WATER RESOURCES RESEARCH, VOL. ???, XXXX, DOI:10.1029/,

Assimilation of Terrestrial Water Storage from GRACE in a Snow-Dominated Basin

B. A. Forman[1,2], R. H. Reichle[1], and M. Rodell[3]

Index Terms: 1847 (Modeling); 1855 (Remote sensing); 1863 (Snow); 3315 (Data Assimilation); GRACE;

B. A. Forman, Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD 20771, USA. (Barton.A.Forman@nasa.gov)

R. H. Reichle, Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD 20771, USA.

M. Rodell, Hydrological Sciences Branch, NASA Goddard Space Flight Center, Code 614.3, Greenbelt, MD 20771, USA.

[1]Global Modeling and Assimilation Office,
NASA Goddard Space Flight Center

[2]Oak Ridge Associated Universities

[3]Hydrological Sciences Branch, NASA
Goddard Space Flight Center

Abstract. Terrestrial water storage (TWS) information derived from Gravity Recovery and Climate Experiment (GRACE) measurements is assimilated into a land surface model over the Mackenzie River basin located in northwest Canada. Assimilation is conducted using an ensemble Kalman smoother (EnKS). Model estimates with and without assimilation are compared against independent observational data sets of snow water equivalent (SWE) and runoff. For SWE, modest improvements in mean difference (MD) and root mean squared difference (RMSD) are achieved as a result of the assimilation. No significant differences in temporal correlations of SWE resulted. Runoff statistics of MD remain relatively unchanged while RMSD statistics, in general, are improved in most of the sub-basins. Temporal correlations are degraded within the most upstream sub-basin, but are, in general, improved at the downstream locations, which are more representative of an integrated basin response. GRACE assimilation using an EnKS offers improvements in hydrologic state/flux estimation, though comparisons with observed runoff would be enhanced by the use of river routing and lake storage routines within the prognostic land surface model. Further, GRACE hydrology products would benefit from the inclusion of better constrained models of post-glacial rebound, which significantly affects ~~estimation~~GRACE estimates of interannual hydrologic variability in the Mackenzie River basin.

1. Introduction

Snow is an important component of the hydrologic cycle that accounts for a large fraction of the available freshwater resources in many parts of the northern hemisphere [*Barnett et al.*, 2005]. Accurate estimation of snow mass, or snow water equivalent (SWE), across space and time using point-scale, ground-based techniques is a difficult task. Therefore, in an effort to better quantify this potential freshwater supply, many researchers have turned to remote sensing estimates derived from space-based instrumentation used in conjunction with land surface models.

Despite recent popularity in the utilization of passive microwave and visible spectrum imagery for the purpose of snow pack estimation (e.g., *Andreadis and Lettenmaier* [2006]; *Durand and Margulis* [2006]; *Dong et al.* [2007]; *Su et al.* [2008]), satellite-derived measurement techniques possess significant limitations. Passive microwave estimates, for example, are prone to large errors for snow packs that are either wet, deep (> 1 m), or contain ice and/or depth hoar layers [*Clifford*, 2010]. Similarly, visible imagery often provides little information outside of the initial accumulation and final ablation periods of the snow season [*Clark et al.*, 2006].

An alternative to passive microwave and visible spectrum-based SWE estimation is the use of gravimetry. Gravimetric techniques focus on the measurement of gravitational anomalies associated with the accumulation (or loss) of mass near the Earth's surface. In the context of snow, changes in the Earth's gravitational field are associated with the accumulation of snow during the snow season and the subsequent ablation and runoff of the snow mass during the melt season. Gravimetry is capable of capturing snow mass throughout the accumulation season, including peak accumulation when SWE information is most valuable to water resource managers. Unfortunately, the drawback of space-based gravimetry is its coarse spatial (∼hundreds of km) and temporal (∼monthly) resolution that limits its applicability for smaller domains. When satellite gravimetric measurements are combined with a land surface model as part of a data assimilation (DA) framework, however, there is the potential to effectively downscale gravimetric estimates in time and space while simultaneously providing useful information content when passive microwave and visible spectrum measurements cannot.

2. Background

One such satellite gravimetry mission is the Gravity Recovery and Climate Experiment (GRACE). GRACE provides approximately monthly estimates of variations in terrestrial water storage (TWS), which includes snow, ice, surface water, soil moisture, and groundwater. The mission is a major step towards understanding regional TWS dynamics [*Tang et al.*, 2010] and offers significant insight into regional- and continental-scale hydrologic processes [*Syed et al.*, 2009; *Rodell et al.*, 2009].

Relatively few studies have been conducted that utilize GRACE measurements within a DA framework. The first study by *Zaitchik et al.* [2008] assimilated GRACE information into a land surface model of the Mississippi River basin. When compared against in-situ groundwater observations, *Zaitchik et al.* [2008] found reduced errors and increased temporal correlations as a result of the assimilation. Further, the results suggested the potential to downscale the coarse-scale GRACE measurements via use of a relatively fine-scale land surface model. However, due to the fact that snow contributes little to TWS in the Mississippi River basin, there was limited opportunity to study the impact of GRACE data assimilation on snow pack characterization.

More recently, *Su et al.* [2010] studied the impact of GRACE data assimilation on TWS estimates in North America for the express purpose of improved snow pack estimation. They found that GRACE assimilation improved SWE estimation in many of the North American basins where snowfall is a significant contributor to the hydrologic cycle. However, *Su et al.* [2010] also found that many issues remain to be addressed, including: 1) the cause of model degradation in some high-latitude basins as a result of GRACE assimilation, 2) the impact of GRACE observational error on DA results, and 3) the impact of GRACE assimilation on components of TWS other than snow.

This study expands on the work by *Zaitchik et al.* [2008] and *Su et al.* [2010] via extended examination of GRACE DA performance within a snow-dominated hydrologic basin. Namely, additional verification activities using independent, ground-based data sets are explored, a number of different GRACE products are tested during assimilation, the impact of GRACE measurement error on DA results is investigated, an analysis of DA innovation sequences is included, and a longer period of record is utilized, which allows for a better assessment of inter-annual variability.

The following sections introduce the methods used in the assimilation framework (section 3), highlight the study domain (section 4), discuss the GRACE measurements and forward model used during the assimilation (section 5), highlight the independent data sets used for ~~model verification~~validation (section 6), present ~~model~~assimilation results (section 7), and conclude with summarized findings and implications (section 8).

3. Data Assimilation Framework

A DA framework is an effective means of merging model estimates with measurements that often yields an improved estimate beyond that of the model or measurements alone [*McLaughlin*, 2002]. Not only does DA provide a conditioned estimate that accounts for both model and measurement uncertainty, but it offers the potential to effectively downscale the measurements in space and time via utilization of the finer-scale information associated with the prognostic model formulation, its parameters, and its forcing data [*Reichle et al.*, 2001; *Zaitchik et al.*, 2008].

Selection of the most appropriate DA system depends on feasibility, robustness, and computational efficiency. In that regard, we choose to employ an Ensemble Kalman Smoother (EnKS) in part because of its ability to handle non-linear models coupled with its flexible, modular structure [*Dunne and Entekhabi*, 2006] as well as the ability to leverage *Zaitchik et al.* [2008] as a precursor study. In general, an EnKS has two basic components: 1) a physically-based, forward model to propagate the model states as an ensemble in order to provide background error covariances, and 2) an update scheme that combines the model states and the satellite measurements in a way that accounts for their respective uncertainties. The work conducted in this current study adapts the EnKS presented in *Zaitchik et al.* [2008] for a snow-dominated basin thereby contributing to the methodological development of GRACE DA (see section 5.3). The EnKS is first introduced below whereas the assimilated measurements and forward model are discussed in section 5.

3.1. Ensemble Kalman Smoother

The prior (unconditioned) estimate of the model states, $\cancel{\mathbf{x}_t^{i-}}\mathbf{x}_\tau^{i-}$, is derived from a prognostic land surface model. This is illustrated in the left-hand side (i.e., Step 1) of Figure 1. The nonlinear model, $\mathcal{F}_t(\cdot)$, propagates the posterior (conditioned) model states, $\cancel{\mathbf{x}_{t-1}^{i+}}\mathbf{x}_{\tau-1}^{i+}$, forward in time, t, from ~~$t-1$ to t~~one month to the next (i.e., from $\tau - 1$ to τ) using an ensemble of N realizations with prescribed model errors $\mathbf{w}_t^i$ as

$$\mathbf{x}_\tau^{i-} = \mathcal{F}_t\left(\mathbf{x}_{\tau-1}^{i+}, \mathbf{w}_t^i\right) \text{ for } i \in N. \tag{1}$$

We adopt the convention where bold lowercase symbols denote vectors, bold uppercase symbols denote matrices, non-bold symbols denote scalars, and calligraphic symbols represent operators. Uncertainties in the model are defined by the model error term, $\mathbf{w}_t^i$ with covariance $\mathbf{Q}_t$. In the ensemble framework, model errors are represented by perturbations that are applied to model states and forcings (section 5).

