Response to reviewers' comments
Manuscript number (hess-2023-25)

Note: Below are the reviewers' comments in black, and our responses in blue, all line numbers in our responses refer to the revised version of the manuscript. In the marked version, the changes made are displayed as blue, underlined text.

We thank both reviewers for taking the time to review our manuscript and for the insightful comments. The suggestions and feedback we received are highly valuable and have contributed to a significant improvement. As evident in the revised version, we have performed extensive edits and revisions. Notably, we have added considerably more details regarding the model construction and calibration process, and we have adopted the correct interpretation of green water and the reported economic valuations reflect this change.

Referee#1 comments:

The scientific importance of the article is high. Congratulations to the authors! They covered a very interesting and important topic. As a novelty, they tried to link ecosystem services derived from integrated hydrological model results to monetary evaluation.

Grammatical and technical errors are not typical, the article is of high quality from a stylistic point of view. The written part of the publication is fine.

Thank you for taking the time to review our paper and for your positive evaluation! We appreciate your acknowledgment of the scientific significance of our work, and your compliments on the quality of our writing. This feedback is very valuable to us as we continue to refine our research activities.

I feel it necessary to place some supporting literature references in some places. I marked them in the attached document.

There are parts in the description of the modeling work that are not completely understandable, and it is essential to clarify them. I marked them in the attached document.

After the clarifications and suggested references have been replaced, the article can definitely be recommended for publication.

Thank you for your recommendation. As part of the revision, we have added the suggested literature references and revised the model description section to make it more comprehensive. The specific additions pertaining to this comment are located at lines 164-176:

"It should be noted that the fidelity of the HGS outputs are also dependent on the model scale, with large scale models generally having lower spatial resolution than small scale models as a result of computational constraints, and in some cases, data constraints. For example, a model of a 766,000 km^2 river basin (e.g., Xu et al., 2021a)) is best suited to answer big picture questions (i.e., basin water balance, regional groundwater), while a model built at similar scale to the SNW (e.g., Frey et al., 2021)) can be used to address questions pertaining to localized processes (i.e., individual wetland influences, groundwater recharge and discharge patterns, aquifer conditions, and soil moisture conditions). If even more localized insights are required, HGS models can be constructed for field or plot scale domains (up to approximately 10 km^2), where highly detailed questions pertaining to things such as riparian zones, soil structure, manure application, and tile drainage influences on both water quantity and quality can be evaluated (Fig. 2). Thus, HGS is a scalable and robust model for ecosystem services analysis across a range of different spatial scales and different levels of hydrologic process detail. For the SNW, HGS is used to simulate watershed surface water outflow, transpiration (green water), subsurface water storage, and land surface water storage (reflecting water held in wetlands and reservoirs) using the model construction framework presented in Frey et al. (2021)."

Line 18: You should define what green water means in your article. It can be a bit confusing in this form.

Thank you for this particularly valuable comment. We apologize for any confusion caused by not providing a clear definition of the term "green water". In fact, the term 'green water' is somewhat ambiguous in existing literature, with some authors referring to evapotranspiration as green water (Schyns et al., 2019), while others refer to productive green water solely as transpiration (Casagrande et al., 2021). In the initial version of the manuscript, we based our calculations on the former definition (evapotranspiration). However, in light of your feedback and after conducting a thorough review of the literature, we have come to embrace the view that portion of green water, which supports terrestrial ecosystem services, primarily refers to transpiration, as this is more consistent with hydrologic-focused ES work. Now we clearly define the green water on lines 86-92:

"Transpiration is also called productive green water—the fraction of the rainfall on the land that eventually returns to the atmosphere via plants. It is a source of nutrition for vegetation/ecosystems (Casagrande et al., 2021), and plays a key role in the production of biomass and ecosystem services (Zisopoulou et al., 2022; Schyns et al., 2019). Green water is essential for functioning and growth of ecosystems, and thus supports and maintains terrestrial ecosystem services (Lowe et al., 2022). Hence, transpiration serves as a key driver in providing ecosystem services (Liu and El-Kassaby, 2017), and is a fundamental process by which to model/map terrestrial ecosystem services production."

