Author replies to Anonymous Referee #2

Overall, I think the authors have done a very good job. The manuscript is clearly written, the structure is logical, and the figures are generally of high quality. With some minor revisions, I believe the paper should be ready for publication. Please see the in-text comments for specific suggestions to improve clarity and to make certain figures more rigorous.

A: We thank the reviewer for the positive and constructive assessment of our manuscript, and the kind words. We fully understand the concerns regarding model robustness and impression of cherry-picking a specific test year and a specific day to illustrate model performance. Below we describe in detail how we plan to address this concerns in the revised version. We also considered the in-text comments and will implement the suggested changes in the revised version.

One point I would like to raise concerns the choice of using a single year (2016) as the test set. The manuscript does not provide a clear justification for this decision. Using only one test year raises the concern that the evaluation may depend on a particularly “lucky” year, or alternatively on a year with atypical behavior. In either case, it becomes difficult to convincingly demonstrate the model's generalisability.

As a reader without specific expertise in the Greenland Ice Sheet (GIS), I am unsure whether 2016 is representative of typical conditions or whether it may have experienced unusual or extreme events. From a scientific robustness perspective, it would strengthen the study to repeat the prediction-versus-observation scatter plots for one or two additional test years. This would allow the reader to assess whether, for example, the melt overestimation by the autoregressive and modular neural network models is a persistent feature or specific to 2016.

If the authors intentionally selected 2016, I would encourage them to provide a clear justification. For example, an appendix figure showing the distribution of SMB over the GIS compared to other years could help demonstrate whether 2016 is representative or exceptional.

A: We appreciate this important comment and want to explain our selection process for the test year. We note that evaluating one additional single year may not necessarily provide stronger evidence of generalisability, since performance on any individual year may itself be influenced by natural interannual variability.

The train/validation/test split was chosen based on several considerations: First, we aimed to use a sufficiently large training dataset to capture a wide range of atmospheric conditions and interannual variability, thereby improving the robustness of parameter estimation. Second, to avoid information leakage between training and testing, we selected the last available year in the dataset as the test set, which is a common strategy for evaluating models using temporally ordered data.

In addition, when selecting the test year (and also the validation year), we checked whether these years were extreme or atypical. The table below shows the total melt (Gt per year), the RMSE, MAE and MBE (in mm w.e. per day) relative to the climatology over 1990-2013 (i.e., the training period), together with the coefficient of determination between the specific year and the climatology. The total melt in the test year (726 Gt) lies well within the historical range of the training climatology (294–1059 Gt) and is comparable to several years in the record (e.g., 1998, 2003, 2005, 2007, 2008). It deviates from the climatological mean by only about 0.75 standard deviations. The deviation metrics between the climatology and the daily melt for 2016 (RMSE = 2.30 mm w.e. per day, MAE = 0.56 mm w.e. per day, MBE = 0.17 mm w.e. per day, R^2 = 0.78) are consistent with values obtained for other years in the record, showing that 2016 is indeed a typical year relative to the climatological reference period.

We will add this table with some explanation in the revised version of our manuscript.

	total melt (Gt)	RMSE	MAE	MBE	R2
1990	576	2.08	0.50	-0.06	0.77
1991	547	2.11	0.52	-0.10	0.74
1992	294	2.48	0.60	-0.48	0.28
1993	570	1.93	0.48	-0.07	0.78
1994	476	1.96	0.49	-0.21	0.74
1995	626	2.50	0.58	0.02	0.71
1996	456	2.15	0.53	-0.24	0.65
1997	519	2.25	0.54	-0.15	0.69
1998	709	2.32	0.55	0.14	0.76
1999	546	2.22	0.54	-0.11	0.72
2000	548	2.46	0.57	-0.10	0.68
2001	502	2.05	0.50	-0.17	0.73
2002	686	2.47	0.58	0.11	0.73
2003	732	2.62	0.59	0.18	0.74

2004	596	2.05	0.51	-0.03	0.77
2005	723	2.50	0.60	0.17	0.74
2006	592	2.49	0.59	-0.04	0.68
2007	740	2.43	0.58	0.19	0.76
2008	723	2.55	0.59	0.16	0.75
2009	571	2.12	0.53	-0.07	0.76
2010	797	2.63	0.61	0.28	0.74
2011	672	2.16	0.51	0.09	0.79
2012	1059	3.61	0.93	0.68	0.65
2013	510	1.94	0.49	-0.16	0.77
2014 (validation)	650	2.22	0.53	0.05	0.77
2015 (gap)	611	2.28	0.57	-0.01	0.73
2016 (test)	726	2.30	0.56	0.17	0.78

Table: total melt in Gt per year, and discrepancy between the daily melt of the specific years relative to the melt climatology 1990-2013 reported as RMSE, MAE, and MBE (in mm w.e. per day), and R2.

A similar concern applies to Figure 5, which focuses on 21 July 2016. While this date is interesting, showing only a single summer day risks giving the impression of a carefully selected ("lucky") example. I would encourage the authors to include additional dates in the appendix, ideally covering different seasons, for example, shoulder seasons or winter periods when little or no melt is expected. This would provide a more comprehensive picture of model behavior across varying surface mass balance regimes.

A: We will provide additional examples in the Supplements.

In addition, it could be helpful to show aggregated diagnostics, such as spatial plots of the mean SMB over several months (or seasonal averages) for the best-performing model. Such analyses would provide stronger evidence that the model captures robust patterns rather than performing well on isolated dates.

A: We will provide additional monthly and seasonal aggregates for the test year, and averaged across the entire dataset in the revised version.

Overall, I consider this a valuable contribution, but addressing these points would further strengthen the manuscript.