where the new USV "observations" originate from (e.g., **Figs. 4 and 7**). The additional USV observations thus most improve predictions in surrounding areas because the augmented training set contains these similar points. However, note that the lower error values shown for the training set do not impact the final error metrics presented in our study, as ~ 99 % of the reconstruction consists of 'unseen' data points (**Figs. S1, S2**).

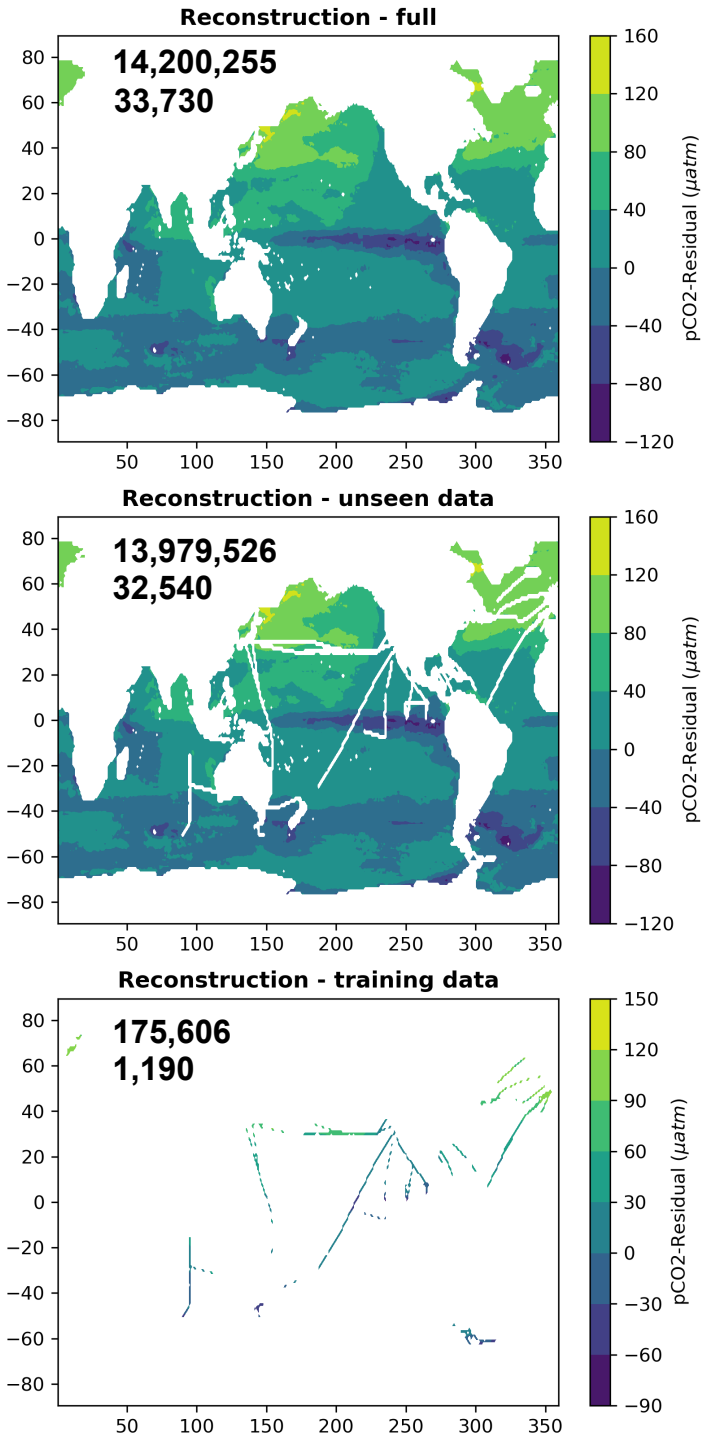


**Figure S1**: Example of one-month of reconstructed and training data to illustrate data sparsity. The full $pCO_2$-Residual reconstruction (all 1°x1° grid cells of the testbed, except for those masked or filtered out; see **Section 2.1** and **2.2**) (top), 'unseen' reconstruction (middle), and training data from the testbed (bottom). The maps show output from CESM member 001 for the month of March 2016 for the 'SOCAT-baseline'-run. Numbers on panels represent the total monthly 1°x1° grid cells for this month and for the entire testbed period (1982-2016).

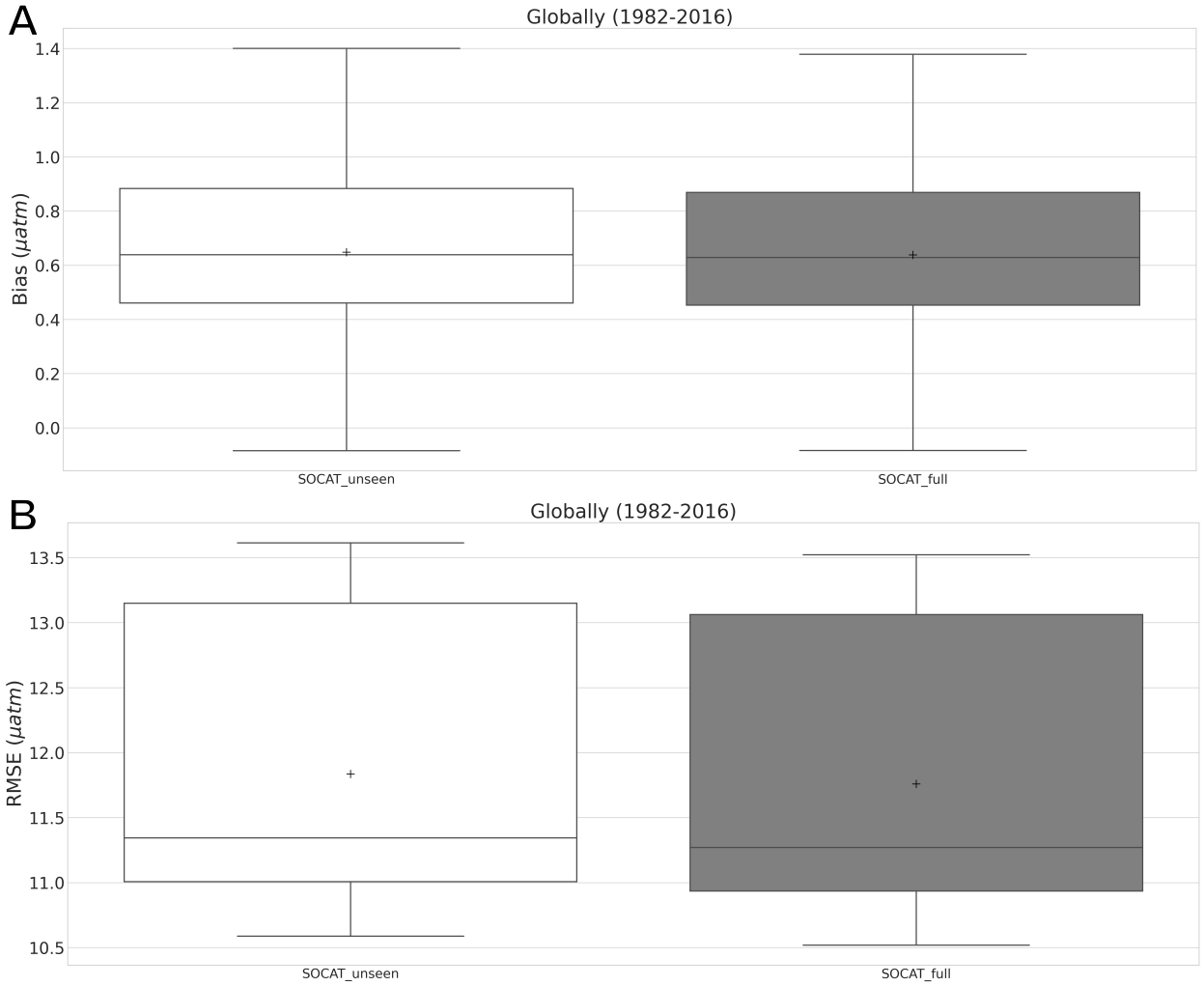


**Figure S2**: Spread of bias (**a**) and RMSE (**b**) for the 75 members of the Large Ensemble Testbed for the 'unseen' and full reconstruction for the 'SOCAT-baseline'-run. The 'unseen' reconstruction represents independent data, i.e., all 1°x1° grid cells that do not correspond to SOCAT or Saildrone USV observations, and is not part of the training set.