Next, the prior model states are updated using the observations available for the time period of interest $\tau \in [t_o, t_f]$ (where t_o and t_f are the beginning and end of the assimilation ~~period~~window, i.e., first and last day of the month in this application). This is illustrated in the right-hand side (i.e., Step 2) of Figure 1. The following linear update equation is employed as

$$\mathbf{x}_\tau^{i+} = \mathbf{x}_\tau^{i-} + \mathbf{K}_\tau\left[\mathbf{y}_\tau + \mathbf{v}^i - \mathbf{H}\mathbf{x}_\tau^{i-}\right], \tag{2}$$

where $\mathbf{K}_\tau$ is the Kalman gain matrix, $\mathbf{y}_\tau$ is the measurement vector, and $\mathbf{H}$ is the predicted measurement model that linearly maps the model states into measurement space. Random perturbations, $\mathbf{v}^i$, representing measurement error are added to the measurement vector [*Burgers et al.*, 1998]. The Kalman gain, $\mathbf{K}_\tau$, is a weighted average between the uncertainty of the prior states and the measurements such that

$$\mathbf{K}_\tau = \mathbf{P}_\tau^- \mathbf{H}_\tau^T \left(\mathbf{H}_\tau \mathbf{P}_\tau^- \mathbf{H}_\tau^T + \mathbf{R} \right)^{-1}, \qquad (3)$$

where $\mathbf{P}_\tau^-$ is the background error covariance computed from $\mathbf{x}_\tau^{i-}$ for $i \in [1N]$, and $\mathbf{R}$ is the measurement error covariance. The analysis increments, $\mathbf{x}_\tau^+ - \mathbf{x}_\tau^-$, are applied evenly over each day of the month as illustrated in Step 2 of Figure 1. The update procedure ignores non-Gaussian characteristics and relies only on the first two moments of the distribution. In practice, however, it may only be feasible to accurately compute the first and second moments of the system state [*Khare et al.*, 2008]. Additional details regarding the EnKS update procedure applied in Equation (2) are found in Figure 5 of *Zaitchik et al.* [2008] as well as in section 5.3 further below.

4. Study Domain

The study domain used here is the Mackenzie River basin (MRB) located in northwestern Canada (Figure 2) and consists of 4 individual sub-basins. Sub-basin delineation was based on topographic control and adhered to the topology of the river network. Each sub-basin was extracted from the original GRACE product in order to produce sub-basin-averaged TWS estimates. The smallest sub-basin is 280,000 km^2, which is larger than the minimum area of roughly 150,000 km^2 that can be resolved by GRACE at mid-latitudes [*Rowlands et al.*, 2005; *Swenson et al.*, 2006]. Additional details regarding the GRACE measurements and measurement preprocessing activities are found in section 5.1 and section 5.2, respectively.

As a whole, MRB is $\sim 1.8\times10^6$ km^2 in drainage area ($\sim 1.6\times10^6$ km^2 for land areas only; see Table 1) with the main branch of the Mackenzie River running from the highlands in the southwestern corner of the domain northward toward the Arctic Ocean. The snow classification scheme of *Sturm et al.* [2010] suggests MRB snow type is dominated by taiga-type snow with smaller areas of tundra- and alpine-type snow found in the northwest and southern regions, respectively (see Figure 2b).

5. Assimilated Measurements and Forward Model

5.1. GRACE Measurement~~s~~ Background

Several different GRACE hydrology products were investigated in this study. TWS anomalies from 1) the Space Geodesy Research Group (GRGS) product [*Bruinsma et al.*, 2010; *Horwath et al.*, 2011], 2) the Tellus product available from the NASA Jet Propulsion Laboratory (Tellus) [*Wahr et al.*, 2004; *Swenson and Wahr*, 2006], and 3) the mass concentration product generated at the NASA Goddard Space Flight Center (MasCon) [*Rowlands et al.*, 2005, 2010]. Each product utilizes the same Level 1 range-rate measurements from GRACE, but is processed in a different manner in order to yield mass change estimates in terms of equivalent water thickness.

Each product is available as gridded TWS anomalies (i.e., deviations from the temporal mean at each location). The GRGS and Tellus products are provided on a $\sim 1^\circ \times 1^\circ$ grid whereas the MasCon product is provided on a $\sim 4^\circ \times 4^\circ$ grid. Each product was subsequently converted into sub-basin-averaged total TWS values by adding the location-specific, ~~temporal mean~~long-term average TWS from the land surface model. More information on GRACE measurement preprocessing is provided in section 5.2 and the land surface model is provided in section 5.3.

5.2. GRACE Measurement Preprocessing

Conversion of the GRACE TWS anomalies into sub-basin-averaged TWS estimates that are compatible with modeled TWS values begins with generating a single-replicate of the forward model for the period 1 September 2002 to 1 September 2009. No model errors are prescribed in this simulation unlike that shown in Equation (1). ~~Temporally~~Long-term (i.e., 2002-2009) averaged , sub-basin-averaged estimates of TWS derived from the forward model are subsequently added to the sub-basin-averaged monthly GRACE TWS anomalies, which yields monthly estimates of TWS for each modeled sub-basin that are eventually assimilated using Equation (2). Additional details on the utilization of the GRACE measurements in Equation (2) are found in *Zaitchik et al.* [2008].

One notable aspect of GRACE preprocessing is the consideration of a secular trend associated with post-glacial rebound (PGR). The Tellus product accounts for PGR using the methods of *Paulson et al.* [2007]. However, the GRGS and MasCon products do not account for PGR. Therefore, model output from *Paulson et al.* [2007] is applied here to the GRGS and MasCon products in a similar manner as done for the Tellus product. Preliminary DA results suggest PGR is overestimated by the model of *Paulson et al.* [2007] in both the Slave and Peace+Athabasca sub-basins, but this cannot be verified as the exact amount of PGR in these regions is unknown. Unfortunately, PGR models are difficult to validate due to a lack of independent data, thus the errors are not well quantified. Therefore, in an effort to better understand the impacts of PGR estimates on GRACE DA performance within the MRB, two different versions of each GRACE product were used in the DA experiments: 1) PGR correction applied using *Paulson et al.* [2007] and 2) PGR correction *not* applied (i.e., PGR correction was removed from the

Tellus product). These two approaches effectively bound the extent of PGR impacts on GRACE DA performance.

Finally, one requirement for optimal data assimilation is an accurate representation of measurement error. Given the multiple sources of error present within the GRACE measurements [*Bruinsma et al.*, 2010; *Horwath et al.*, 2011; *Rowlands et al.*, 2005; *Swenson and Wahr*, 2006; *Wahr et al.*, 2006], this task is not trivial. GRACE TWS errors arise from a combination of measurement errors, processing errors, and errors in the geophysical models used to de-alias the GRACE measurements [*Wahr et al.*, 2004]. The error estimates used in this study ~~generally agree with~~are based on those of *Swenson and Wahr* [2006] and *Swenson* [In Prep.], and are comparable to those used in *Zaitchik et al.* [2008]. Even though the spatially-distributed error estimates provided in *Swenson* [In Prep.] are only for the Tellus product, we believe they are fairly representative of the measurement error in all the GRACE products since each product utilizes the same Level 1 range-rate measurements. The time-invariant GRACE measurement error used in this study is less than that used in *Zaitchik et al.* [2008] due to the increased number of satellite overpasses near the poles. The measurement error covariance for each sub-basin of interest is provided in Table 1. The impact of measurement error covariance on DA performance is further discussed in section 7.4.

5.3. Catchment Land Surface Model

The prognostic model used in this application is the Catchment Land Surface Model (Catchment) developed by *Koster et al.* [2000]. Catchment employs a catchment deficit prognostic variable rather than the more commonly-used soil water content variable to estimate subsurface water storage, and explicitly models sub-grid scale soil moisture variability and its effect on hydrological processes such as runoff and evaporation. Further, the inclusion of a three-layer snow model [*Stieglitz et al.*, 2001] provides additional capability in the estimation of terrestrial water storage in areas where snow is a significant contributor to the hydrologic cycle. These attributes create a unique capability for Catchment in the assimilation of terrestrial water storage data assimilation.

The predicted measurement model, **H**, (see Equation (2)) maps the Catchment model states into the GRACE measurement space. **H** not only spatially aggregates the model states into the 4 sub-basins as described in section 5.1, but it also integrates the model states to yield a vertically integrated estimate of TWS. Catchment-based estimates of TWS include changes in the unconfined water table, root-zone soil moisture, surface soil moisture, SWE, and canopy interception. A schematic of Catchment-derived TWS is shown in Figure 3. Catchment-derived TWS was computed in a similar manner as done in *Zaitchik et al.* [2008] except with the additional consideration of canopy interception. Even though lake water storage can be a significant storage component of TWS, Catchment does not account for mass changes within surface water impoundments.

