Responses to Reviewer #2

I echo the summary and comments of the first Reviewer, so I will not repeat them here. This is a solid analysis, but there are some points that should be addressed.

General comment:

The goal of this work (from my understanding) is to create benchmark snow datasets using an offline model, which is potentially more consistent than reanalysis or coupled model snow products that can be affected by uncertainties and errors related to forcing, data assimilation, model bias, and coupling. The authors assert that offline modeling can "isolate the role of meteorological driving" from these other issues. This is largely true. However, I would caution the authors that the model they are using still has a number of parameters whose values they chose, and which affect the snow output from the model. With a different set of parameters, the model could (or arguably will) provide a different indication of the amount of error introduced by meteorological forcing, because the model dynamics will change. So, it's not entirely possible to disentangle forcing uncertainty from model construction and parameter uncertainty, without an exhaustive analysis of model sensitivity. I would recommend that the authors qualify their statements by noting that this is only one parameter set for this model, and their findings might be different if the parameters in Table 1 were changed.

Thank you for the thoughtful consideration of this paper and the context in which we discussed the results. Studies of snow mass/SWE uncertainty are most frequently done by comparing (fixed versions of) various snow models, including output from reanalyses. Sometimes, this precludes investigation of forcing biases and structural biases caused by model choices.

As noted, analyzing a model's sensitivity to parameter (or process) changes can be and has been done systematically in some cases (e.g. Essery, 2015 or Raleigh et al., 2015). While it is outside of the scope of this study to develop the B-TIM, recent increases in the availability and quality of in situ SWE, snow depth, and snow density information may feed into future development. We will incorporate discussion of this structural/parameter uncertainty into the manuscript.

"Isolat[ing] the role of meteorological driving" is meant in the sense that the offline inter-product differences are not a function of snow modeling differences, while the online inter-product differences are. These products may still be biased relative to ground-truth, which is explored in Fig. 3 and the new supplementary figure AC Fig. 3 included below.

Specific comments:

Lines 66-70: Past studies have attempted to assess the influence of various factors on snow model uncertainty, including forcing, and it would be appropriate to cite one or more here (e.g., Raleigh et al., 2015, https://doi.org/10.5194/hess-19-3153-2015).

Thank you for this idea. The Raleigh et al., 2015 study and others (e.g. Cho et al., 2022, Essery, 2015, Günther et al., 2019, and Menard et al., 2021) have explored these various factors. They find that snow modeling is sensitive to forcing biases and parameter changes, and provide avenues to assessing these sensitivities. We will include some reference to this work around L442. Our method offers a way to decompose or investigate snow biases when it is impossible to run additional simulations or directly interact with the model. This is the case for reanalysis snow, as in this study.

Fig. 1: The 20% of precipitation loss seems quite arbitrary. I realize that this constant derives from the Brown et al. 2003 paper, but there is no reason to assume that this loss rate would be consistent across sites. This parameter could have a strong influence on the magnitude of snow accumulation. Can the authors give some indication of why a constant 20% is the best choice?

This is an important open question. The 20% reduction does derive from previous studies and tuning that was done with respect to in-situ snow data at a limited number of sites. For some time, a varying loss parameter was used in the B-TIM for different snow classes (following Sturm et al., 2010). The 20% reduction was applied to tundra, prairie, and taiga snow-climate zones as the most likely regions where blowing snow and sublimation could dominate. This was simplified by extending the same reduction to the rest of the NH land, as these are the snow-climate zones that take up most of the Northern Hemisphere. This practice has continued due to robust performance of the modeled snow, but regional performance is almost certainly affected.

While it is outside the scope of the current study, the spatial variability of and sensitivity to this precipitation loss factor have not been recently characterized. Given recent increases in the availability and quality of in situ SWE, snow depth, and snow density information, future B-TIM development should revisit this 20% reduction.