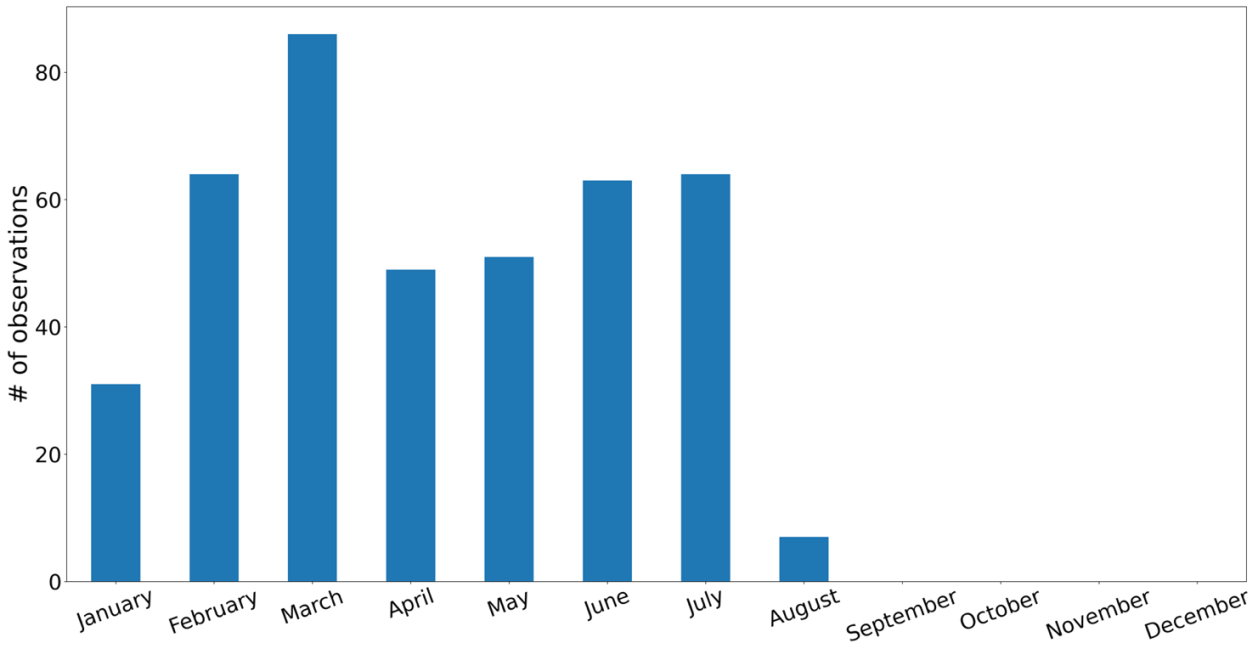


**Figure S3**: Number of monthly 1°x1° observations from the 'real-world' Saildrone USV journey in Sutton et al. (2021).

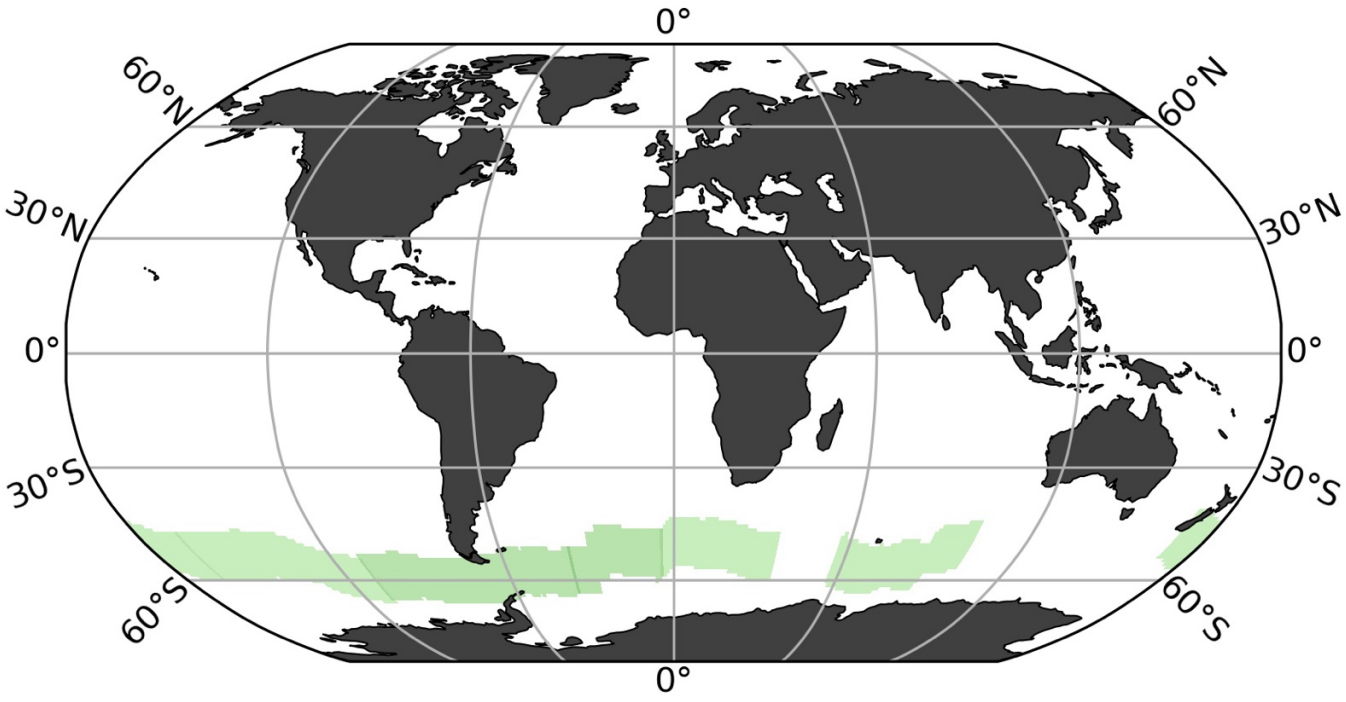


**Figure S4**: Map showing the spatial extent of sampling with 13 Saildrone USVs for run 'x13_10Y_J-A' and 'x13_10Y_W'. The 'real-world' USV track from Sutton et al. (2021) was repeated six times by 1°, covering an additional 6° both north and south.

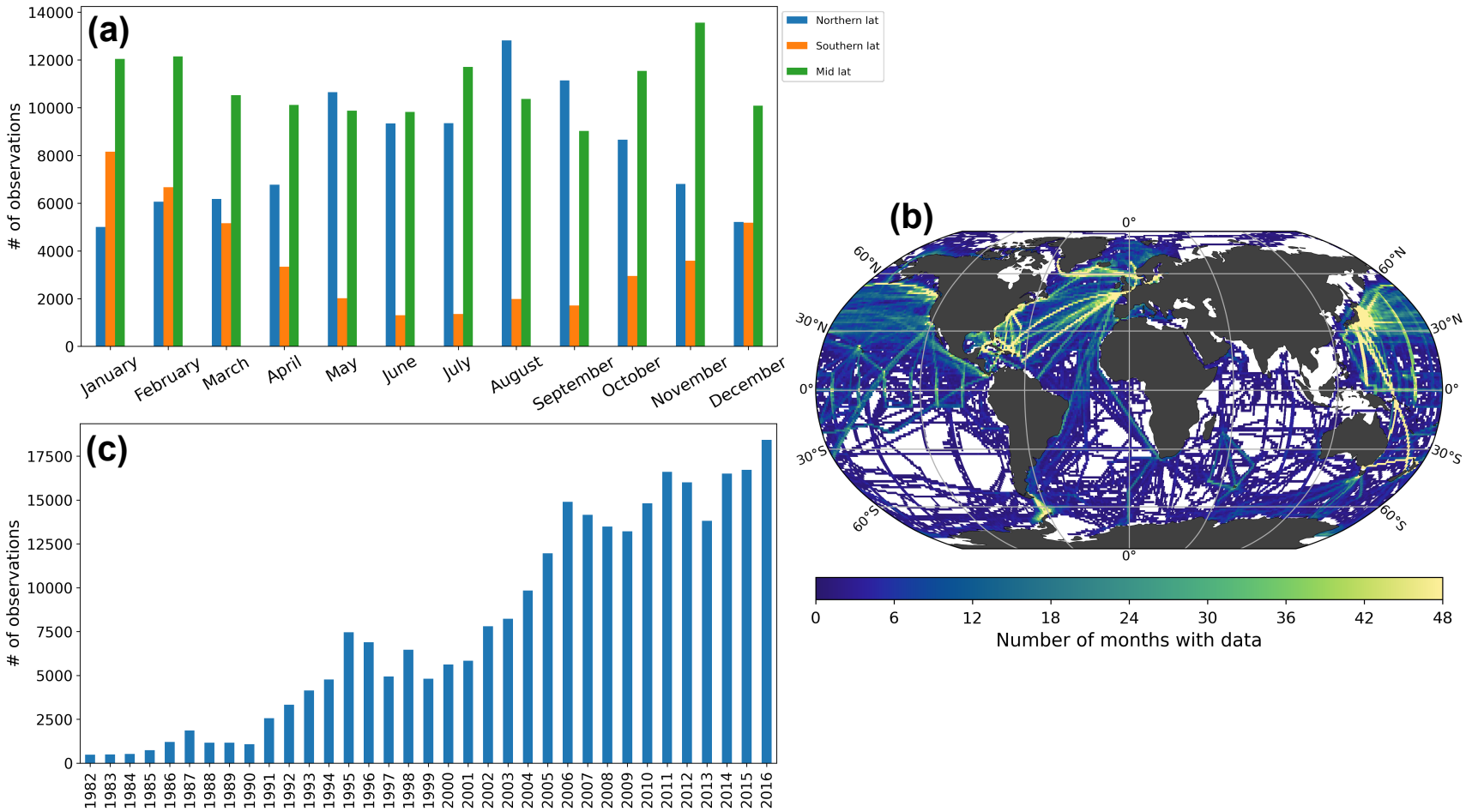


**Figure S5**: SOCAT observations from 1982-2016. Number of monthly 1°x1 SOCAT observations per month divided by latitude band (**a**). "Northern lat" = north of 35° N. "Southern lat" = south of 35° S. "Mid lat" = between 35° S and 35° N. Spatial extent of SOCAT tracks and number of monthly observations per 1°x1° grid cell over 1982-2016 (**b**). Number of monthly 1°x1 SOCAT observations per year (**c**).

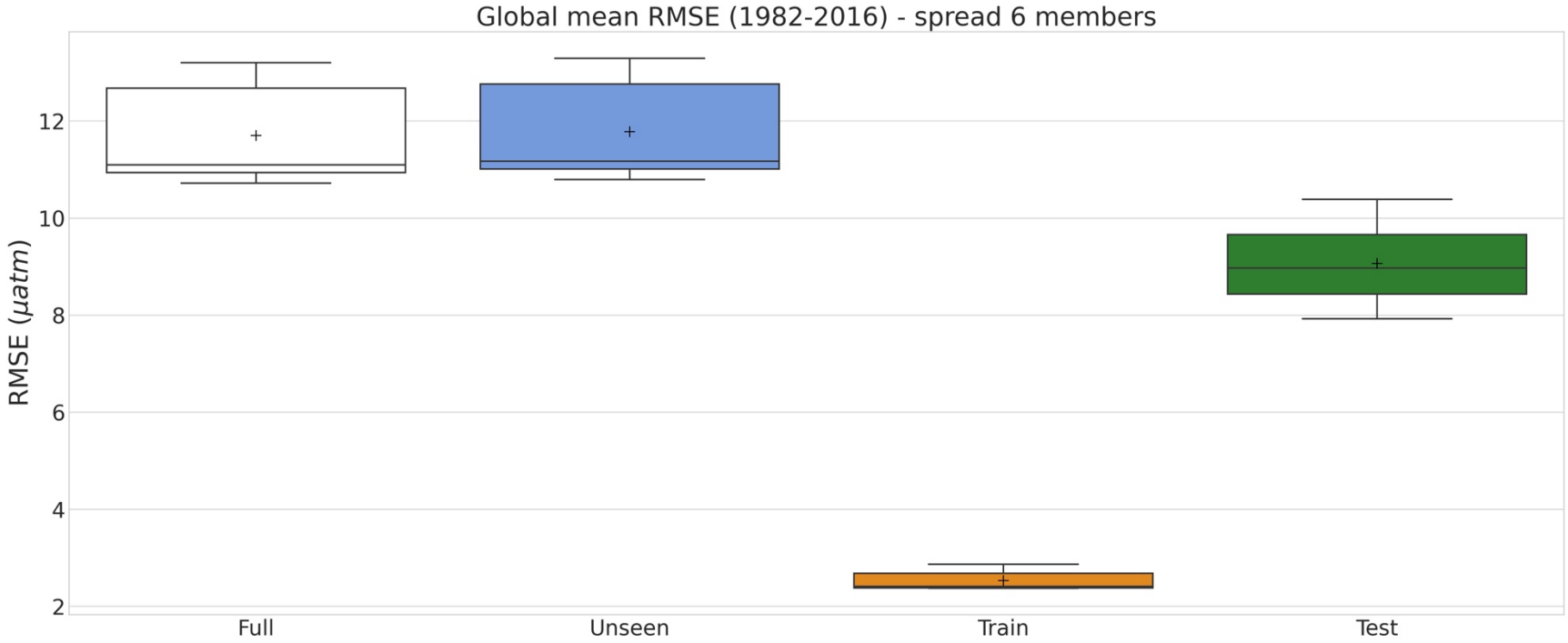


**Figure S6.** Global mean RMSE (1982-2016) for full, unseen, train and test sets for the 'SOCAT-baseline' experiment (reconstructions compared to the 'model truth'). The boxplots show the ensemble spread of six members of the LET, two from each of the three Earth System Models. Boxes = interquartile range (IQR). Crosses = mean. Horizontal bars inside boxes = median. Horizontal bars outside boxes = minimum and maximum value.

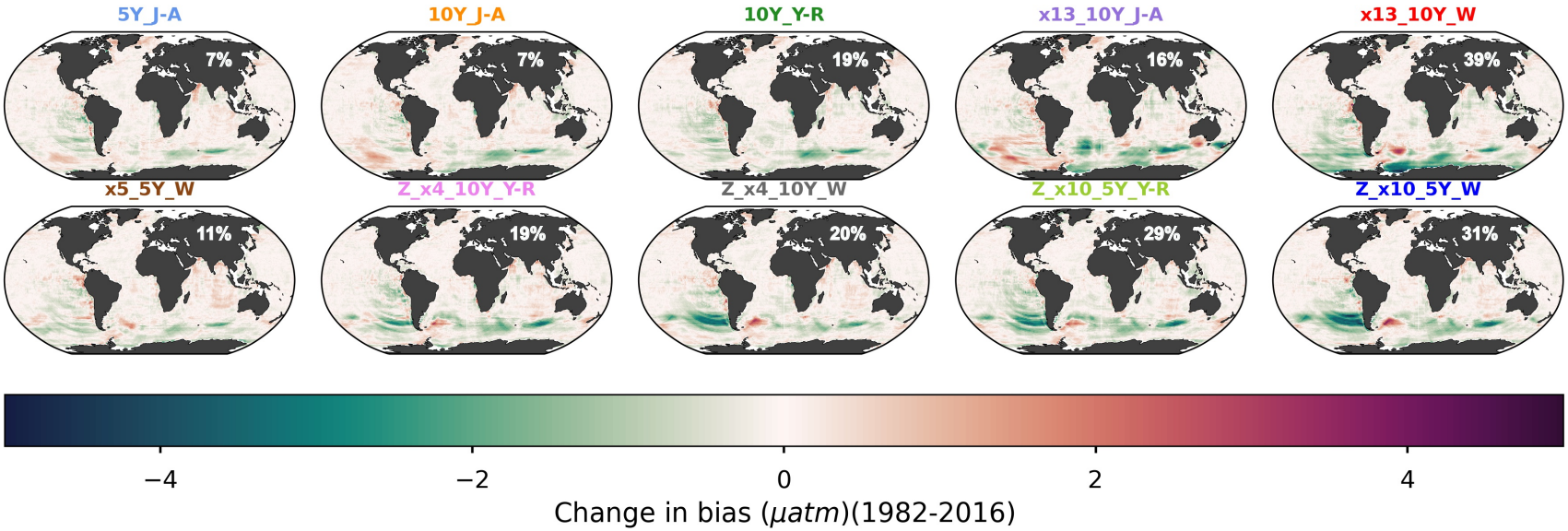


**Figure S7**. Change in bias when comparing the different Saildrone USV sampling strategies to the 'SOCAT-baseline'. The improvement compared to the 'SOCAT-baseline' is shown by percent values (absolute value, global average for 1982-2016). The 'one-latitude'-run 'x13_10Y_W' demonstrates the most significant improvement in absolute bias (i.e., 39%).