The ~~Goddard Earth Observing System Version 5.2.0 (GEOS-5)~~ Modern Era Retrospective-Analysis for Research and Application (MERRA) product [*Rienecker et al.*, 2011]~~, of which Catchment is a part,~~ was used to force the land surface model. MERRA is provided at an hourly temporal resolution and a $1/2° \times 2/3°$ (latitude/longitude) spatial resolution. An alternative forcing data set by *Reichle et al.* [2011] was investigated for use, which is the same as MERRA except that the precipitation estimates have been corrected towards estimates from the Global Precipitation Climatology Project (GPCP) [*Huffman et al.*, 1997] ~~through rescaling of the MERRA precipitation such that the total amount of precipitation matched that found in the original GPCP~~. No significant difference in the performance of the DA experiments was found between the two forcing data sets. Therefore, only the results utilizing the MERRA forcing are presented here.

Perturbations to specified model states and forcings were prescribed in order to adequately represent model error. Both multiplicative and additive perturbations were utilized as listed in Table 2. Model state perturbations were applied every 20 minutes (i.e., at each model time step) whereas model forcing perturbations were applied every 60 minutes (i.e., at each forcing time step). Temporal correlations were imposed using a first-order auto-regressive model (AR(1)) within the perturbed fields as discussed in *Reichle et al.* [2008]. Following the work of *Reichle and Koster* [2003], a horizontal error correlation length of 2.0° was applied. The root zone soil moisture excess prognostic variable was not perturbed to avoid the introduction of unwanted bias in the subsurface. Cross-correlations between perturbations were included in an analogous manner as conducted in *Reichle et al.* [2007].

To better manage perturbations made to the Catchment ensemble, a number of modifications were made to the DA framework from that originally used in *Zaitchik et al.* [2008]. Perturbations applied to the 3 snow layers were only applied to SWE and not to snow depth or snow heat content. Perturbed snow depth was subsequently recomputed as the perturbed SWE divided by the unperturbed snow density. Snow heat content was also recomputed such that the perturbed SWE yielded the same snow pack temperature as the unperturbed SWE. This was done to ensure physical consistency within the snow pack associated with the prescribed SWE perturbations. In addition, perturbations to the catchment deficit (subsurface) were modified based on the presence of snow in conjunction with frozen soil conditions. More specifically, if snow is present and the surface (~5 cm) soil temperature is below freezing, perturbations are applied to the SWE states only; if the surface soil temperature is at or above freezing, perturbations are applied to the SWE states as well as the catchment deficit. Conversely, if snow is absent and the surface soil temperature is below freezing, perturbations applied to the catchment deficit state were scaled by a factor <1 in order to mimic the attenuated soil moisture dynamics associated with reduced soil permeability; if the surface soil temperature is at or above freezing, perturbations were applied normally to the catchment deficit. Collectively, the changes better maintain physical consistency within the snow pack while better simulating an attenuated soil moisture response when frozen soil conditions persist.

Model spin-up and initialization consisted of a two-step approach. The first step involved a repeated, one-year (i.e., May 2001 to May 2002) cycle of a single replicate without model perturbations for ten years to yield a reasonable estimate of TWS. The second step involved running the model as an open-loop (OL) ensemble from May 2002 to September 2002 in order to yield a reasonable estimate of cross-correlations between different model states as well as to produce an adequate amount of uncertainty (spread) within the OL ensemble. From September 2002 to September 2009, the model was run in either OL mode or with GRACE DA enabled. Finally, an ensemble size of 16 was used based on the convergence of the TWS standard deviation of the prior ensemble. Ensemble sizes greater than 16 showed no significant change in ensemble standard deviation, hence it was determined that 16 replicates was sufficiently large.

6. Validation Approach

A variety of observational data sets were used to evaluate the GRACE DA output. However, due to the observation sparsity within the MRB, particularly in the northern regions, not all pertinent model states could be verified. Most notable amongst the observational data gap is a lack of groundwater and soil moisture measurements. Despite the lack of some observational types, a series of modeled and measured estimates are available that provide a reasonable assessment of the MRB hydrologic response as a function of space and time.

6.1. CMC Daily Snow Analysis Product

Snow observations were based on the Canadian Meteorological Centre (CMC) daily snow depth product [*Brasnett*, 1999; *Brown and Brasnett*, 2010] obtained via ftp server at `sidads.colorado.edu` . The CMC product yields snow depth estimates throughout the northern hemisphere at a horizontal resolution of ∼24 km for the period of March 1998 to the present, and is often considered the best available snow product for evaluating model output [*Su et al.*, 2010]. It is based on optimal interpolation of in situ daily snow depth observations and aviation reports with a first-guess field generated from a simple snow model driven by analyzed temperatures and forecast precipitation from the Canadian forecast model [*Brasnett*, 1999]. SWE estimates were derived from the CMC daily snow depth estimate in conjunction with the climatological snow density parameterization of *Sturm et al.* [2010] as a function of snow depth, day of year, and snow class (Figure 2b).

6.2. INAC Snow Surveys

An additional set of ground-based observations was made available by the Indian and Northern Affairs Council (INAC). This observational dataset consists of snow surveys at

42 different locations, predominantly within the Slave Basin (Figure 2b). Each survey consisted of snow depth and snow water equivalent measurements at ∼10 different points that were then averaged together to yield a single survey estimate at each of the 42 different survey locations. In general, surveys were conducted annually when the snow pack reached peak accumulation. Therefore, these ground-based observations are only available once per year and only within a small portion of the MRB. Between the CMC measurement product and the INAC observational dataset, however, a reasonable comparison of SWE estimates may be conducted over the entire MRB domain throughout the course of the snow season with particular emphasis placed on peak accumulation.

6.3. GRDC Runoff Observations

Runoff estimates were provided by The Global Runoff Data Center (GRDC) via `http://www.bafg.de/GRDC/EN/Home/homepage__node.html`. GRDC estimates are available at hundreds of locations within the MRB at a daily timescale. However, only a handful of stations were selected based on a minimum upland drainage area of ≥250,000 km^2 and a minimum of six (6) years of measurements (Figure 2a). Daily estimates were subsequently aggregated to a monthly timescale for comparison against the DA results utilizing monthly GRACE observations. Table 3 lists the stations used in this study along with the approximate sub-basin aggregation (in terms of integrated upland area) in accordance with the sub-basins shown in Figure 2a. GRDC discharge estimates in the MRB are, in general, based on measurements of river stage height, which were then converted into volumetric flux using assumptions of river cross-sectional area and flow velocity. During the winter time when ice floes are common in the MRB, river discharge measurement error likely increases.

6.4. Validation Metrics

Using the independent, ground-based observations described above, a number of validation metrics were computed. Mean difference (MD) was computed as $MD = \frac{1}{T}\sum_{t=1}^{T}(M_t - O_t)$ where M_t is the modeled ensemble mean and O_t is the ground-based observation, respectively, at time t and where T is the total number of time steps. Similarly, root mean squared difference (RMSD) was computed as $RMSD = \frac{1}{T}\sqrt{\sum_{t=1}^{T}(M_t - O_t)^2}$. Finally, the anomaly correlation coefficient (R) was computed by first determining the climatological seasonal cycle over the course of the simulation period, then the anomaly time series is computed by subtracting the climatological seasonal cycle from the original time series, and finally the anomaly R is computed as the correlation coefficient between the modeled ensemble mean anomalies and the corresponding anomalies of the ground-based observations. For all 3 metrics, the modeled values are obtained from either the open-loop (OL) or data assimilation (DA) simulations. In addition, only times and locations with values $M_t > 0$ or $O_t > 0$ were used in the computation. That is, coincident zeros were excluded (e.g. omitting summertime values when no snow is present in both the model and observations).

Statistical significance of R is determined using the Hotelling-Williams Test, which investigates the equality of two dependent correlations [*Steiger*, 1980]. In this study, the dependent correlations are between: 1) the ground-based observations and the OL results (R_{12}), and 2) the ground-based observations and the DA results (R_{13}). It begins with the hypothesis that the two dependent correlations are equal (i.e., $H_o : R_{12} = R_{13}$). Next, a t-statistic is computed as

$$t_{N-3} \sim (R_{12} - R_{13}) \sqrt{\frac{(N-1)(1+R_{23})}{2\frac{N-1}{N-3}|R| + \overline{R}^2 (1-R_{23})^3}}, \tag{4}$$

where N is the approximate number of degrees of freedom, $\overline{R} = \frac{R_{12}+R_{13}}{2}$, R_{23} is the correlation between the OL and DA results, and $|R| = 1 - R_{12}^2 - R_{13}^2 - R_{23}^2 + 2R_{12}R_{13}R_{23}$. If the computed t-statistic is greater than the corresponding Student t-statistic for a given N at a given confidence level, then one can reject the null hypothesis, H_o, and in turn say that the computed correlation coefficients are statistically different. It is important to note that the t-statistic computed here is only an approximation and likely overestimates the value because of the presence of serial error correlations, which imply that the actual number of degrees of freedom is less than the number of data points.