Fig. 1: What are delta rho_c and delta rho_w? I do not see them mentioned anywhere else in the text?

Thank you for catching this omission. These two variables represent the change in snowpack density under “cold” and “warm” compaction processes. Equations 7a, 7c, and 8 will be corrected.

$$\Delta\rho_c = C_1\,(\mathrm{SWE}^*)\,\exp[\,C_3\,(T_{melt} - T_{snow})\,]\,\exp[\,-C_2\rho^*\,]\,,\ T < T_{melt} \tag{7a}$$

$$\Delta\rho_w = (\rho_{max} - \rho^*)(1 - e^{-a\Delta t}),\ T \geq -1°C. \tag{7c}$$

$$\rho_f = \rho^* + \Delta\rho_w \text{ if } T \geq -1°C \text{ else } \rho_f = \rho^* + \Delta\rho_c, \tag{8}$$

Lines 279-280: The authors state that ERA5 outperforms JRA-55 and MERRA-2, based on uRMSE and correlation. However, Figure 3g seems to show that the bias is higher for ERA5. Shouldn’t the bias be important here as well? What about raw RMSE (without removing the bias)? I would guess that most users of these datasets are unlikely to unbias them before using them.

The bias is important, but because it is calculated as the sum of differences, positive and negative differences can cancel and yield a small bias. The RMSE measures the average magnitude of the error (weighting larger errors more heavily), so it avoids the cancellation problem. However, RMSE and bias are not independent pieces of information, as any bias that exists contributes to the RMSE. That is why we report uRMSE instead.

$$RMSE = \sqrt{uRMSE^2 + bias^2} \tag{1}$$

Generally, the products considered here have small bias compared to uRMSE, as provided in the table below. Each B-TIM product has a lower RMSE than its reanalysis counterpart and the greatest RMSE arises from the JRA-55 online snow product. A comment about this will be incorporated in the manuscript to supplement Fig. 2.

	BIAS	**URMSE**	**RMSE**
BRE5	-11	32	34
BRJ55	10	33	34
BRM2	8	36	37
ERA5	-9	38	39
JRA55	4	61	61
MERRA2	9	46	47

Line 345: The authors find that the "B-TIM products provide more consistent descriptions of key snowpack climatology metrics". This is true, but consistency does not necessarily mean accuracy. It's possible that one of the reanalyses is a more accurate reflection of reality. The authors could use their in-situ data to evaluate this, but have not yet sufficiently done so in this manuscript.

Thank you for requesting further clarification of these statements. We have done some additional assessments with the in-situ data to supplement these claims (AC Fig. 3).