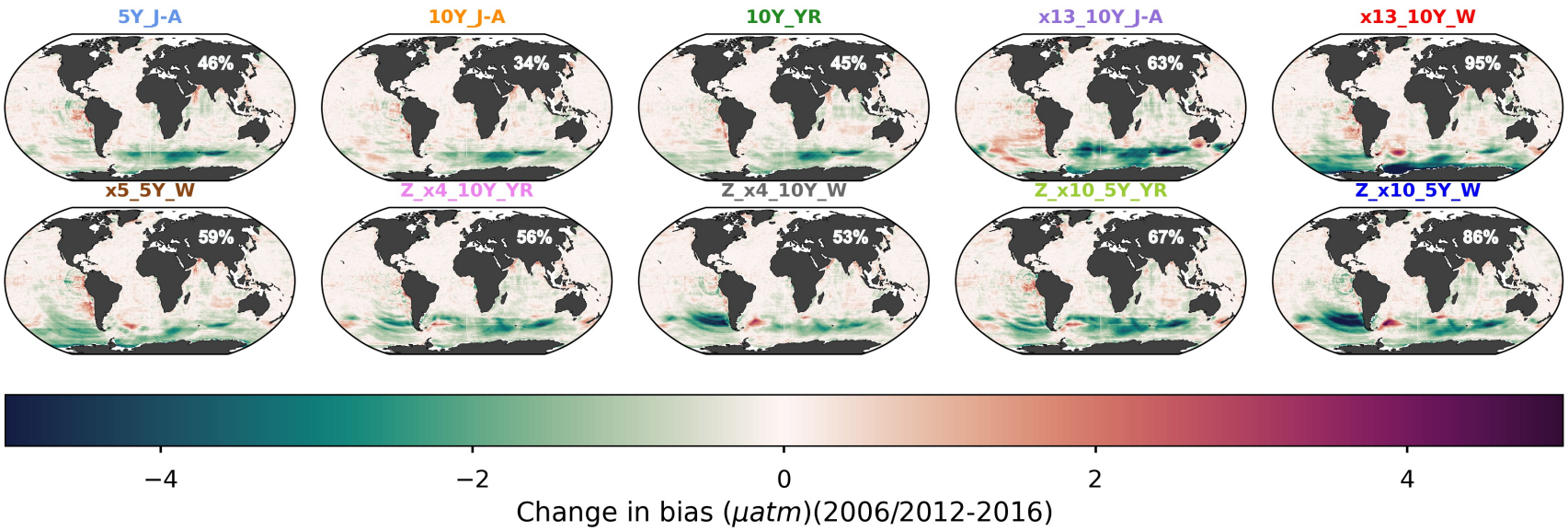


**Figure S8**: Same as **Figure S7**, but averaged over the years of USV sample addition (2006-2016 or 2012-2016). There is significant improvement in bias for all runs compared to the entire testbed period (1982-2016; **Fig. S7**).

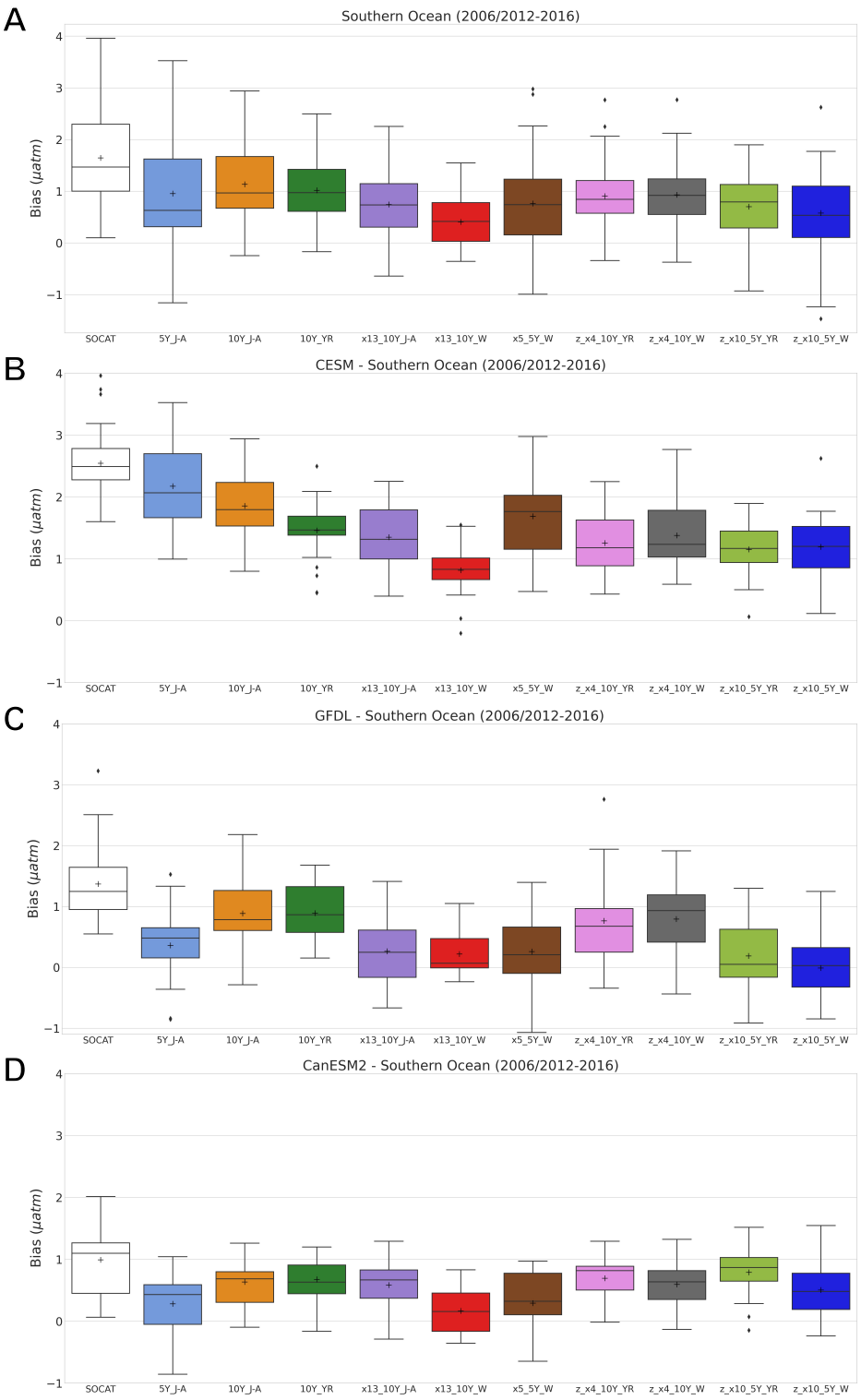


**Figure S9**: Spread in bias for the full Large Ensemble Testbed (75 members; **A**), 25 CESM members (**B**), 25 GFDL members (**C**) and 25 CanESM2 members (**D**) over the Southern Ocean (< 35° S) and duration of Saildrone USV sampling (2006-2016 or 2012-2016). The 'SOCAT-baseline' is averaged over 2006-2016. Colored boxes = interquartile range. Bars = median. Crosses = mean. Diamonds = outliers.

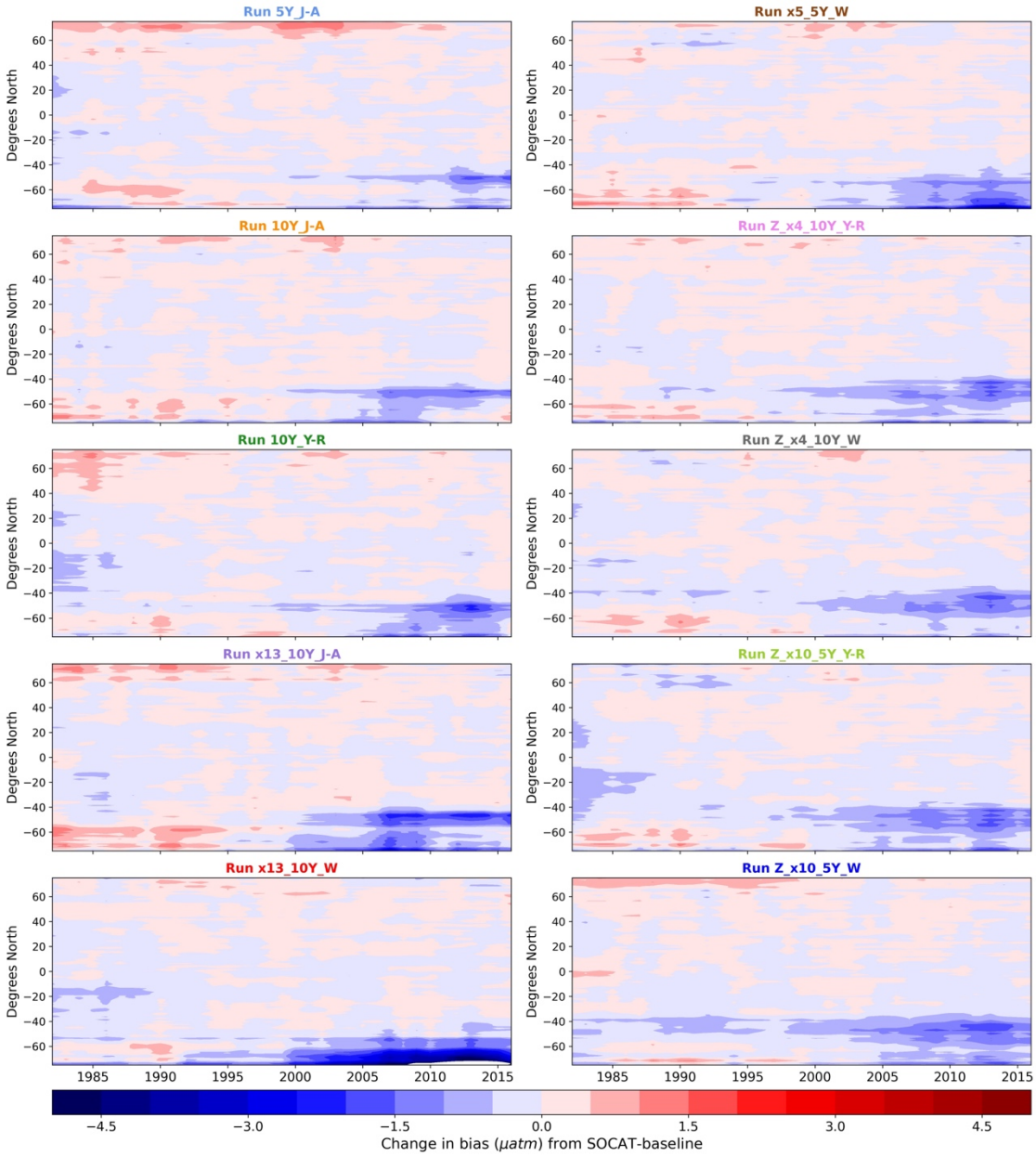


**Figure S10**: Zonal mean, annual mean Hovmöller of the change in bias when comparing the different Saildrone USV sampling strategies to the 'SOCAT-baseline'. Bias improvements expand backwards to year 2000 for all runs, which is well beyond the duration of USV additions (shown by arrow on panels).