Line 67: Maybe you should mention the most simplest approaches like matrix models as well.

Thank you for your suggestion. We have now included references to matrix models within the text pertaining to other ES modeling approaches on lines 63-68:

"At present, common modeling tools available for ecosystem services mapping include relatively simple matrix models (i.e., Decsi et al., 2022), and more complex models such as ARtificial Intelligence for Environment & Sustainability (ARIES) (Villa et al., 2021), Co$ting Nature

(Mulligan, 2015), Envision (Bolte, 2022), and Integrated Valuation of Ecosystem Services and Tradeoffs (InVEST) (Natural Capital Project, 2022), with InVEST being by far the most prominent in the scientific literature (Ochoa and Urbina-Cardona, 2017)."

Lines 70-77: You should emphasize the uncertainties of these tools from a hydrologic point of view. There are studies that highlighted their limitations.

Line 86-89: Perhaps the best support for the weakness and unreliability of these models is when they yielded the same results as simple matrix models (without any hydrological calculations). Maybe this article raises your interest: https://doi.org/10.1016/j.ecolind.2022.109143
Maybe you should refer to the model.
This one seems to be appropriate:
https://doi.org/10.1111/j.1745-6584.2011.00882.x

or this:
https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=HydroGeoSphere:+A+Three-Dimensional+Numerical+Model+Describing+Fully-Integrated+Subsurface+and+Surface+Flow+and+Solute+Transport&btnG=

Are there other models which would be capable to handle hydrology similar? E.g.: MIKE SHE..
You may pay attention for this:
https://www.frontiersin.org/articles/10.3389/feart.2021.721009/full#B3

Thank you for this comment. We have now added references to a set of fully-integrated hydrologic model intercomparison studies, wherein models with similar capabilities to HGS are presented and discussed. Please see lines 99-104:

"While a few common hydrological models can weakly capture the subsurface water dynamics and subsurface-surface water interactions (Clark et al., 2015), fully integrated subsurface-surface hydrologic models can dynamically resolve water exchange between groundwater, surface water, and soil moisture, and evaporation and transpiration fluxes with much higher levels of spatial and temporal fidelity. Benchmarking studies have been conducted wherein the most common subsurface-surface hydrologic models have been described in detail, and their simulation behavior compared (Maxwell et al., 2014; Kollet et al., 2016)."

Moreover, we include dedicated text in the Discussion section (Lines 430-461) that specifically centers on the future implications of the study regarding the valuation of ecosystem services.

Lines 430-461:
"While the study herein advances the scientific utility of physics-based fully-integrated groundwater–surface water models, it is essential to acknowledge the inherent uncertainty associated with such an analysis, along with factors that could potentially reduce this uncertainty. It is well known that highly parameterized, structurally complex models can have many degrees of freedom, high data requirements, and non-uniqueness challenges (Beven, 2006). However, the parameterization of physics-based models can also be viewed as a strength due to the constraining

relationship between physically measurable characteristics and parameter values (Ebel and Loague, 2006). For the SNW, soil and subsurface materials are well characterized and hence the spatial distribution and magnitudes of the associated hydraulic parameters are generally well represented in the HGS model. Incorporating meteorological variability into structurally complex model calibration and performance evaluation can also act to reduce uncertainty (Moeck et al., 2018). Because the SNW simulation extended over an 18-year time frame that included multiple droughts and floods, there is confidence that the model structure and parameterization is suited for a wide spectrum of hydrologic conditions, and that the model can dynamically capture transitions from wet-to-dry and dry-to-wet conditions, which is a critical part of the SNW analysis.