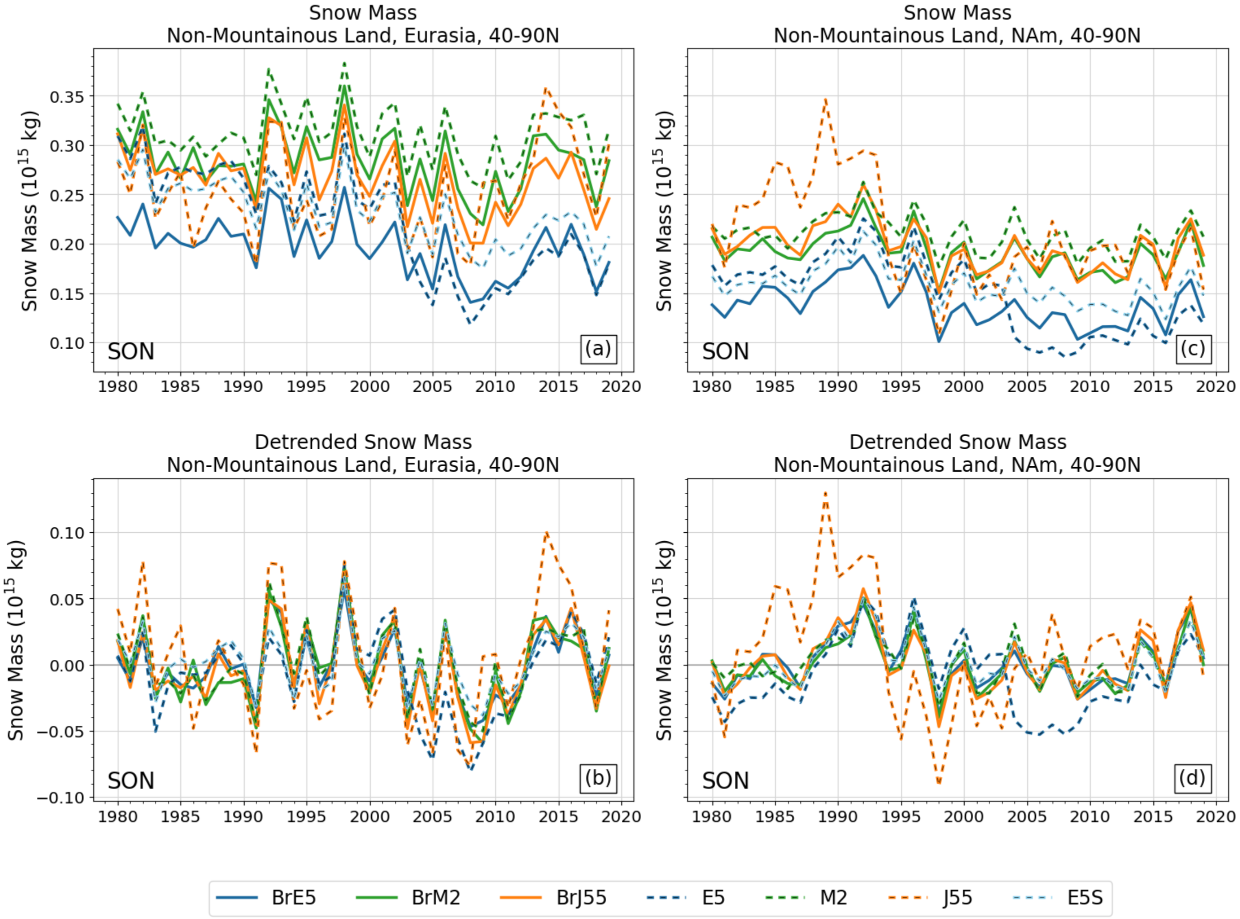
Line 366: Why did the authors not use a more robust trend method, like Theil-Sen slope (which is less influenced than OLS by outliers, and the start and end of time series), for detrending?

The Theil-Sen estimator gives a robust linear regression. As was noted, it is less influenced by outliers and the start/end of time series than the OLS minimization method. To the authors' knowledge, Theil-Sen slope is well defined, but there are several definitions of the y-intercept in the literature. The definition of y-intercept that we used to produce the figure below is the one implemented in the scipy stats Python module.

$$median(y) - TheilSenSlope \times \ median\ (x) \tag{2}$$

Regardless of detrending method, the same qualitative results are seen. All the B-TIM datasets display the same variability and diverge notably from JRA-55 (throughout 1980-2020) and ERA5 (after 2004).

We propose to include: Detrending by another method yields similar results (e.g. using the Theil-Sen estimator, which is robust to outliers and shifts to the start or end of the time series, not shown).