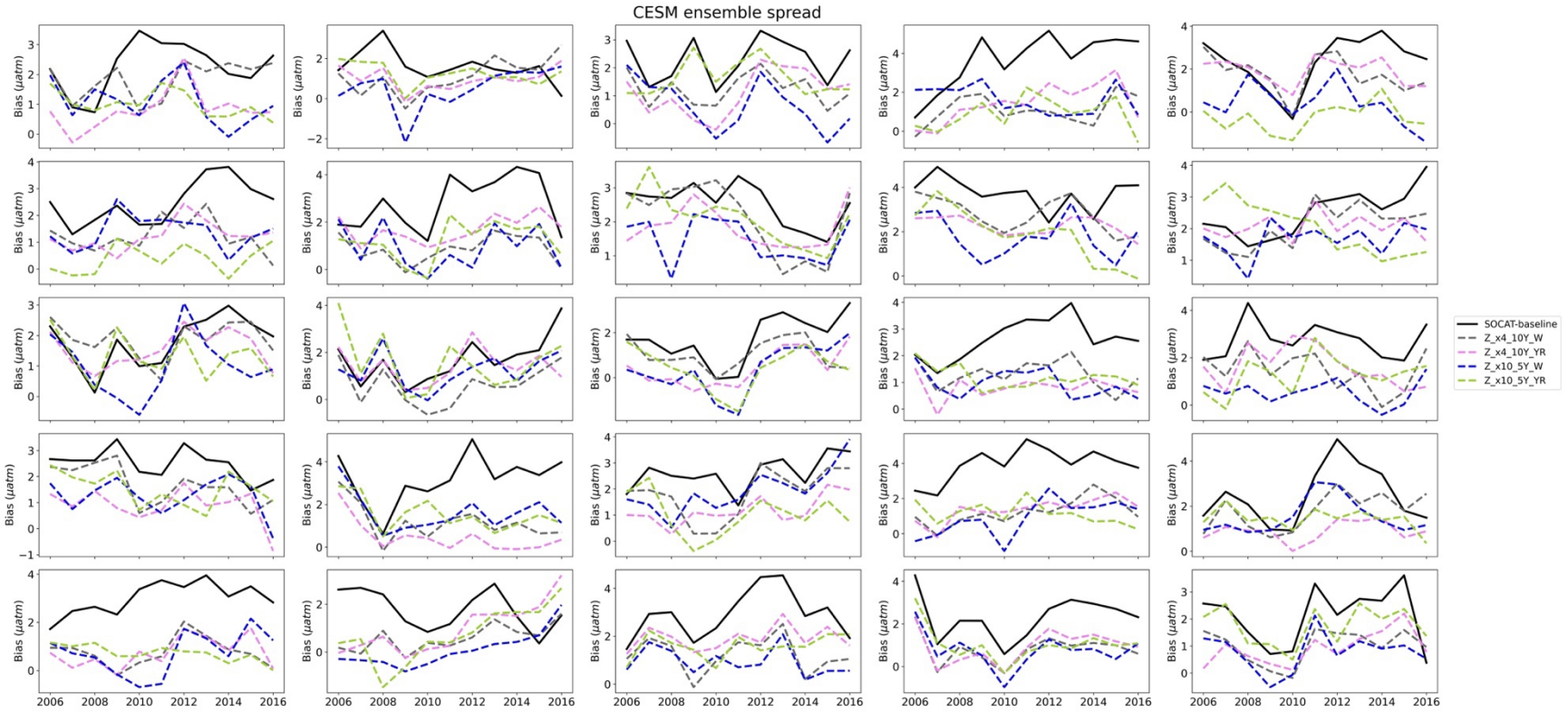


**Figure S11.** Annual mean bias (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 CESM members in the Large Ensemble Testbed.

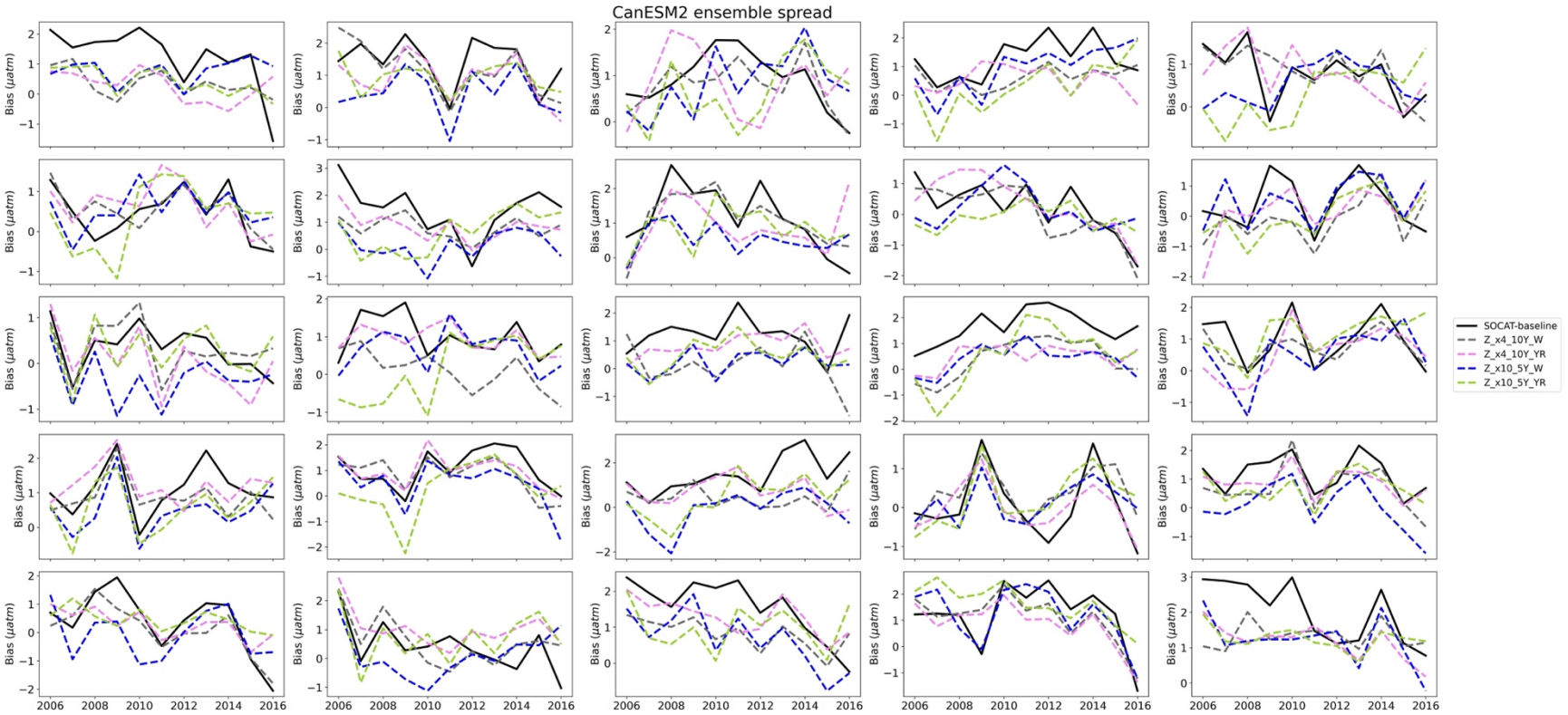


**Figure S11 continued.** Annual mean bias (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 CanESM2 members in the Large Ensemble Testbed.

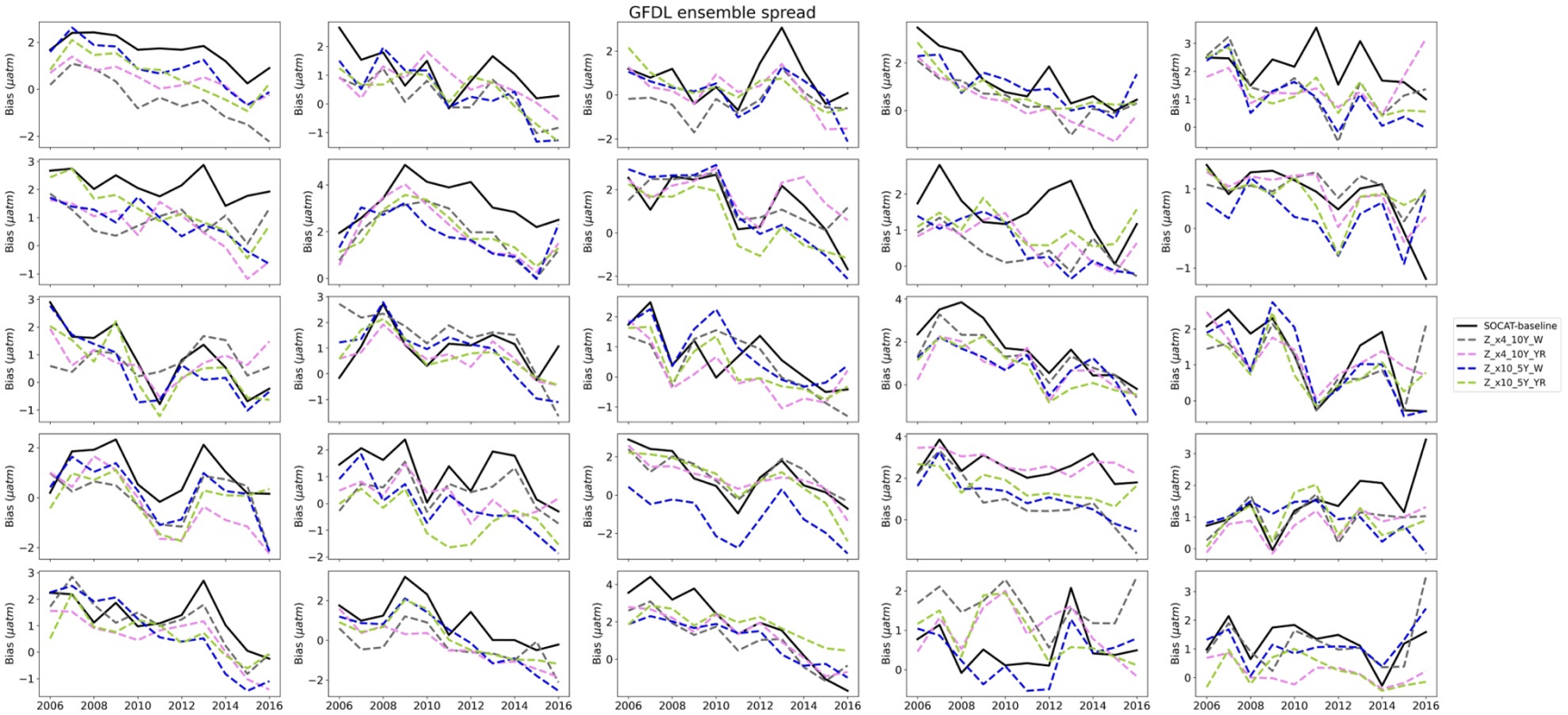


**Figure S11 continued.** Annual mean bias (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 GFDL members in the Large Ensemble Testbed.