7. Results and Discussion

7.1. Terrestrial Water Storage (TWS)

Comparison of model results begins with a comparison against the assimilated GRACE TWS measurements used during the conditioning phase. Theory predicts that if information transfer from the GRACE observations into the model estimates takes place during conditioning, then a better agreement between the conditioned estimates and the GRACE observations should occur. If not, the lack of change is either due to a near-zero covariance structure in **K** or is due to close agreement between the GRACE TWS and the model-predicted TWS.

Figure 4 shows the ensemble OL and DA simulations relative to the GRGS (without PGR correction) GRACE TWS observations for the 4 assimilated sub-basins along with the MRB as a whole. The dark gray and light gray regions represent the range of the OL and DA ensembles, respectively. The GRACE observations are shown as solid, black dots with the error bars representing the time-invariant standard deviation of the observation error. The thick dashed and solid lines represent the ensemble means for the OL and DA ensembles, respectively.

In general, there is good agreement between the OL ensemble mean and the GRACE measurements with the exception of the Slave basin during 2002-2004. When DA is enabled, the ensemble mean moves toward the GRACE observations as a result of conditioning. The presence of positive, non-zero covariances in **K** coupled with differences between the GRACE observations and the model-based TWS estimates allows for a significant correction in the DA ensemble toward the GRACE observations. However, it should also be noted that significant differences exist between the model estimates (OL and DA) and the GRACE observations near the annual minimum of TWS. This is in part due to a bias in the variability between the OL model and the observations. That is, the Catchment model has a tendency to "dry out" beyond what the GRACE measurements would suggest. As is discussed in more detail in section 7.3 and 8, a lack of hydraulic routing and lake storage routines in Catchment leads to a more rapid hydrologic response, which results in a more variable (i.e., larger dynamic range) estimate of TWS. Assimilation of the GRACE measurements serves to constrain some of this variability. In addition, when the snow melts and subsequently runs off, the model-derived background error variance is smaller (due to a lack of snow and snow errors) than the prescribed measurement error variance, which ultimately leads to a significant reduction in the Kalman gain (see Equation (4)) and hence a ~~relatively small~~smaller update towards the GRACE measurements.

After conditioning, another notable feature is that the ensemble spread is significantly reduced between the OL and DA simulations. This is indicative of the DA procedure having an impact on the model-derived ensemble and suggests increased confidence in the TWS estimates via assimilation. Collectively, these findings compose a useful sanity check on the efficacy of the assimilation framework and lends some credibility to its ability to model TWS in a snow-dominated basin.

7.2. Snow Water Equivalent (SWE)

7.2.1. Comparison to CMC Product

Monthly-averaged CMC values of SWE for each of the four sub-basins as well as the entire MRB are compared against the OL and DA simulations. As discussed in section 5.2, multiple versions of each GRACE product were generated that include PGR corrections as well as exclude PGR corrections using the model of *Paulson et al.* [2007]. For brevity only the GRGS product is discussed herein as it is representative of the other GRACE products and because it yields the most complete timeseries (i.e., fewest monthly gaps) for the study simulation period. Further, only the results for the GRGS product excluding PGR corrections are shown in Figure 5. The sensitivity to the PGR corrections will be discussed later.

Differences in Figure 5 between the OL and DA simulations are apparent, most notably the reduction in ensemble ~~spread (uncertainty)~~standard deviation (spread) associated with GRACE assimilation. In general, the conditioning procedure moves the DA ensemble mean closer to the CMC estimates relative to the OL simulation. This is more apparent in the Liard basin where the snowfall accumulation is greatest, particularly in 2005-2007 and 2009 when the model has a tendency to overestimate SWE. Changes are less apparent in the other sub-basins because less snow is present, hence the changes are much smaller in magnitude, and because in general, the OL does a reasonable job of estimating SWE. This is further discussed in section 7.5 where it is shown that the updates to SWE are near-zero during much of the accumulation phase, hence differences in OL and DA SWE are relatively small.

Figure 6 shows statistics of MD, RMSD, and anomaly R for each of the sub-basins. Metrics are shown for the open loop (white), and for assimilation of GRGS GRACE TWS without (light gray) and with (dark gray) PGR correction. In terms of MD and RMSD without PGR correction, the greatest improvement is witnessed in the Liard basin. MD relative to the CMC product is reduced through assimilation by ∼30% (MD=13.2 mm for OL and MD=9.3 mm for DA) with a >15% reduction in RMSD (RMSD=24 for OL and RMSD=19.6 for DA). The other sub-basins, including the MRB as a whole, contain less snow and receive a much smaller amount of correction compared to the Liard basin. In general, the other sub-basins receive a small reduction in MD with little or no change to RMSD. Changes in MD and RMSD of SWE are essentially the same no matter which GRACE product is assimilated and no matter whether PGR correction is or is not applied (results not shown).

Unlike MD and RMSD, changes to anomaly R are typically degraded as a result of the assimilation. When excluding PGR correction, the differences are not statistically significant at the 5% level based on the Hotelling-Williams Test, but there are apparent reductions in the ability to capture the inter-annual variability of SWE when invoking the DA procedure. These results suggest the DA simulations do a reasonable job of estimating the amount of SWE in each basin but that the timing of the accumulation/ablation phases are slightly degraded when incorporating information from GRACE. When PGR correction is applied to the GRACE observations, the anomaly R degradation becomes much more pronounced, particularly in the Slave basin where PGR is most prominent in the model of *Paulson et al.* [2007] (R=0.70 for DA without PGR correction and R=0.64 for DA with PGR correction). More specifically, assimilation of the GRGS product with PGR correction yields the lowest anomaly R values among basins in both the Slave sub-basin and the MRB as a whole with values that are statistically different from the OL results via the Hotelling-Williams test.

7.2.2. Comparison to INAC Surveys

On average both the OL and DA simulations underestimate SWE when compared against the INAC ground-based observations with MD=-28 mm for OL and MD=-33 mm for DA estimates (Table 4). Each comparison was conducted by first comparing all of the surveys at a given location in space against the model output collocated in time. Then, the MD and RMSD was computed across time and subsequently presented in Table 4. The assimilation of GRACE data typically removes snow mass near peak accumulation thereby further increasing the negative bias. The INAC observations are in direct contrast to the CMC product results, which suggest a positive bias in the OL and DA results. However, given that the CMC product is conditioned on snow depth observations collected in open areas such as airports that are subject to wind-blown snow redistribution, there is a potential to introduce a negative bias into the CMC estimates (relative to the truth). Snow at the stations used in the CMC optimal interpolation routine tends to be shallower and melt earlier than in surrounding terrain [*Brown et al.*, 2003]. Hence, the disparity between the CMC product and the INAC observations within the Slave basin can be partly explained by the CMC negative bias (relative to the truth) as well as by the differences in the sampling domain between point-scale observations and the ∼24-km pixel resolution of the CMC product.

7.3. Runoff

Comparison against GRDC runoff measurements were conducted in a similar manner as with the CMC SWE estimates. However, rather than comparing by sub-basin, runoff estimates are compared against individual gauging stations. Table 3 lists the upland area and approximate sub-basin integration for each station of interest. Results are displayed in Figure 7. One notices many distinct features. Namely, all simulations (OL or DA) suffer from a significant negative bias relative to the runoff observations. This is mostly due to insufficient baseflow runoff within the model for all but the smallest of sub-basins. This is clearly demonstrated in Figure 7 at the downstream observation locations during the winter when melt flux and overland flow are near-zero because the surface water (e.g. SWE) is restrained in solid phase. Hence, the baseflow component is the dominant contributor to winter runoff. Since the observed runoff at the downstream locations is considerably larger than the modeled runoff, it is reasonable to assume that the model generates an insufficient amount of baseflow at these locations during the winter season when overland flow is minimized. One also notices an overestimation of annual peak flow, particularly during the spring freshet. This is partly due to a lack of runoff routing and lake storage routines, which contributes to a more rapid runoff response within the model. No discernible difference between the OL and DA ensemble means is witnessed in Figure 7 as the DA line effectively overlaps the OL line. However, a small (∼5-10%) reduction in ensemble standard deviation (spread) is witnessed in most sub-basins as a result of the assimilation procedure.

Figure 8 shows the corresponding computed statistics of MD, RMSD, and anomaly R at a monthly timescale at each of the gauging stations. In general, MD is slightly more negative as a result of assimilation, but to a lesser degree when PGR correction is excluded (light gray) relative to the inclusion of PGR correction (dark gray). The decrease in negative MD results from the removal of SWE during peak accumulation, which results in less runoff production during ablation. The removal of SWE is essentially counter-balanced by an increase in subsurface storage (further discussed in section 7.5), but does not translate into any significant increase in baseflow production or infiltration excess runoff, hence the slightly more negative MD. RMSD, in general, is reduced or remains unchanged in all of the sub-basins and is effectively the same between the different GRACE products (results not shown).