It can be posited that physics-based models are best suited for the type of challenge addressed in the work herein because simpler models lack process representation critical within the problem conceptualization (Ebel and Loague, 2006). This can be deemed especially true when considering difficulties associated with quantifying large scale evaporation and transpiration (Stoy et al., 2019), and groundwater–surface water interaction (Barthel and Banzhaf, 2016). Structurally complex models have been shown to perform better than simple models when simulating evapotranspiration (Ghasemizade et al., 2015) and groundwater recharge (Moeck et al., 2018), and previous work by Hwang et al. (2015) demonstrated the utility of HGS for constraining ET at the watershed scale within the same geographic region as the SNW. Further confidence in the SNW HGS model can be established through comparison with other studies. In terms of overall water balance, results from the study herein compare closely with data compiled as part of regional water management study encompassing the SNW (EOWRMS, 2001). Although the study time frames differ (the EOWRMS (2001) study utilized pre-2000 data), the results are similar, with ETa accounting for approximately 45 % and 62 % of annual precipitation in EOWRMS (2001) and the study herein, respectively. While there is limited previous work investigating the partitioning of ETa into transpiration and evaporation that can be directly compared, it is useful to refer to highly detailed analysis based off Fluxnet data (Pastorello et al., 2020) as reference for transpiration and evaporation partitioning in landcover settings representative of those within the SNW. For example, Xue et al. (2023) reported that transpiration as a percentage of ET ranged from 21-56 % and 39-83 % in Fluxnet cropland and mixed forest settings, respectively, whereas the HGS model predicts an aggregate range of 45-65 % across the SNW watershed, which supports the use of HGS transpiration estimates in subsequent ecosystem services valuation."

Line 120-122: Maybe you should refer to- or describe in a nutshell the definition Strahler order to help readers from other disciplines (economy, ecology).

Thank you for the suggestion. We have added detail to this sentence that describes the relative size of the Strahler order values (Lines 124-126).

"The SNW surface water flow network is approximately 6,489 km long and consists of 1,606 km of Strahler order 3+ (relatively large), 1,548 km of Strahler order 2, and 3,335 km of Strahler order 1 (smallest) river and stream features (Fig. 2A)."

Line 153-163: This paragraph should be supported with some reference related to the topic of applicable models on different spatial scales.

Thank you for this suggestion. We have included appropriate references and have modified the first sentence in this paragraph to clarify the connection between model scale and spatial resolution. Please see lines 164-176 in the revised manuscript.

Line 192: Did you carry out some kin of harmonization on input spatial data regarding to their resolution?

In the revised manuscript we describe that land surface and subsurface hydraulic properties were mapped into the HGS model's unstructured FEM using a dominant component approach, meaning that if two or more property classes exist within the input data set for a single element, the majority class is represented. Please see lines 211–214 in the revised manuscript.

Lines 211-214:
"Each landcover category utilizes a unique surface roughness (Manning's n coefficient) value, ranging from 0.001 (urban) to 0.03 s/m1/3 (forest). Land cover properties, as well as subsurface hydraulic properties, were mapped into the HGS model's unstructured FEM using a dominant component approach, such that when two or more property classes exist within the input data set for a single finite element, the majority class is represented."

Line 217-219: How do these stations operate? At what intervals is data recorded?

In the revised manuscript we note that the surface water flow monitoring stations provide daily temporal resolution, while the groundwater monitoring network data provide hourly temporal resolution. Please see lines 231-234:

"The observation data is derived from daily temporal resolution surface water flow monitoring conducted at nine Water Survey of Canada (WSC) hydrometric stations (Figure 4a) and groundwater level data from 10 Provincial Groundwater Monitoring Network wells that was collected hourly but aggregated into daily average values (Figure 4b)."

Line 220-221: Why did you use other metrics to evaluate groundwater performance?
Why did you used this one? Maybe you should take into account other statistical evaluation tools. In the results section we can see that, the coefficient of determination is almost perfect, but the difference between the mean GWLs are significant. Maybe you should bring in the RMSE as well.

We agree that it would be valuable to incorporate additional statistical evaluation tools in our analysis of groundwater performance and now include RMSE values for the simulated vs observed groundwater levels in the revised manuscript.

Lines 234-237: "The Nash-Sutcliffe Efficiency (NSE) and Percent Bias (Pbias) metrics (Moriasi et al., 2007) are used to evaluate surface water flow simulation performance, while the coefficient of determination (R2) and root mean square error (RMSE) is used to evaluate groundwater simulation performance."

Line 266-267: Good accuracy, according to what? You should cite a reference.

In the revised manuscript, we provide the reference (Moriasi et al., 2007) to support our model performance interpretation. Please see lines 285-287 in the revised manuscript .