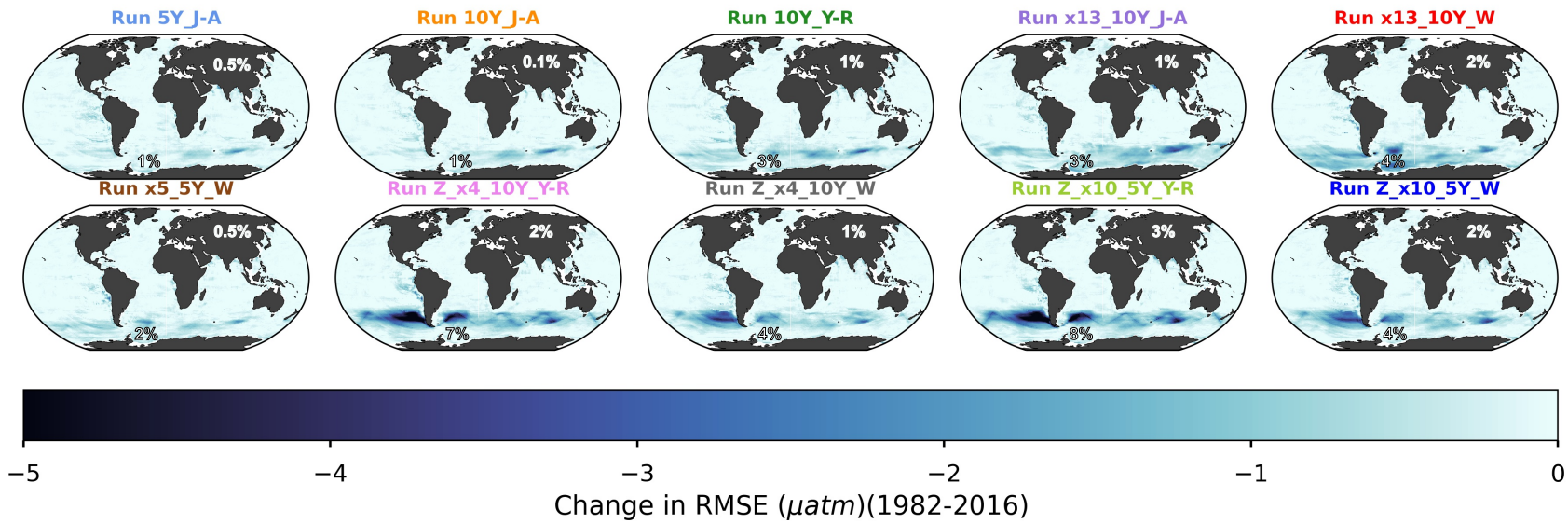


**Figure S12**: Change in RMSE when comparing the different Saildrone USV sampling strategies to the 'SOCAT-baseline'. Improvement in RMSE occurs mainly in southern latitudes (<35°S), where the baseline reconstruction demonstrated high RMSEs (**Fig. 3b**). Percent values represent improvement compared to the 'SOCAT-baseline' globally (top value) and in the Southern Ocean (< 35° S) (bottom value).

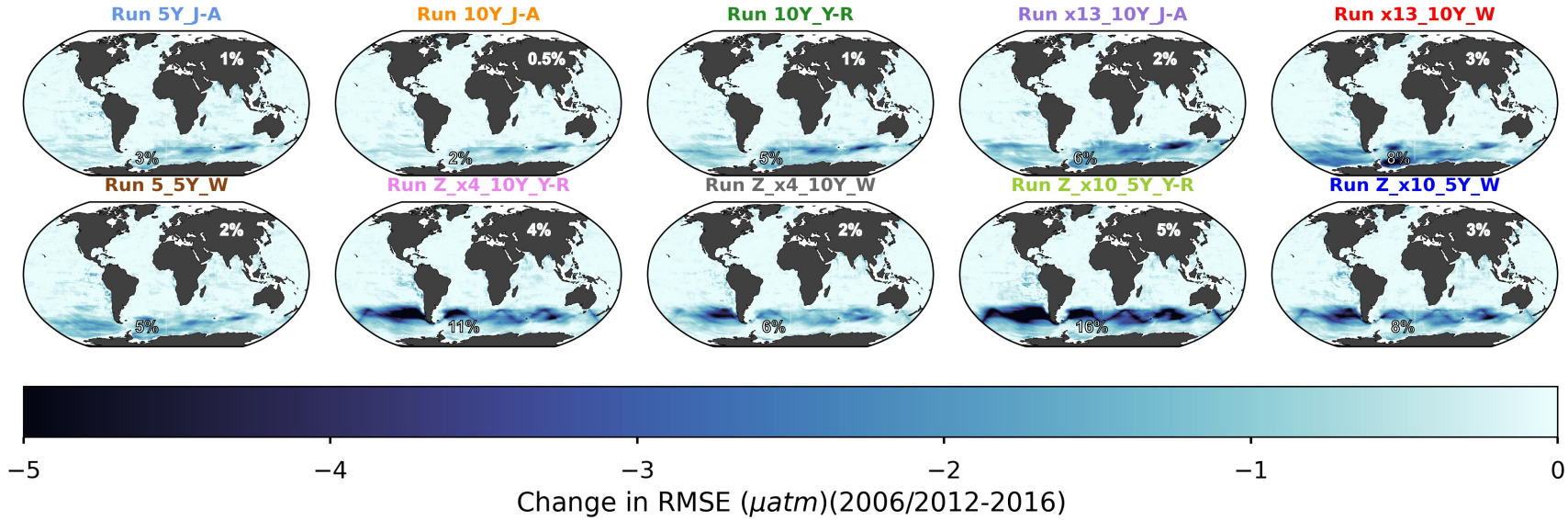


**Figure S13**: Same as **Figure S12**, but averaged over the years of USV sample addition (2006-2016 or 2012-2016). Compared to the mean of the entire testbed period (1982-2016; **Fig. S12**), there is improvement in RMSE for all runs. Percent values represent improvement compared to the 'SOCAT-baseline' globally (top value) and in the Southern Ocean (< 35° S) (bottom value).

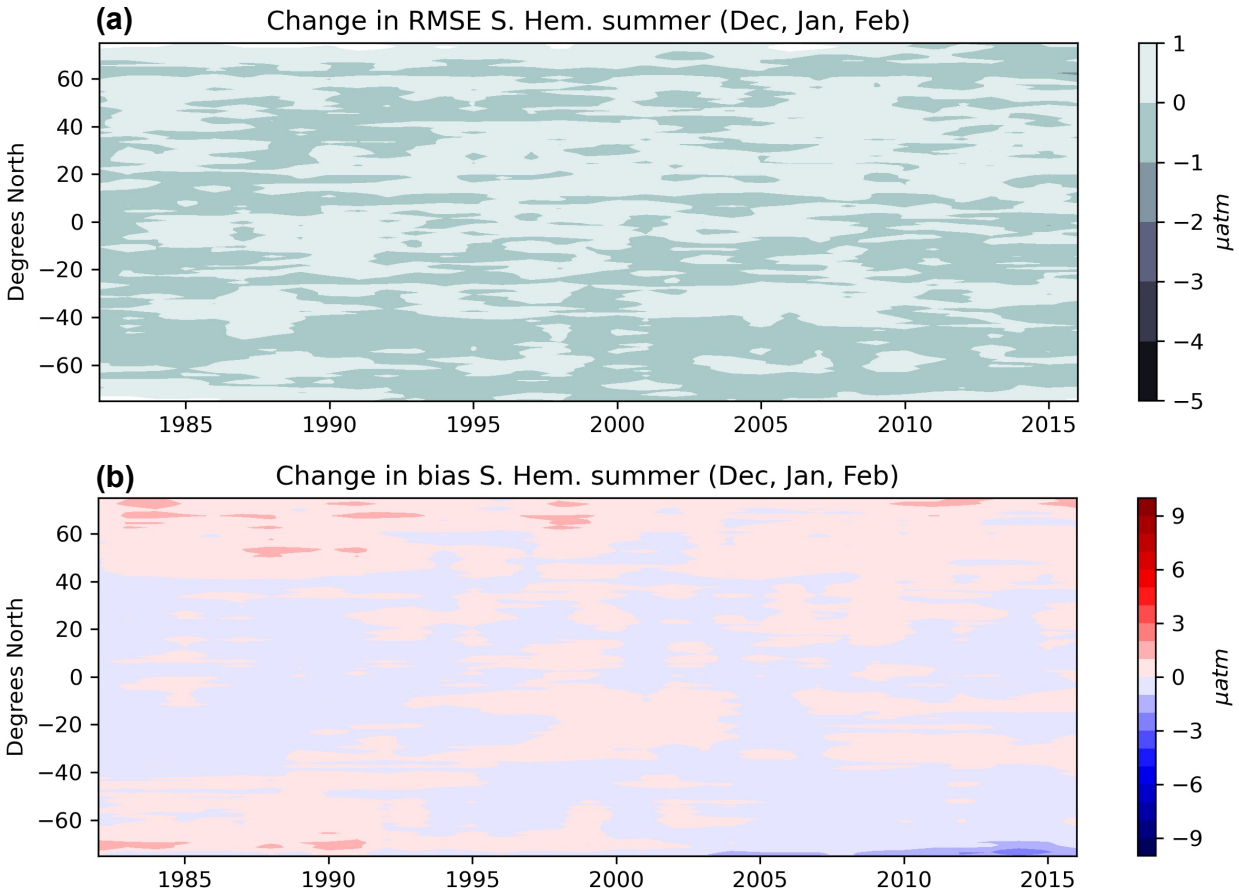


**Figure S14**: Zonal mean, DJF (December, January, February) mean Hovmöller change in RMSE (**a**) and bias (**b**) for run 'x13_10Y_W' compared to the 'SOCAT-baseline'. There is minimal improvement in RMSE and bias during southern hemisphere summer months.