The greatest discord between the different assimilation experiments is found in the anomaly R values. The GRGS product without PGR correction, in general, yielded the best results. However, 2 out of 6 station locations are degraded as a result of GRACE assimilation relative to the OL results. Station number 4 (S+L+P+A in Figure 8c) undergoes a statistically significant level of improvement (R=0.25 for OL and R=0.30 for DA without PGR correction), but at the cost of statistically significant degradations at the first station (L in Figure 8c; R=0.71 for OL and R=0.64 for DA without PGR correction) and fifth station (S+L+P+A+B in Figure 8c; R=0.50 for OL and R=0.46 for DA without PGR correction). When PGR correction is included, more stations are degraded than are improved with most station degradations being significant at the 5% level. These results, in conjunction with the SWE results, suggest assimilation of the

GRGS product excluding PGR correction yields the greatest amount of improvement (and least amount of degradation) in terms of inter-annual variability.

Finally, in order to investigate the potential impact of a river routing scheme, an analysis was conducted in which runoff estimates (OL or DA) were computed using a simple, fixed-lag smoother. For a given month, the fixed-lag smoother computed the runoff as the average of the given month and the preceding n months ~~preceding~~. This effectively delays the hydrologic runoff response in a manner analogous to that of a runoff routing scheme. Based on the anomaly R and RMSD statistics between the GRDC observations and the runoff computed from the fixed-lag smoother (results not shown), the greatest improvements typically occur with a temporal lag of 1-2 months. However, the general conclusions with or without application of the fixed-lag smoother remain the same in that the runoff response with GRACE assimilation is improved, albeit by a small amount. Therefore, even though the results displayed in Figure 8c do not account for hydraulic routing, the results serve as a good proxy of the impact of GRACE assimilation on runoff estimation.

7.4. Normalized Innovation Sequence

A filter innovation is the difference between the ensemble mean observation and model forecast, $\mathbf{y}_t - \mathbf{H}\mathbf{x}_t^-$, at a given time, t. Investigation of filter innovations is a useful tool for assessing whether or not measurement (Table 1) and model (Table 2) error parameters have been appropriately selected. If a model is linear and all errors are Gaussian, then the normalized innovations, NI, should appear similar in form to white noise (i.e., zero mean, unit variance, and temporally uncorrelated). Even though the application used here is a smoother rather than a filter and the forward model is non-linear, the investigation of the normalized innovations can provide useful information as to the performance of the DA procedure.

The normalized innovation may be written as

$$NI_t = \frac{\mathbf{y}_t - \mathbf{H}\mathbf{x}_t^-}{\sqrt{\mathbf{H}\mathbf{P}_t^-\mathbf{H}^T + \mathbf{R}}}, \tag{5}$$

where the numerator represents the difference between the assimilated measurement and the predicted measurement, and the denominator represents a combination of the background error covariance and the measurement error covariance. Normalized innovations are collected as a function of time and then the mean is computed as $\overline{NI} = \frac{1}{T}\sum_{t=1}^{T} NI_t$ while the standard deviation computed as $\sigma_{NI} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(NI_t - \overline{NI}\right)^2}$.

Figure 9 plots the mean versus the standard deviation of the normalized innovations for each of the four (4) sub-basins using the GRGS product excluding PGR correction. The different colors represent different amounts of measurement error standard deviation used during the DA experiments relative to the nominal values listed in Table 1. The most striking feature is that all of the mean innovations are negative regardless of the sub-basin or the measurement error. This suggests the DA procedure attempts to correct a systematic bias where the model contains too much water relative to the GRACE observations during certain times of the year. This can be seen via inspection of Figure 4e where the individual sub-basin GRACE updates effectively remove mass most years at peak accumulation, particularly after January 2005. During the ablation and runoff phase, GRACE DA attempts to add mass in the subsurface, but the amount of mass added is, in general, less than the amount of SWE removed. Hence, the result is a posterior ensemble with less TWS. This behavior is further discussed in the following section via inspection of the analysis increments.

The second feature of note in Figure 9 is the wide range in σ_{NI} resulting from changes to the measurement error standard deviation. As expected, an increase in measurement error causes an increase in the denominator of Equation (5), which causes a corresponding reduction in the spread (or standard deviation) of the normalized innovation sequence. If the design of $\mathbf{HP}_t^-\mathbf{H}^T$ is assumed reasonable, Figure 9 suggests that 2× the nominal measurement error standard deviation of Table 1 is too large. A large measurement error variance (relative to the background error variance) results in a small value of the gain $\mathbf{K}$ (Equation (3)), which leads to only minimal assimilation updates. Conversely, a value of 0.5× the nominal measurement error standard deviation is too small, which causes the assimilation to overly "trust" the measurement quality and effectively make too large of an update toward the GRACE measurements. Based on σ_{NI}, application of 1.0× to 1.5× the estimated measurement error appears reasonable and is similar to the GRACE measurement errors used in *Zaitchik et al.* [2008] and *Su et al.* [2010].

7.5. Analysis Increments

Investigation of the analysis increments (i.e., difference between $\mathbf{x}_t^+$ and $\mathbf{x}_t^-$) can provide valuable insight into the behavior of the assimilation procedure. It enables one to track mass within the relevant TWS components in order to see how much and at what time mass is being added to or removed from the system. Figure 10 shows the analysis increments from the assimilation of the GRGS product excluding PGR correction. The thin, solid line shows the increments made to the subsurface TWS component as the negative of the catchment deficit prognostic variable. Assimilation updates were not applied to the surface soil excess or root zone soil excess states. However, this is inconsequential as the efficacy with which Catchment redistributes water in the subsurface is overwhelmingly dominated by the catchment deficit variable [*Zaitchik et al.*, 2008]. Averaged over time and space the increments are positive for a total of 12.5 mm, which means assimilation results in increasing the amount of water in the subsurface. This is most evident during the spring and summer. The thick, dashed line shows the increments for SWE summed across the three individual SWE layers. Averaged over time and space SWE is removed during the accumulation phase with a small amount added back during the ablation and runoff phase for a total SWE increment of -45.1 mm. Collectively, the analysis increments to the catchment deficit and SWE serve to reduce mass during snow accumulation and then increase the mass during ablation and runoff. These two phenomena essentially constrain the amplitude of the modeled TWS dynamics such that better agreement with the GRACE observations is achieved.

8. Conclusions

GRACE-derived estimates of TWS were assimilated into a land surface model for the purpose of improved snow pack characterization in northwestern Canada. It was shown that the conditioning procedure, in general, could reduce MD and RMSD in the SWE estimates (prior versus posterior) when compared against the CMC snow product. However, anomaly R values were typically degraded as a result of the assimilation. Even though the anomaly R differences were not statistically significant at the 5% level, they suggest some degree of reduced skill at simulating inter-annual variability when using the DA procedure. A comparison of model results against GRDC runoff observations suggested relatively little change to runoff MD and RMSD statistics. Anomaly R values for runoff, however, were improved at several locations and remain essentially unchanged at the basin outlet. Improvements to anomaly R values for runoff are mostly attributable to a more delayed runoff response with assimilation.

These results are encouraging, but it is important to highlight shortcomings and discuss potential improvements that could be made in future developments. For example, the land surface model used in this study does not contain a river routing scheme. Runoff is effectively routed to the outlet instantaneously. However, given the size and extent of the MRB, runoff residence times near the basin outlet can be conservatively estimated to be on the order of a couple of months. The improvements to runoff anomaly R values are generally associated with a delayed runoff response that effectively retains water within the basin for a longer period of time. That is, the assimilation acts to correct some of the limitations in the model physics that could likely be addressed via inclusion of a runoff routing routine. Similarly, the land surface model does not contain a lake storage routine. Changes in lake storage can be a significant contributor to TWS and can also be an important factor in attenuating hydrologic runoff response at the basin outlet. Analogous to a runoff routing routine, inclusion of a lake level storage routine could likely improve runoff timing relative to the GRDC observations. Development and testing of runoff routing and lake storage routines are beyond the scope of this current study, but would be worthwhile addressing in future work.

In addition, another limitation of this study is a lack of subsurface observations (i.e., soil moisture and groundwater) to evaluate model results. Updates to the catchment deficit prognostic variable can only be discussed in a qualitative sense without a valid dataset to make quantitative comparisons. Unfortunately, soil moisture and groundwater level measurements are not readily available in hydrologic basins located in the high latitudes thereby making such a comparison difficult if not impossible. The lack of subsurface observations severely limits the conclusions that can be made about the ability of the assimilation to effectively disaggregate TWS into snow, soil moisture, and groundwater components.

Despite these shortcomings, the GRACE DA procedure did improve MD and RMSD statistics of SWE in the MRB as well as improved some runoff estimates in terms of inter-annual variability. These preliminary findings are encouraging and suggest the potential for further improvements via merger with passive microwave and visible spectrum remote sensing products to further downscale the GRACE observations in time and space while simultaneously disaggregating the GRACE observations into individual, vertical components of TWS. Finally, additional improvements could be achieved through refining the GRACE measurement error model, investigating the effects of different horizontal error correlation lengths within the land surface model forcings, determining a more optimal GRACE measurement scale, utilizing a more optimal GRACE averaging kernel, and better constraining of PGR model estimates used during GRACE preprocessing.