Line 268-271: What was the temporal resolution of the compared data (daily, weekly, monthly, yearly)? What does the R2 value refer to? The large difference between the observed and modeled average groundwater level depth is worrisome. Especially knowing that in L107-L109 you wrote about a shallow GW depth of 1-3 m. This can also significantly affect the modeled actual evapotranspiration values.

Regarding the temporal resolution of the compared data, the observed groundwater level data was collected at a hourly resolution from the 10 groundwater monitoring wells across the SNW; however, the hourly data was aggregated to daily average prior to being used for model performance evaluation. Both groundwater and surface water simulation performance was evaluated based on daily temporal frequency. In the revised manuscript we will rewrite this section to be more clear in regards to the temporal frequency.

The R^2 value refers to the proportion of the variance in the observed groundwater level that is predicted from the simulated groundwater level. A high R^2 value indicates a good fit between the observed and simulated data.

We acknowledge your concern regarding the large difference between the observed and modeled average groundwater level depth. However, we would actually regard the average difference of 2.8 m between simulated and observed groundwater levels to be very good for two primary reasons. Firstly, the model covers 3830 km^2 and has element edge lengths that vary from ~100 m to 300 m, hence subtle variability in local topography (from which groundwater depths are referenced) are not perfectly captured in the model geometry. Secondly, because groundwater extractions were not represented in the model, simulated groundwater levels are biased higher, and this bias will be most pronounced in groundwater production areas, where the monitoring wells tend to be placed.

Finally, because the majority of soils are clay and silt loam with high capillary suction, 2–3 m fluctuation in water table depth will have little influence on groundwater movement into the plant root zone.

We have added text to the revised manuscript that acknowledges the groundwater simulation performance on lines 287–291:

"Groundwater levels were also reproduced across the SNW with reasonable accuracy for the 2000 to 2017 interval. The R^2 between simulated and observed water levels in the 10 observation wells is 0.98, with the simulated values having a mean value 2.8 m higher than the observed values. Groundwater simulation performance at the individual wells is presented in Table 1.”

Line 366-368: What about the limitations of the fine-grained models? Are they applicable anywhere with any spatial scale? Data needs, other requirements (resource, financial, expert, so on).

We have completely rewritten this section of the discussion. We no longer refer to ‘fine-grained models’ as this is an ambiguous term, and we specifically mention that our results extend only to watershed areas similar to ours (i.e. 4,000 km^2). Please see lines 403-407 in the revised manuscript:

"Incorporating green-blue water resource consideration at the watershed scale helps with the characterization and quantification of the role water plays in land use and terrestrial ecosystem function. Based on the study herein, fully-integrated groundwater – surface water models, such as HGS, have potential to facilitate better management of watershed scale (approximately 4,000 km2) water resources for ecosystem services endpoints, and to evaluate the contributions of terrestrial water storages towards green water supply."

Referee#2:

The paper on Monetizing the role of water in sustaining watershed ecosystem services using a fully integrated subsurface–surface water model by Tariq Aziz et al presents an interesting case study of integrating subsurface–surface water model with valuation of ecosystem services. However, there are few queries about the methodology adopted as well as the results presented and discussed. A line-by-line comment is given below:

We appreciate your positive evaluation of our work. We will address all the queries you have about the methodology and results presented in the revised version of the paper.

Introduction

Page 1, L 24-25: What is the relationship between subsurface water and ecosystem services? Kindly extend on this point in the introduction to provide a clear picture of how subsurface water is linked to ecosystem function and, as a result, the production of ecosystem services. Furthermore, a conceptual diagram connecting subsurface water with various ecosystem services would help readers connect the paper by providing a clear picture.

Thank you for this comment. In the extensively revised manuscript, we have added additional description of how subsurface water links to ecosystem services, and as well, we have reorganized the existing text so that the subsurface water connection is more clear. Please see the revised introduction section.

Methodology

217-219: How the observed data is used to run the model. Did you run the model for all 9 sites for surface water flow calibration, or did you run it in an integrated fashion? This is unclear. Please clarify the same for Groundwater Monitoring Network wells.