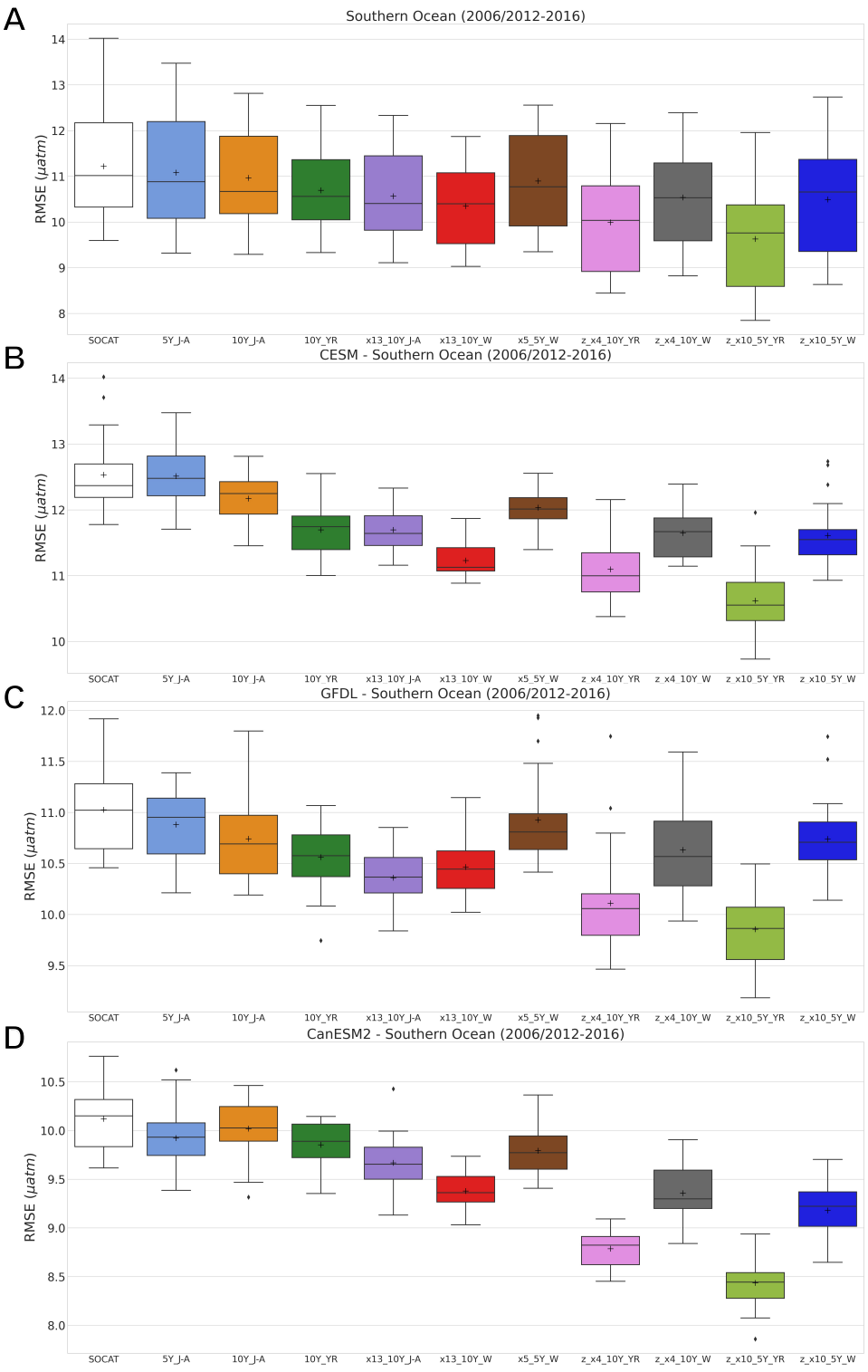


**Figure S15**: Spread in RMSE for the full Large Ensemble Testbed (75 members; **A**), 25 CESM members (**B**), 25 GFDL members (**C**) and 25 CanESM2 members (**D**) over the Southern Ocean (< 35° S) and duration of Saildrone USV sampling (2006-2016 or 2012-2016). The 'SOCAT-baseline' is averaged over 2006-2016. Colored boxes = interquartile range. Bars = median. Crosses = mean. Diamonds =outliers.

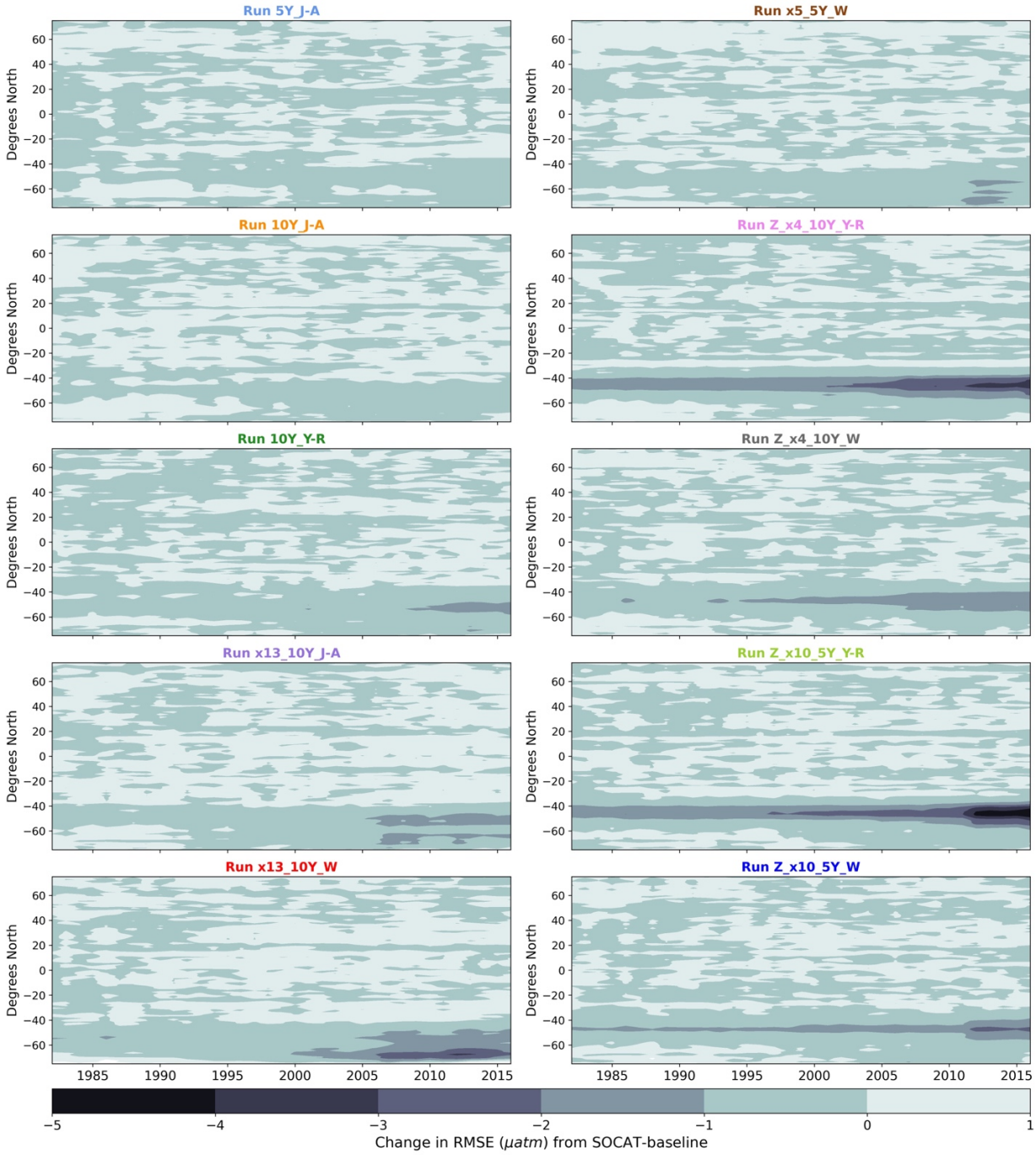


**Figure S16**: Zonal mean, annual mean Hovmöller of the change in RMSE when comparing the different Saildrone USV sampling strategies to the 'SOCAT-baseline'. The most significant improvement in RMSE is shown for the 'zigzag'-runs and the 'one-latitude'-run 'x13_10Y_W' in the Southern Ocean during the last 5-10 years when the additional Saildrone observations are introduced.

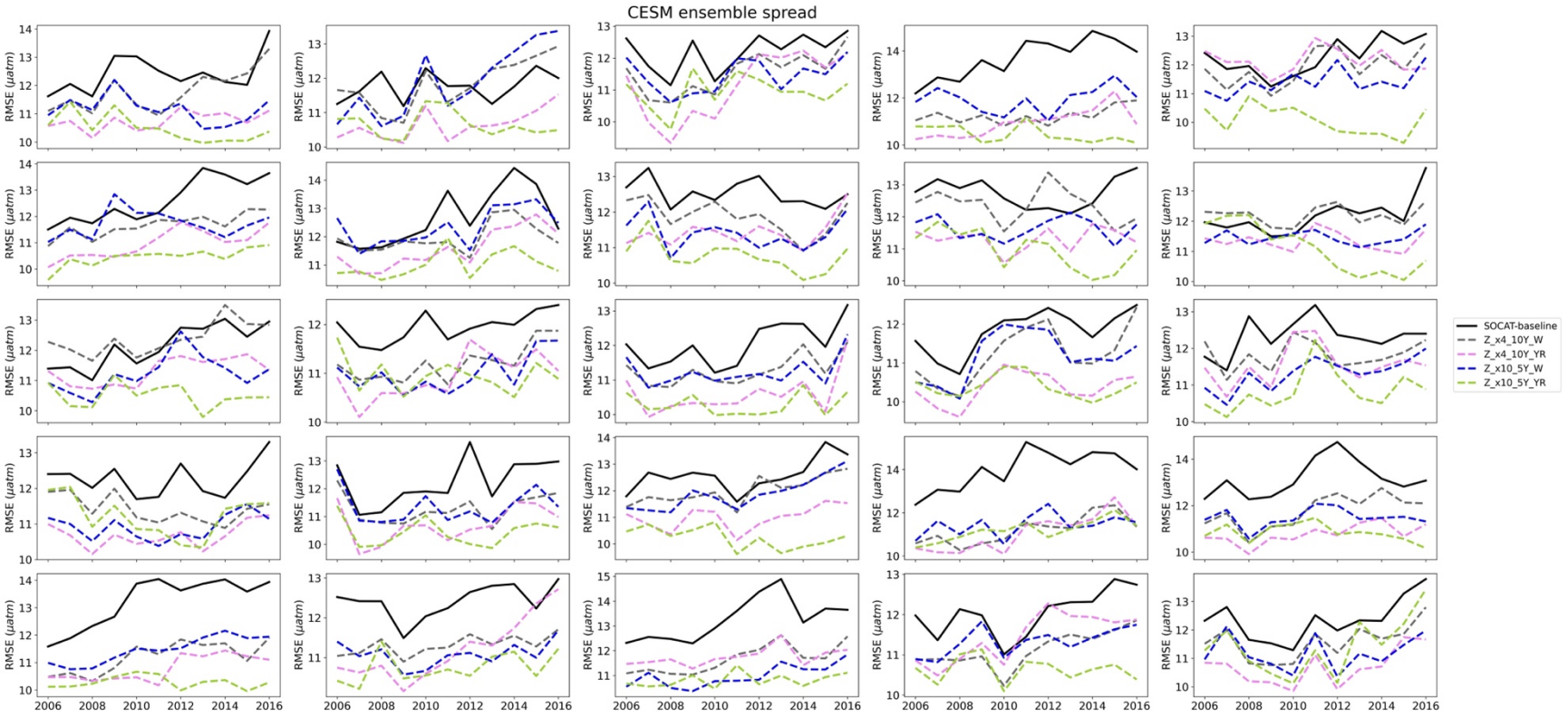


**Figure S17.** Annual mean RMSE (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 CESM members in the Large Ensemble Testbed.

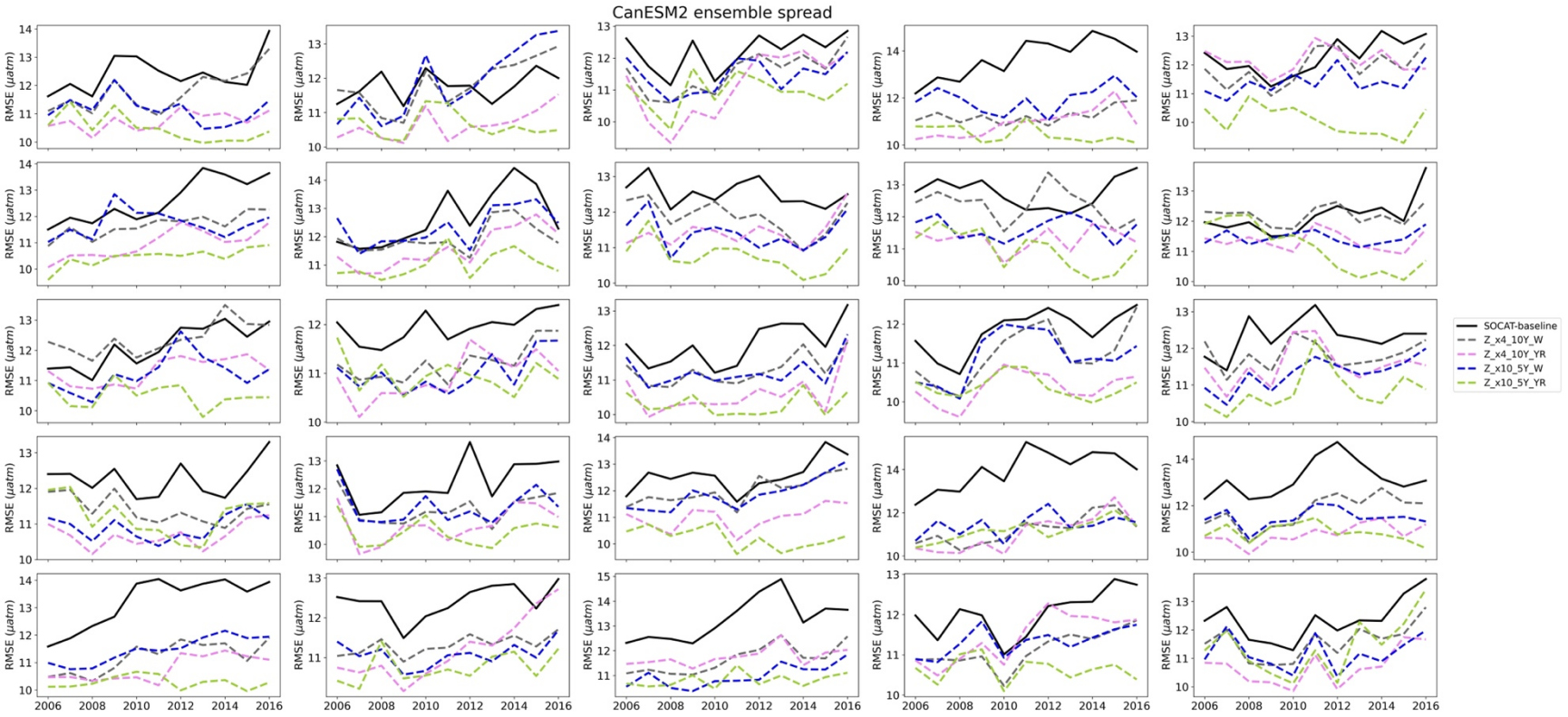


**Figure S17 continued.** Annual mean RMSE (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 CanESM2 members in the Large Ensemble Testbed.