Acknowledgments. Funding provided by the NASA Postdoctoral Program Fellowship (Contract NNH06CC03B). Additional thanks go to Ross Brown for answers to our questions regarding the CMC snow analysis product; Derek Faria for access to the INAC snow surveys; Ulrich Looser for access to the GRDC database; John Wahr, Sean Swenson, and Felix Landerer for discussions on PGR; and Bailing Li for many constructive comments.

References

K.M. Andreadis and D.P. Lettenmaier. Assimilating remotely sensed snow observations into a macroscale hydrology model. *Adv. Water Resour.*, 29:872–886, 2006.

T.P. Barnett, J.C. Adam, and D.P. Lettenmaier. Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, 438:303–309, 2005.

B. Brasnett. A global analysis of snow depth for numerical weather prediction. *J. Appl. Meteorol.*, 38:726–740, 1999.

S. Bruinsma, J.-M. Lemoine, R. Biancale, and N. Valès. CNES/GRGS 10-day gravity field models (release 2) and the evaluation. *Adv. Space Res.*, 45:587–601, 2010.

R.D. Brown and B. Brasnett. Canadian Meteorological Centre (CMC) daily snow depth analysis data. Environment Canada, 2010. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.

R.D. Brown, B. Brasnett, and D. Robinson. Gridded North American monthly snow depth and snow water equivalent for GCM evaluation. *Atmos. Ocean*, 41:1–14, 2003.

G. Burgers, P.J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Mon. Weather Rev.*, 126:1719–1724, 1998.

M.P. Clark, A.G. Slater, A.P. Barrett, L.E. Hay, G.J. McCabe, B. Rajagopalan, and G.H. Leavesley. Assimilation of snow covered area information into hydrologic and land-surface models. *Adv. Water Resour.*, 29:1209–1221, 2006.

D. Clifford Global estimates of snow water equivalent from passive microwave instruments: History, challenges, and future developments. *Int. J. Remote Sens.*, 31:3707–3726, 2010.

J. Dong, J.P. Walker, P.R. Houser, and C. Sun. Scanning multichannel microwave radiometer snow water equivalent assimilation. *J. Geophys. Res.*, 112:doi:10.1029/2006JD007209, 2007.

M.T. Durand and S.A. Margulis. Feasibility test of multi-frequency radiometric data assimilation to estimate snow water equivalent. *J. Hydrometeorol.*, 7:443–457, 2006.

S. Dunne and D. Entekhabi. Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. *Water Resour. Res.*, 42:doi:10.1029/2005WR004334, 2006.

M. Horwath, J.-M. Lemoine, R. Biancale, and S. Bourgogne. Improved GRACE science results after adjustment of geometric biases in the Level-1B K-band ranging data. *J. Geod.*, 85–23:38, 2011.

G.J. Huffman, R.F. Adler, P.A. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider. The global precipitation climatology project (GPCP) combined precipitation dataset. *B. Am. Meteorol. Soc.*, 78:5–20, 1997.

S.P. Khare, J.L. Anderson, T.J. Hoar, and D. Nychka. An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. *Tellus*, 60A:97–112, 2008.

R.D. Koster, M.J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar. A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure. *J. Geophys. Res.*, 105:24809–24822, 2000.

D.B. McLaughlin. An integrated approach to hydrologic data assimilation: Interpolation, smoothing, and filtering. *Adv. Water Resour.*, 12:1275–1286, 2002.

A. Paulson, S. Zhong, and J. Wahr. Inference of mantle viscosity from GRACE and relative sea level data. *Geophys. J. Int.*, 171:497–508, 2007.

R.H. Reichle, D. Entekhabi, and D.B. McLaughlin. Downscaling of radio brightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. *Water Resour. Res.*, 37:2353–2364, 2001.

R.H. Reichle and R.D. Koster. Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *J. Hydrometeorol.*, 4:1229–1242, 2003.

R.H. Reichle, R.D. Koster, P. Liu, S.P.P. Mahanama, E.G. Njoku, and M. Owe. Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the Scanning Multichannel Microwave Radiometer (SMMR). *J. Geophys. Res.*, 112: doi:10.1029/2006JD008033, 2007.

R.H. Reichle, W.T. Crow, and C.L. Keppenne. An adaptive ensemble Kalman filter for soil moisture data assimilation. *Water Resour. Res.*, 44: doi:10.1029/2007WR006357, 2008.

R.H. Reichle, R.D. Koster, G.J.M. De Lannoy, B.A. Forman, Q. Liu, S. Mahanama, and A. Toure. Assessment and enhancement of MERRA land surface hydrology estimates. *J. Clim.*, In Press, doi: 10.1175/JCLI-D-10-05033.1, 2011.

M.M. Rienecker, M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R.D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C.R. Redder, R.H. Reichle, F.R. Robertson, A.G. Ruddick, M. Sienkiewicz, J. Woolen. MERRA - NASA's modern-era retrospective analysis for research and applications. *J. Clim.*, 24:3624–3648, 2011.

M. Rodell, I. Velicogna, and J.S. Famiglietti. Satellite-based estimates of groundwater depletion in India. *Nature*, 460:999–1002, 2009.

D.D. Rowlands, S.B. Luthcke, S.M. Klosko, F.G.R. Lemoine, D.S. Chinn, J.J. McCarthy, C.M. Cox, and O.B. Anderson. Resolving mass flux at high spatial and temporal resolution using GRACE intersatellite measurements. *Geophys. Res. Lett.*, 32:doi:10.1029/2004GL021908, 2005.

D.D. Rowlands, S.B. Luthcke, J.J. McCarthy, S.M. Klosko, D.S. Chinn, F.G. Lemoine, J.-P. Boy, and T.J. Sabaka. Global mass flux solutions from GRACE: A comparison of parameter estimation strategies – Mass concentrations versus Stokes coefficients. *J. Geophys. Res.*, 115:doi:10.1029/2009JB006546, 2009.

J.H. Steiger. Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, 87:245–251, 1980.

M. Stieglitz, A. Ducharne, R.D. Koster, and M. Suarez. The impact of detailed snow physics on the simulation of snow cover and subsurface thermodynamics at continental scales. *J. Hydrometeorol.*, 2:228–242, 2001.

M. Sturm, B. Taras, G.E. Liston, C. Derksen, T. Jonas, and J. Lea. Estimating snow water equivalent using snow depth data and climate classes. *J. Hydrometeorol.*, 11:1380–1394, 2010.

H. Su, Z.-L. Yang, G.-Y. Niu, and R.E. Dickinson. Enhancing the estimation of continental-scale snow water equivalent by assimilating MODIS snow cover with the ensemble Kalman filter. *J. Geophys. Res.*, 113:doi:10.1029/2007JD009232, 2008.

H. Su, Z.-L. Yang, R.E. Dickinson, C.R. Wilson, and G.-Y. Niu. Multisensor snow data assimilation at the continental scale: The value of Gravity Recovery and Climate Experiment terrestrial water storage information. *J. Geophys. Res.*, 115:doi:10.1029/2009JD013035, 2010.

S. Swenson, P. J.-F. Yeh, J. Wahr, and J.S. Famiglietti. A comparison of terrestrial water storage variations from GRACE with in situ measurements from Illinois. *Geophys. Res. Lett.*, 33:doi:10.1029/2006GL026962, 2006.

S. Swenson and J. Wahr. Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.*, 33:doi:10.1029/2005GL025285, 2006.

S. Swenson. Restoring signal loss in GRACE terrestrial water storage estimates. *Geophys. Res. Lett.*, In Preparation, 2011.

T.H. Syed, J.S. Famiglietti, and D.P. Chambers. GRACE-based estimates of terrestrial freshwater discharge from basin to continental scales *J. Hydrometeorol.*, 10:22–40, 2009.

Q. Tang, H. Gao, P. Yeh, T. Oki, F. Su, and D.P. Lettenmaier. Dynamics of terrestrial water storage change from satellite and surface observations and modeling. *J. Hydrometeorol.*, 11:156–170, 2010.

J. Wahr, S. Swenson, V. Zlotnicki, and I. Velicogna. Time-variable gravity from GRACE: first results. *Geophys. Res. Lett.*, 31:doi:10.1029/2004GL019779, 2004.

J. Wahr, S. Swenson, and I. Velicogna. Accuracy of GRACE Estimates. *Geophys. Res. Lett.*, 33:doi:10.1029/2005GL025305, 2006.

B.F. Zaitchik, M. Rodell, and R.H. Reichle. Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi river basin. *J. Hydrometeorol.*, 9:doi:10.1175/2007JHM951.1, 2008.

Table 1. Sub-basin characteristics for the MRB (land areas only) along with applied GRACE measurement error covariance, **R**.

Sub-basin Name	Land Area [km]	**R** [mm^2]
Peel+Bear	4.1×10^5	18^2
Slave	3.6×10^5	16^2
Liard	2.8×10^5	17^2
Peace+Athabasca	5.7×10^5	16^2
Entire Mackenzie	1.6×10^6	17^2

Table 2. Parameters for perturbations to meteorological forcing inputs and model prognostic variables.