We apologize for any confusion regarding the use of observed data to run the model and the approach taken for surface water flow calibration and groundwater monitoring network wells. The model was run continuously for the 2000 to 2017 time interval, with gridded, daily temporal resolution climate forcing as the primary input to the model. All nine surface water flow gauges were concurrently used for calibration, in conjunction with the 10 groundwater monitoring wells. Please see line 287–288 in the revised manuscript.

217-219: It would be better to indicate on what time scale the model is calibrated/validated? Daily, Monthly, Hourly?

We agree that it is important to indicate the time scale used for calibration and validation in our modeling study. To clarify, we used a daily time scale for model calibration and validation. We now make sure to specify this in the revised text to enhance the clarity of our work. Please see lines 230–234 as below:

"The SNW HGS model was run continuously from 2000–2017 with daily temporal resolution climate forcing, and simulation performance is evaluated using observed surface water flow rates and groundwater levels. The observation data is derived from daily temporal resolution surface water flow monitoring conducted at nine Water Survey of Canada (WSC) hydrometric stations (Fig. 4a) and groundwater level data from 10 Provincial Groundwater Monitoring Network wells that was collected hourly but aggregated into daily average values (Fig. 4b)."

219: 221: Is the model validated? if yes, mention years for calibration and validation

Thank you for raising this point. For the purpose of our study, the model was calibrated over the full 2000 to 2017 time interval, and the performance metrics are calculated/reported for the same time interval. Accordingly, the model performance metrics reflect the same time interval over which the ecosystem service analysis is conducted. A formal validation was not conducted as our intention was to optimize model performance for the full 18-year simulation interval.

Results:

The paper makes no mention of the model's performance. For instance, how the model behaved at various gauge stations.

We appreciate your feedback and agree that the model's performance should be clearly presented in the manuscript, and we note that reviewer 1 also raised this point (see above). In the revised manuscript we now present simulated vs. observed streamflow time series graphs for all nine surface water gauging stations, and RMSE statistics for the groundwater simulation performance. Please see the new figure 5 and the new Table 1, along with text in lines 285–291:

“For the 2000 to 2017 simulation interval, the HGS model reproduces surface water flow rates at the nine WSC hydrometric stations across the SNW with good accuracy per the interpretation guidance provided by Moriasi et al. (2007). Based on daily evaluation frequency, NSE at the individual gauge stations ranges from 0.59 to 0.70, with a mean of 0.66; while PBias ranged from -17.4 % to 17.1 %, with a mean of 3.9 % (Fig. 5). Groundwater levels were also reproduced across the SNW with reasonable accuracy for the 2000 to 2017 interval. The R^2 between simulated and observed water levels in the 10 observation wells is 0.98, with the simulated values having a mean value 2.8 m higher than the observed values.”

268-271: Are these value aggregate for all gauge station and observation well?

The model performance metrics presented in the original version of the manuscript were indeed aggregated across all stations. In the revised manuscript the model performance metrics are presented individually for each surface water gauging station and each monitoring well. Please see lines 285-291 as above.

277-280: Check figure 5(a), Can you show the observed and simulated graph of the stream flow? Similarly for surface water storage as well and mentioned the NSE and PbIAS value for each zone/site.

Per previous responses (above), in the revised manuscript we now include observed vs. simulated graphs for stream flow along with the corresponding NSE and Pbias values for gauging site.

277-280: Check figure5 (b), Is the watershed evaporation one of the outputs from the model? What are others? mention either in methodology or results?

Yes, the model outputs include surface evaporation, subsurface evaporation, and subsurface transpiration. We clarify this in the revised version of the manuscript. Please see lines 155–158 in the revised manuscript as below:

“HGS employs a physically based approach to simulate water movement and the partitioning of precipitation input into surface runoff, streamflow, evaporation, transpiration, groundwater recharge, as well as groundwater discharge into surface water bodies like rivers and lakes (Brunner and Simmons, 2012).”

289: Table 1: Is this value calculated or obtained from secondary sources?