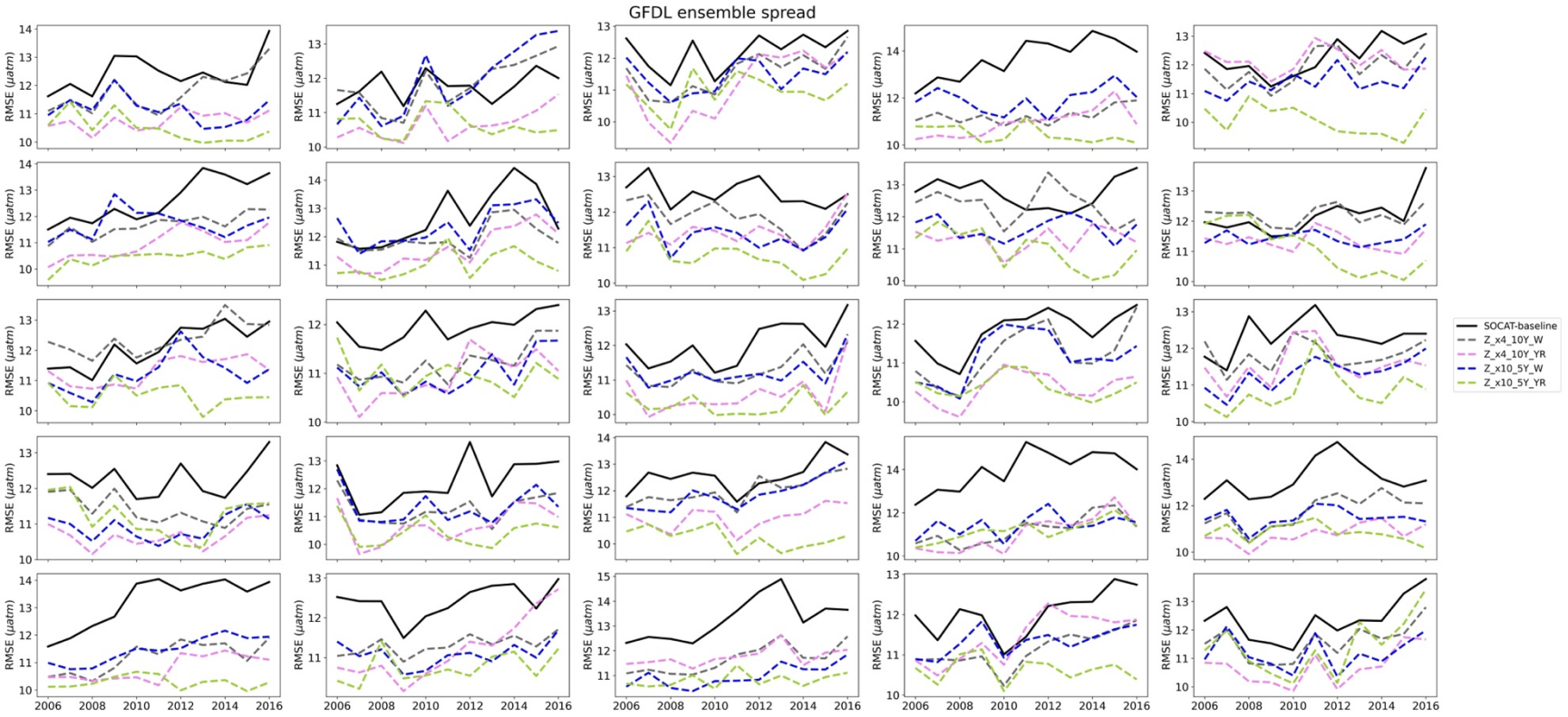


**Figure S17 continued.** Annual mean RMSE (Southern Ocean; > 35° S) for 'zigzag' runs for each of the 25 GFDL members in the Large Ensemble Testbed.

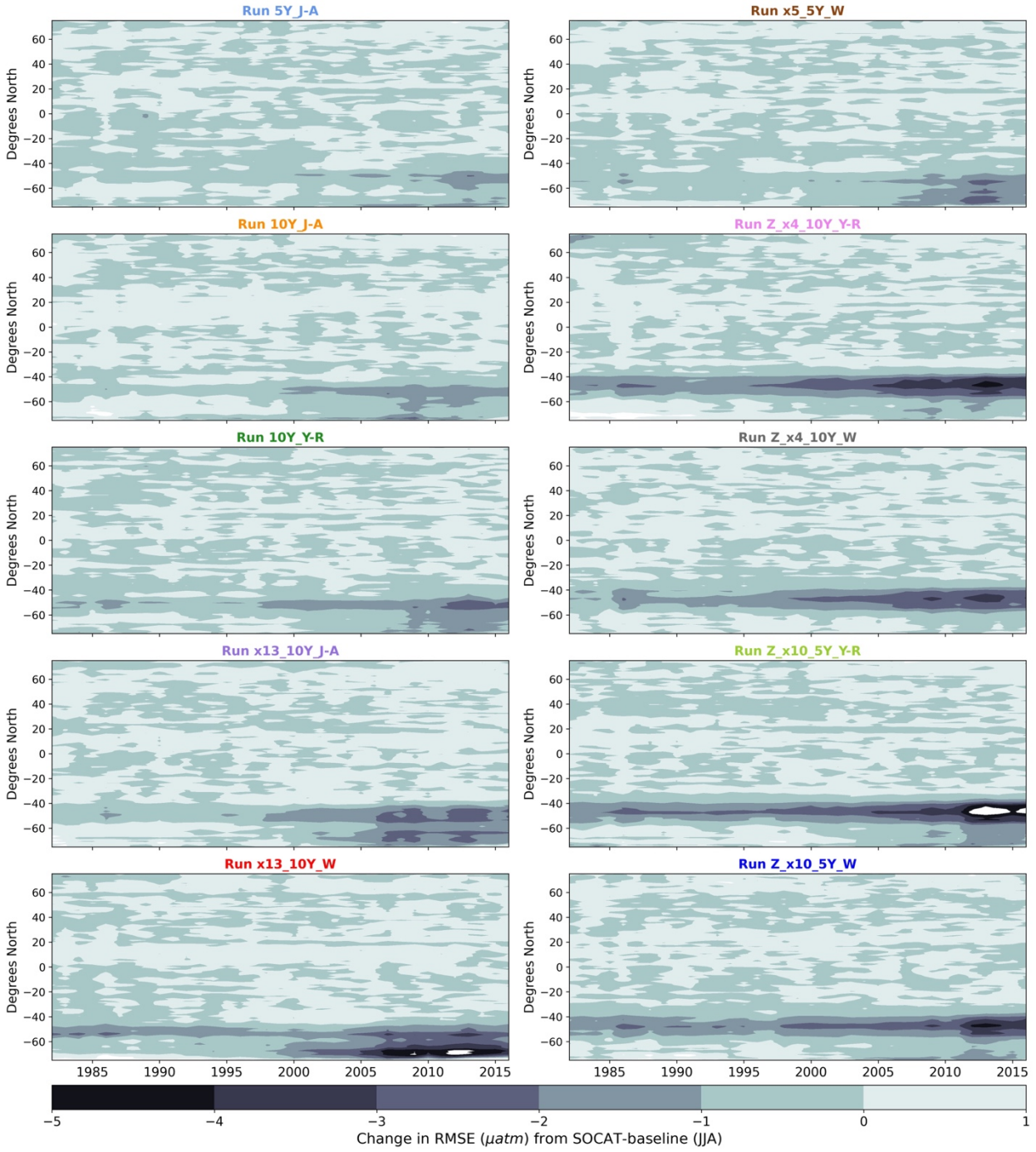


**Figure S18**: Same as **Figure S16**, but for southern hemisphere winter months (i.e., JJA; June, July, August). Compared to **Figure S16** (averaging over all months), there is improvement in RMSE both in terms of magnitude and duration.

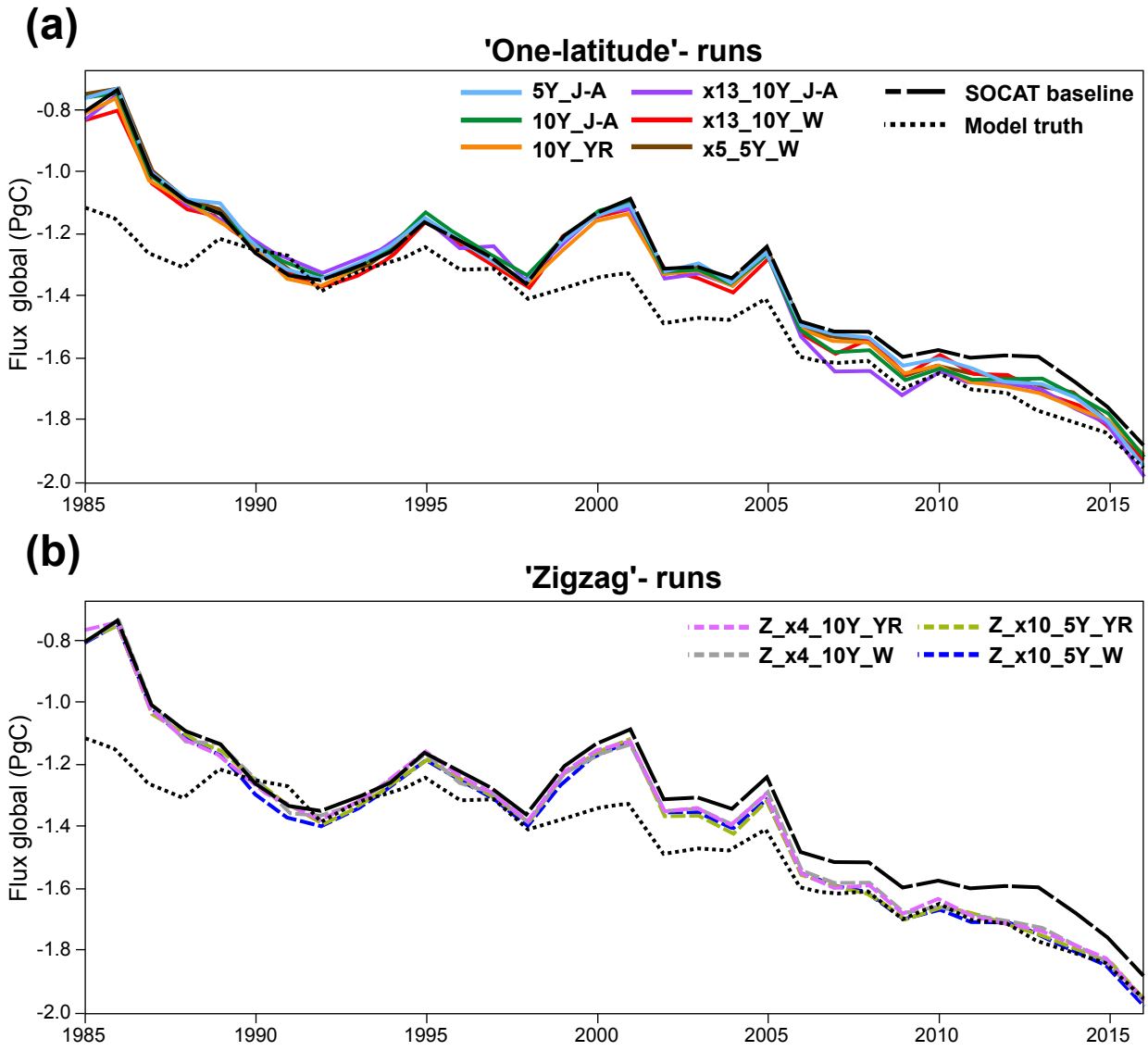


**Figure S19**. Globally annually averaged air-sea $CO_2$ flux for 'one-latitude' (**a**) and 'zigzag' (**b**) runs, including the 'SOCAT-baseline' (black dashed line) and the testbed 'model truth' (black dotted line). Generally, the Saildrone USV additions lead to an increased sink in the Southern Ocean compared to the SOCAT baseline. The 'zigzag' runs (dashed lines) generate a stronger sink compared to the 'one-latitude' runs (solid lines), and closely match the model truth the last ten years of the testbed.