Perturbation	Type	Standard Deviation	Units	**L** [deg]	AR(1) [day]
Precipitation	M	0.5	-	2	3
Shortwave Radiation	M	0.5	-	2	3
Longwave Radiation	A	50	W m^{-2}	2	3
Snow Water Equivalent[a]	M	0.0004	-	2	1
Catchment Deficit	A	0.05	mm	2	1
Surface Soil Excess	A	0.02	mm	2	1

[a]Perturbations made to all three (3) snow layers; M=Multiplicative; A=Additive; **L**=spatial correlation length; AR(1)=first-order auto-regressive temporal correlation

Table 3. GRDC runoff measurement characteristics.

Station Number	Station ID	Upland Area [km^2]	Sub-basin Aggregation
1	4208271	2.75×10^5	Liard
2	4208450	2.93×10^5	Peace
3	4208400	6.06×10^5	Peace+Athabasca
4	4208005	1.27×10^6	Slave+Liard+Peace+Athabasca
5	4208150	1.57×10^6	Slave+Liard+Peace+Athabasca+Bear
6	4208025	1.66×10^6	Entire Mackenzie

Table 4. Statistics for the OL and DA experiments relative to the INAC snow surveys.

Ensemble	MD [mm]	RMSD [mm]
OL	-28	39
DA	-31	41

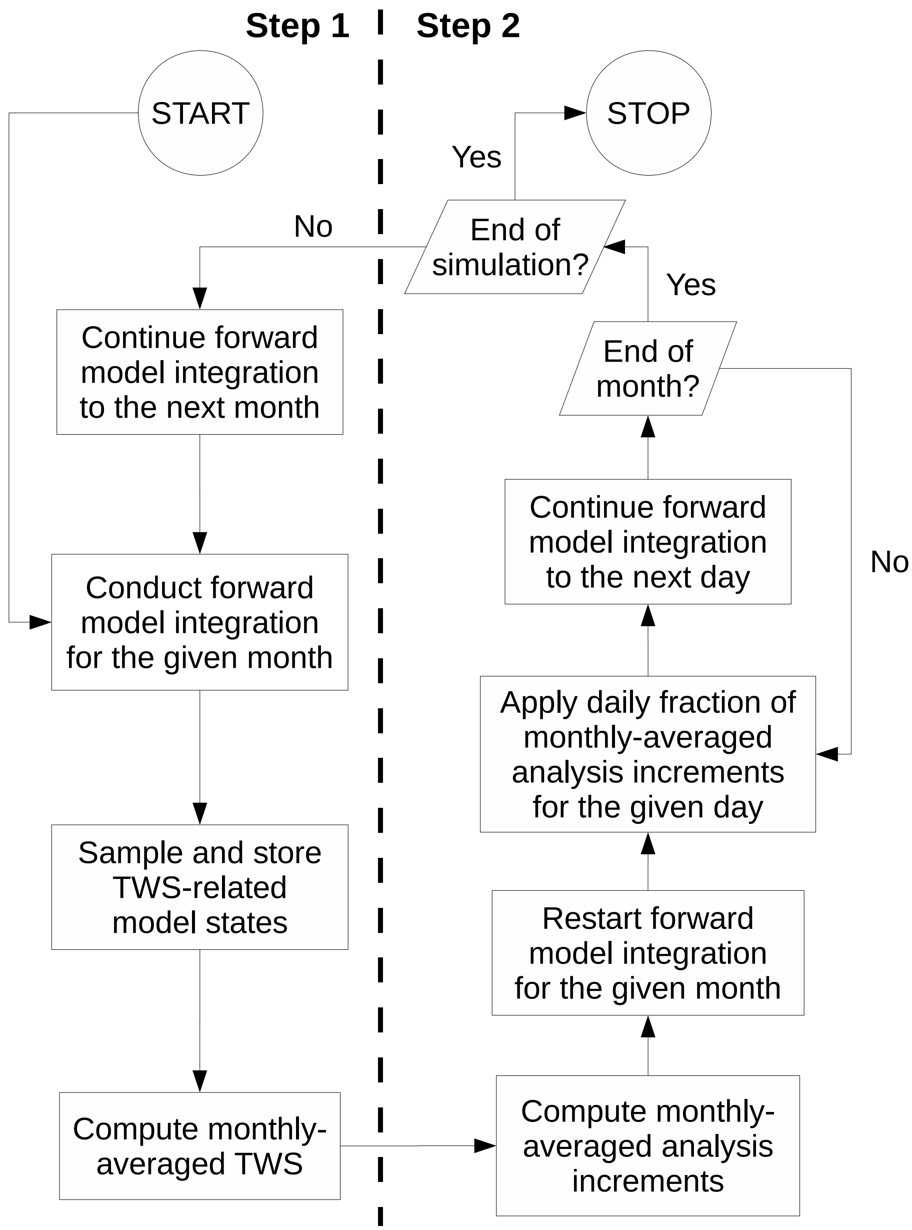
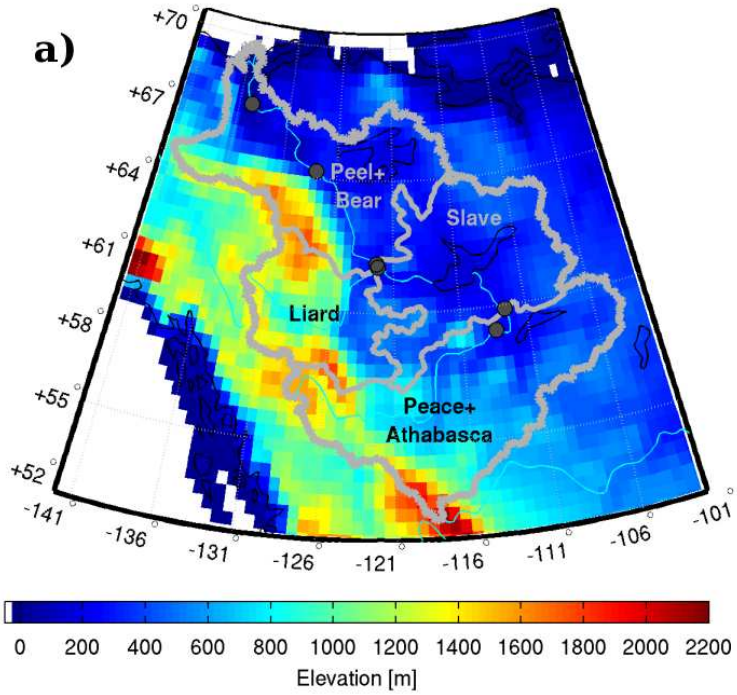


Figure 1. Simplified flowchart of EnKS application.



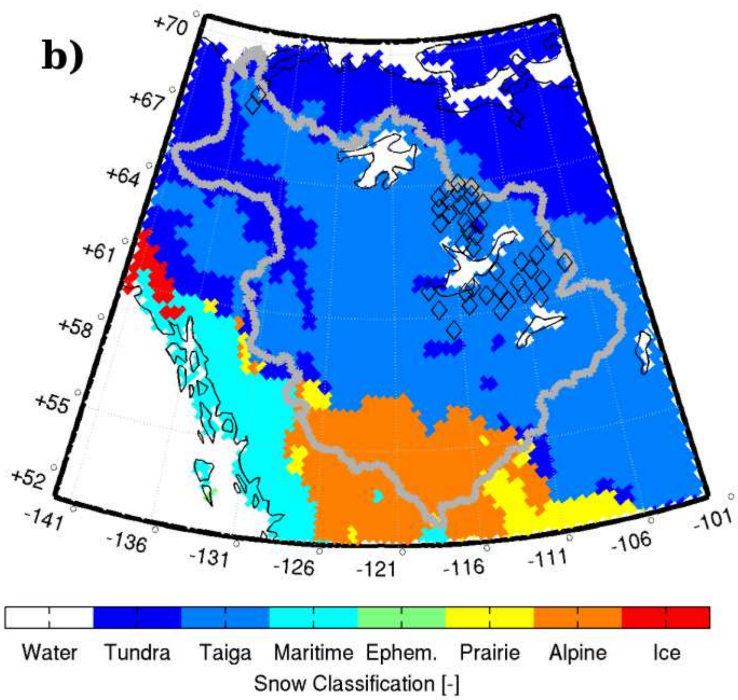


Figure 2. Map of Mackenzie River Basin including a) GEOS-5 topography, sub-basin delineation, and GRDC observation locations (solid dots), and b) *Sturm et al.* [2010] snow type with INAC snow survey locations (hollow diamonds).