The value is calculated using modelling results and values of ecosystem services from 2000-2017 for the SNW. The marginal productivity of water value mentioned in the manuscript is derived from the water production function, which represents the relationship between ecosystem services values and the volume of water consumed in producing them (i.e., transpiration). The slope of this production function gives us the marginal productivity of water value. For SNW, the marginal productivity of water is $0.26/$m^3$.

Discussion:

The discussion section focuses heavily on the results and very little on the validity of the findings. Most important, the authors provide little reflection on uncertainty in their data, models, and underlying assumptions. What does that mean in terms of reliability of the modelled results? The authors should consider where their modeling efforts. shine versus where they fall short, and how the shortcomings can be addressed. I would suggest the authors to discuss the results based on model uncertainty, and future implications of the study in terms of valuation of ecosystem services as well.

We agree with your comment regarding the importance of discussing the validity and reliability of the modeled results in the discussion section. We also appreciate your suggestion to reflect on the uncertainty in the data, models, and underlying assumptions and to discuss the shortcomings of our modeling efforts and how they can be addressed. This comment aligns with a comment made by Reviewer 1 as well. In the revised manuscript, we address these concerns by adding a section related to model limitations and uncertainties associated with our study. Furthermore, we add text to the Discussion section (Lines 430-461) that specifically focuses on the future implications of the study in terms of the valuation of ecosystem services.

Lines 430-461:

"While the study herein advances the scientific utility of physics-based fully-integrated groundwater–surface water models, it is essential to acknowledge the inherent uncertainty associated with such an analysis, along with factors that could potentially reduce this uncertainty. It is well known that highly parameterized, structurally complex models can have many degrees of freedom, high data requirements, and non-uniqueness challenges (Beven, 2006). However, the parameterization of physics-based models can also be viewed as a strength due to the constraining relationship between physically measurable characteristics and parameter values (Ebel and Loague, 2006). For the SNW, soil and subsurface materials are well characterized and hence the spatial distribution and magnitudes of the associated hydraulic parameters are generally well represented in the HGS model. Incorporating meteorological variability into structurally complex model calibration and performance evaluation can also act to reduce uncertainty (Moeck et al., 2018). Because the SNW simulation extended over an 18-year time frame that included multiple droughts and floods, there is confidence that the model structure and parameterization is suited for a wide spectrum of hydrologic conditions, and that the model can dynamically capture transitions from wet-to-dry and dry-to-wet conditions, which is a critical part of the SNW analysis.

It can be posited that physics-based models are best suited for the type of challenge addressed in the work herein because simpler models lack process representation critical within the problem conceptualization (Ebel and Loague, 2006). This can be deemed especially true when considering difficulties associated with quantifying large scale evaporation and transpiration (Stoy et al., 2019), and groundwater–surface water interaction (Barthel and Banzhaf, 2016). Structurally complex models have been shown to perform better than simple models when simulating evapotranspiration (Ghasemizade et al., 2015) and groundwater recharge (Moeck et al., 2018), and previous work by Hwang et al. (2015) demonstrated the utility of HGS for constraining ET at the watershed scale within the same geographic region as the SNW. Further confidence in the SNW HGS model can be established through comparison with other studies. In terms of overall water balance, results from the study herein compare closely with data compiled as part of regional water management study encompassing the SNW (EOWRMS, 2001). Although the study time frames differ (the EOWRMS (2001) study utilized pre-2000 data), the results are similar, with ETa accounting for approximately 45 % and 62 % of annual precipitation in EOWRMS (2001) and the study herein, respectively. While there is limited previous work investigating the partitioning of ETa into transpiration and evaporation that can be directly compared, it is useful to refer to highly detailed analysis based off Fluxnet data (Pastorello et al., 2020) as reference for transpiration and evaporation partitioning in landcover settings representative of those within the SNW. For example, Xue et al. (2023) reported that transpiration as a percentage of ET ranged from 21-56 % and 39-83 % in Fluxnet cropland and mixed forest settings, respectively, whereas the HGS model predicts an aggregate range of 45-65 % across the SNW watershed, which supports the use of HGS transpiration estimates in subsequent ecosystem services valuation."

Conclusion:

The conclusion may be subsequently modified.

Thank you again for your suggestions. We have extensively revised the conclusion section.