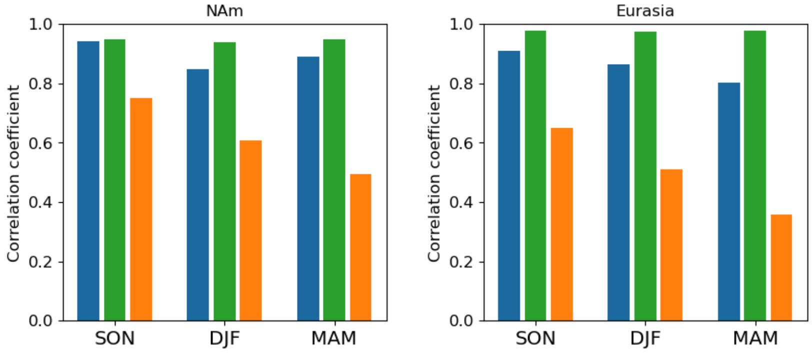


AC Fig. 1: Same as Fig. 7 in the original manuscript, but detrending in panels (b) and (d) was done based on Theil-Sen line fitting.

Fig. 8: Interesting to make point that differences among reanalysis are greater than among B-TIM. Isn't that kind of expected? Wouldn't it also be informative to compare B-TIM vs. reanalysis pairs (same forcing, different models), using more than just correlation (as in Fig. 9, but using bar charts as in figure 8, for example)?

On one hand, it is reasonable to expect that differences among reanalyses are greater than among B-TIM datasets. However, model differences could theoretically be introducing snow biases of opposing sign (e.g. one model with too much melt and another with too little melt could increase or reduce the bias in the snow depending on the overall bias). Therefore, it is important to document the finding in this case.

The suggestion to look at same-forcing pairs is a good one. Spatial correlation is one aspect of their agreement that we show in Fig. 9, and rather than relying on the visual comparison (i.e. looking at the same-colour lines on Figs. 6 and 7), we can include a figure in the supplementary information to address this question. A preliminary version of this figure is below (AC Fig. 2). The pair with JRA forcing (orange) is poorly correlated across all seasons and both continents. It is consistently worse over Eurasia for a given season, and the correlation drops over the snow season. The other pairs have higher correlations, with the MERRA pair being most similar. This lines up with the discussion of Fig. 9 that is based on spatial correlations.



AC Fig. 2: Correlation between SON, DJF, and MAM snow mass time series with the same forcing. BrE5-ERA5 pair in blue, BrM2-MERRA2 pair in green, and BrJ55-JRA55 pair in orange. Values are split for two continents.