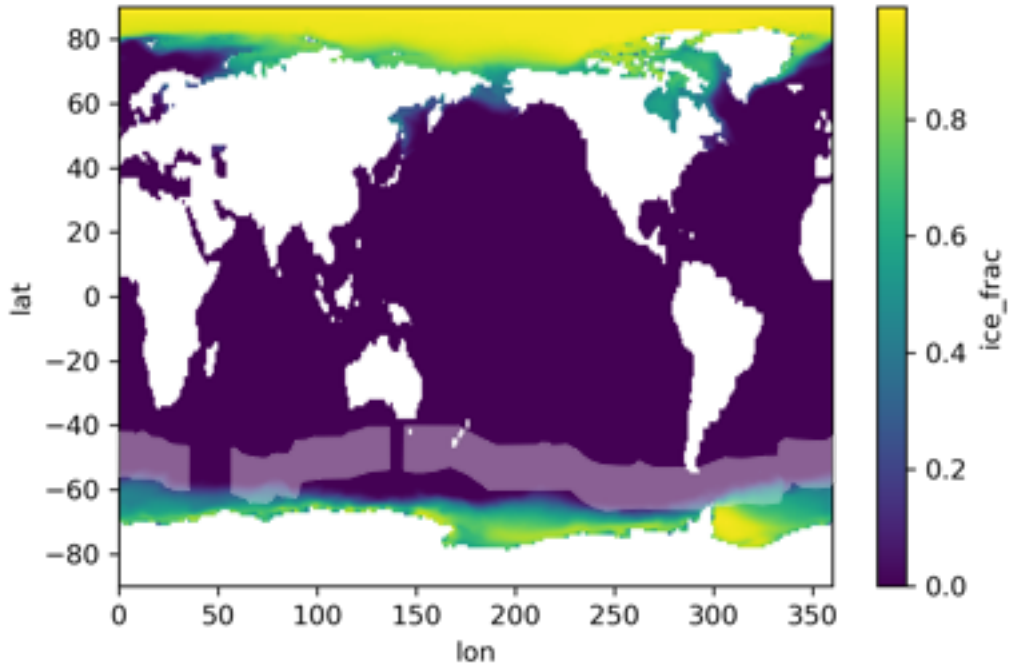


**Figure S20**: Map of the global sea-ice extent as defined by the SeaFlux product (Fay et al., 2021). Light shaded areas represent the spatial extent of Saildrone USV sampling for run 'x13_10Y_J-A' and 'x13_10Y_W', where the 'one-latitude' Saildrone USV track (Sutton et al., 2021) was repeated six times by 1° north and south. The southernmost sampling occurs within the ice-zone as defined by ice fraction ('ice-frac') > 0.

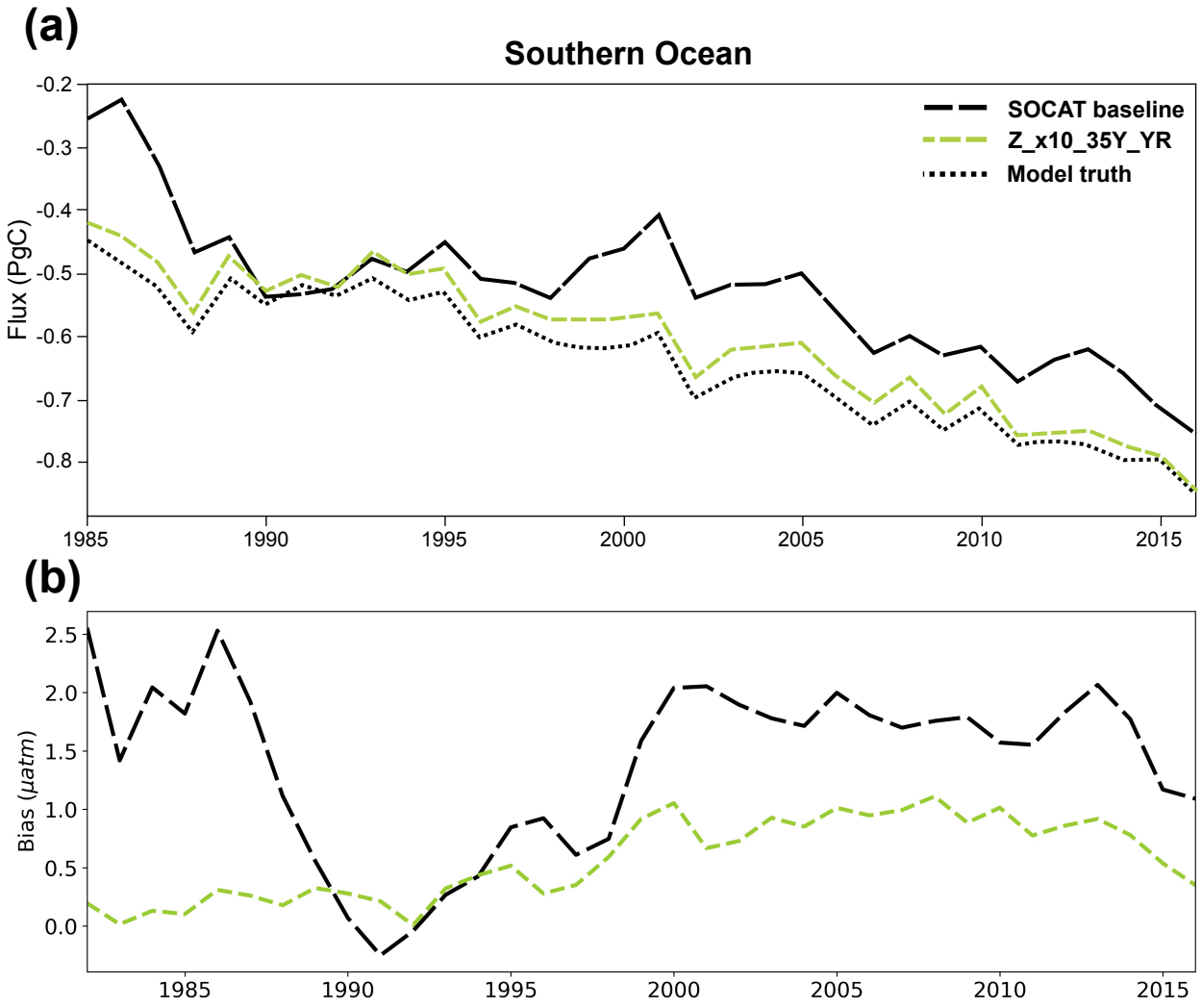


**Figure S21**: Southern Ocean (> 35° S) annually averaged air-sea $CO_2$ flux (**a**) and bias (**b**) for run 'Z_x10_35Y_YR' and the 'SOCAT-baseline'. Run 'Z_x10_35Y_YR' closely matches the 'model truth' (black dotted line) air-sea flux for the entire testbed period (**a**), and shows a significant dampening in bias variability compared to the SOCAT baseline (**b**).

| Mean flux (Pg C/yr) | Model truth | SOCAT baseline |
|---|---|---|
| **Globally** | | |
| 1982-2016 | -1.5 | -1.3 |
| 2006-2016 | -1.7 | -1.6 |
| 2012-2016 | -1.8 | -1.7 |
| **SOUTH (90°S-35°S)** | | |
| 1982-2016 | -0.6 | -0.5 |
| 2006-2016 | -0.8 | -0.7 |
| 2012-2016 | -0.8 | -0.7 |

**Table S1**: Global and Southern Ocean mean air-sea $CO_2$ fluxes (in Pg C yr-1) for the testbed 'model truth' and the SOCAT baseline.

| Run name | 5Y_J-A | | 10Y_J-A | | 10Y_YR | | x13_10Y_J-A | | x13_10Y_W | | x5_5Y_W | | Z_x10_5Y_W | | Z_x4_10Y_W | | Z_x10_5Y_YR | | Z_x10_5Y_W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Saildrone track*** | One-lat | | One-lat | | One-lat | | One-lat | | One-lat | | One-lat | | Zigzag | | Zigzag | | Zigzag | | Zigzag | |
| ***Years of sampling*** | 5 | | 10 | | 10 | | 10 | | 10 | | 5 | | 10 | | 10 | | 5 | | 5 | |
| ***Duration of sampling*** | Jan-Aug | | Jan-Aug | | Shifted | | Jan-Aug | | SO winter | | SO winter | | Year-round | | SO winter | | Year-round | | S winter | |
| ***Additional observations*** | **2,075** | | **4,150** | | **4,150** | | **44,250** | | **25,395** | | **5,022** | | **7,600** | | **2,500** | | **11,400** | | **3,800** | |
| | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** | **% change** | **diff** |
| **Globally** | | | | | | | | | | | | | | | | | | | | |
| **1982-2016** | -1 | 0.01 | -1 | 0.02 | -2 | 0.03 | -2 | 0.03 | -2 | 0.03 | -1 | 0.02 | -3 | 0.04 | -3 | 0.04 | -4 | 0.05 | -4 | 0.06 |
| **2006/2012-2016** | -4 | 0.06 | -3 | 0.06 | -4 | 0.06 | -6 | 0.09 | -3 | 0.06 | -3 | 0.06 | -5 | 0.09 | -5 | 0.08 | -6 | 0.10 | -7 | 0.12 |
| **90°S-35°S** | | | | | | | | | | | | | | | | | | | | |
| **1982-2016** | -2 | 0.01 | -3 | 0.02 | -5 | 0.03 | -6 | 0.03 | -5 | 0.03 | -3 | 0.02 | -8 | 0.04 | -7 | 0.04 | -9 | 0.05 | -11 | 0.06 |
| **2006/2012-2016** | -10 | 0.07 | -9 | 0.06 | -10 | 0.07 | -14 | 0.09 | -9 | 0.06 | -9 | 0.06 | -13 | 0.09 | -13 | 0.08 | -15 | 0.10 | -15 | 0.10 |

**Table S2:** Difference ('diff') between calculated mean air-sea flux of individual Saildrone USV runs from the 'SOCAT-baseline' (in Pg C yr-1), and change shown by %. There is a negative change from the 'SOCAT-baseline', i.e., all Saildrone USV runs predict a stronger sink. 'Additional observations' = 1°x1° monthly Saildrone USV observations (not including SOCAT). 'SO winter' = Southern hemisphere winter months, i.e., June, July, August, and also including September. Testbed period = 1982-2016.