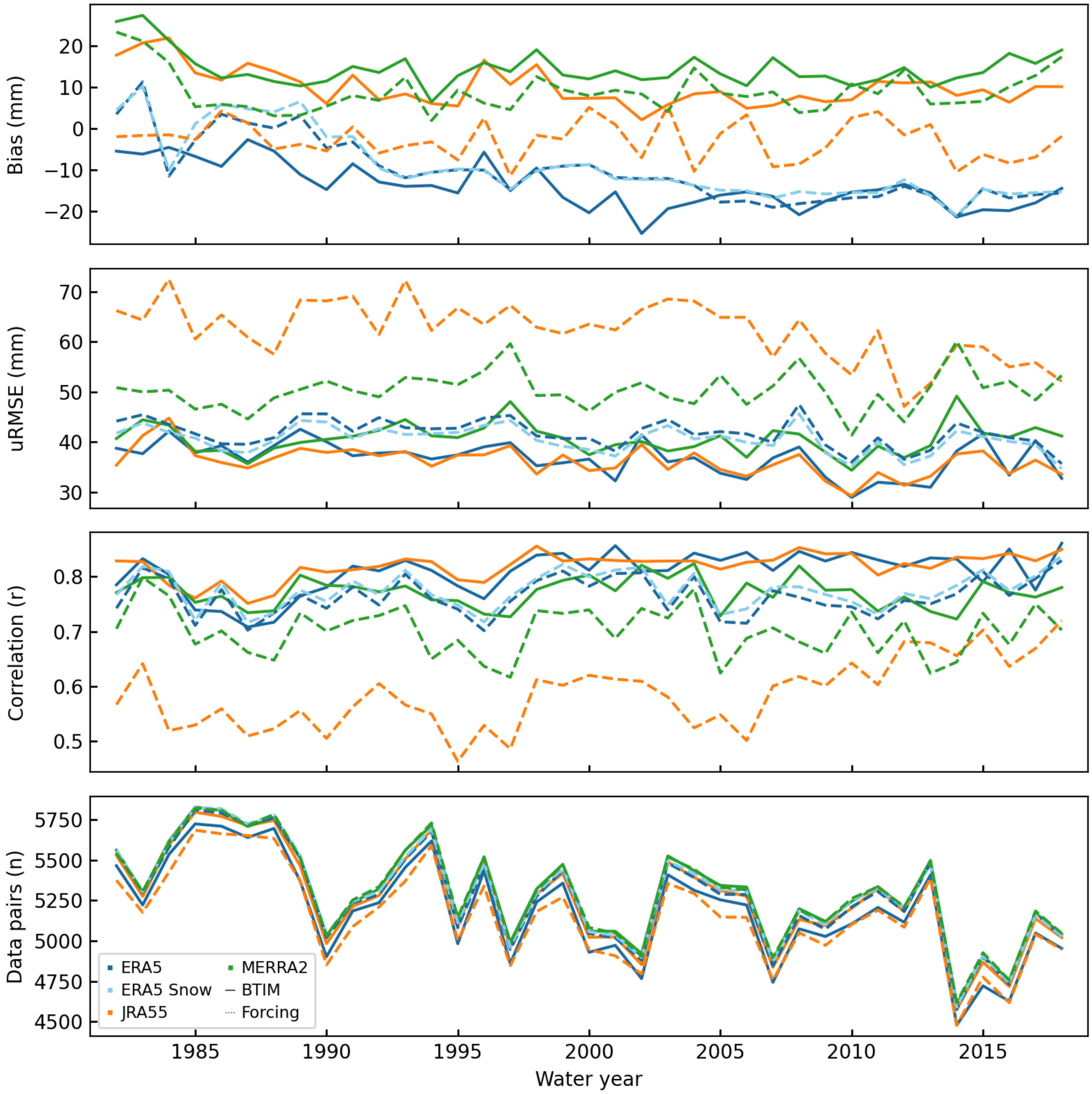
Lines 422-425 (section 4, 3rd bullet): The authors show that the B-TIM model results in "far more consistent interannual variability" than the reanalysis products. However, this does not necessarily mean that the B-TIM interannual variability is more "correct" (i.e. a more accurate representation of the true interannual variability). Can the authors show using their in-situ data that less (or more consistent) interannual variability results in greater accuracy?

Discussion below.

Lines 457-459: Similar comment as above. The authors suggest that there is a "problem" with JRA-55. This is a strong statement to make. It's true that JRA is the least accurate by some metrics and different from the other reanalyses, so it's possible that the authors' suggestion is correct. However, the authors have not shown in this manuscript that the interannual variability of JRA-55 is wrong. Maybe the interannual variability in the other reanalyses is too muted? The authors have in-situ data available to back up their statement, but they have not yet sufficiently done so in the manuscript.

Thank you for the suggestion to fold the in-situ data into the discussion of JRA-55. We have produced some additional analysis covering Dec-Feb that indicates poor performance of JRA-55. AC Fig. 3 indicates that the version of JRA-55 that has less interannual variability (BrJ55, solid orange) also has significantly lower RMSE and higher correlation with the in-situ data than the native JRA-55 (dashed orange). All the products with similar interannual variability have lower RMSE and high correlation with in-situ data.

Finally, further evidence can be found in Mudryk et al. (in discussion). In that comparison of 23 snow products, JRA-55 is consistently among the lowest-ranking of them. This analysis breaks down mountainous, Arctic, and midlatitude regions. A discussion of this can also be included in the discussion.



AC Fig. 3: Time series of validation metrics. DJF measurements only. Dashed lines are native SWE (ERA5, MERRA2, and JRA-55), while solid are B-TIM outputs (BrE5, BrM2, BrJ55).