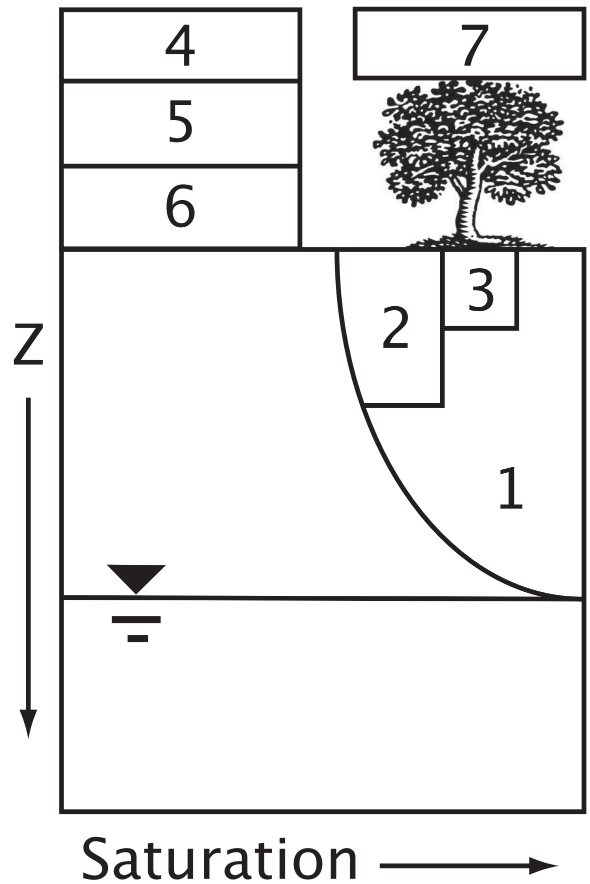


Figure 3. Conceptual representation of the components of Catchment model terrestrial water storage where 1=catchment deficit, 2=root zone excess, 3=surface soil excess, 4-6=individual snow layers, and 7=canopy interception.

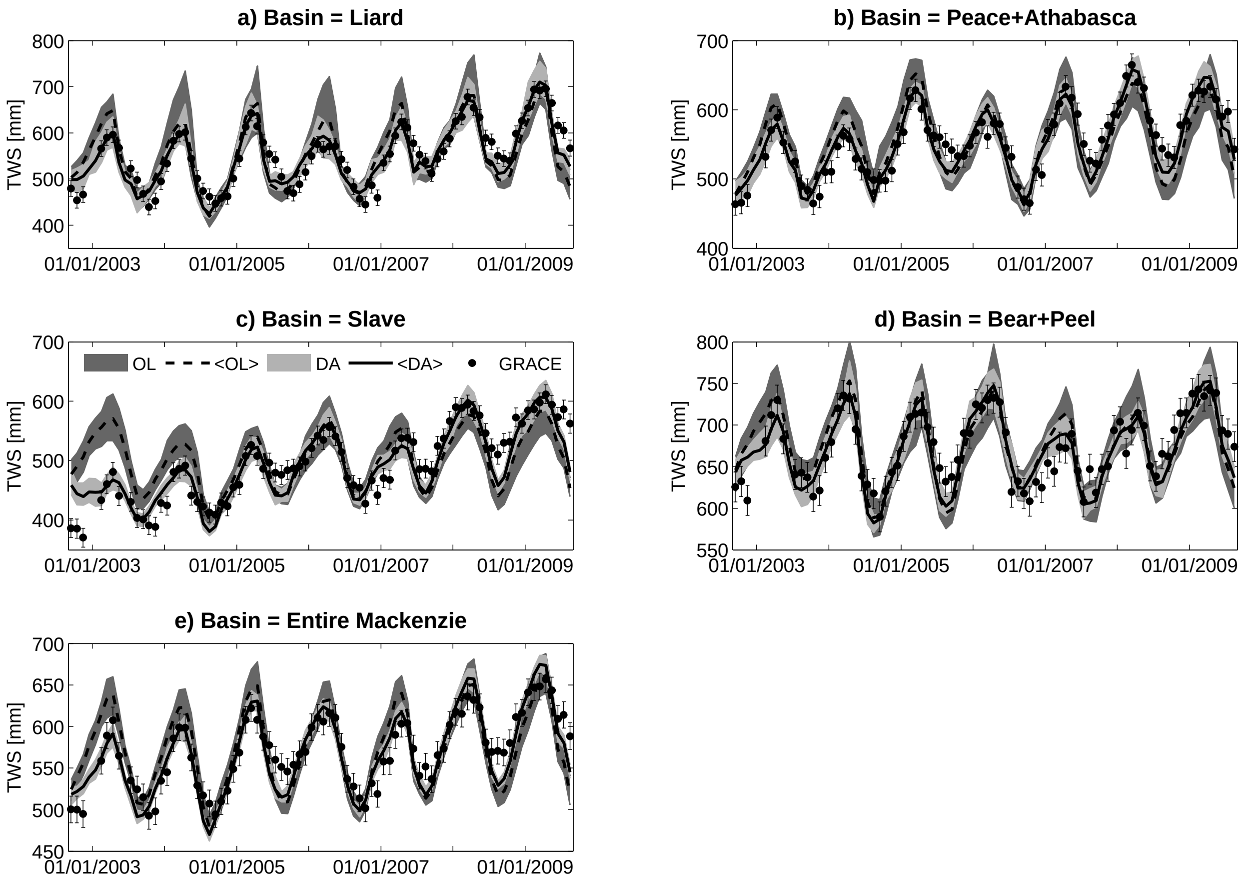


Figure 4. TWS estimates for the OL (dark gray), DA (light gray), and GRACE (dots) for the GRGS product without PGR correction. Each line represents the respective ensemble mean whereas the error bars represent the standard deviation of the GRACE observations.

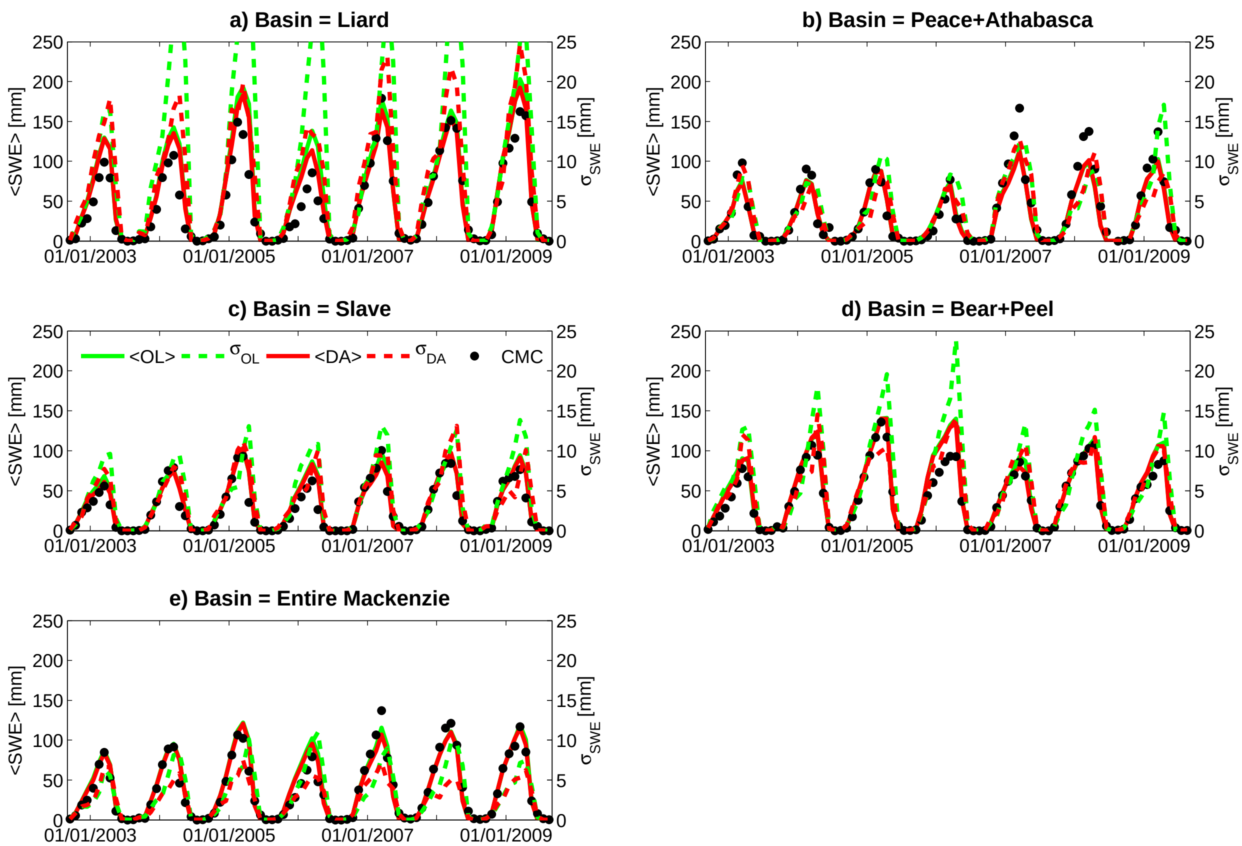


Figure 5. SWE estimates from OL (green), DA (red), and CMC (black dots) for the GRGS product without PGR correction. Solid lines represent the ensemble means (left axis) and dashed lines represent the ensemble standard deviations (right axis). CMC SWE estimates were derived from CMC snow depths using *Sturm et al.* [2010] snow densities.

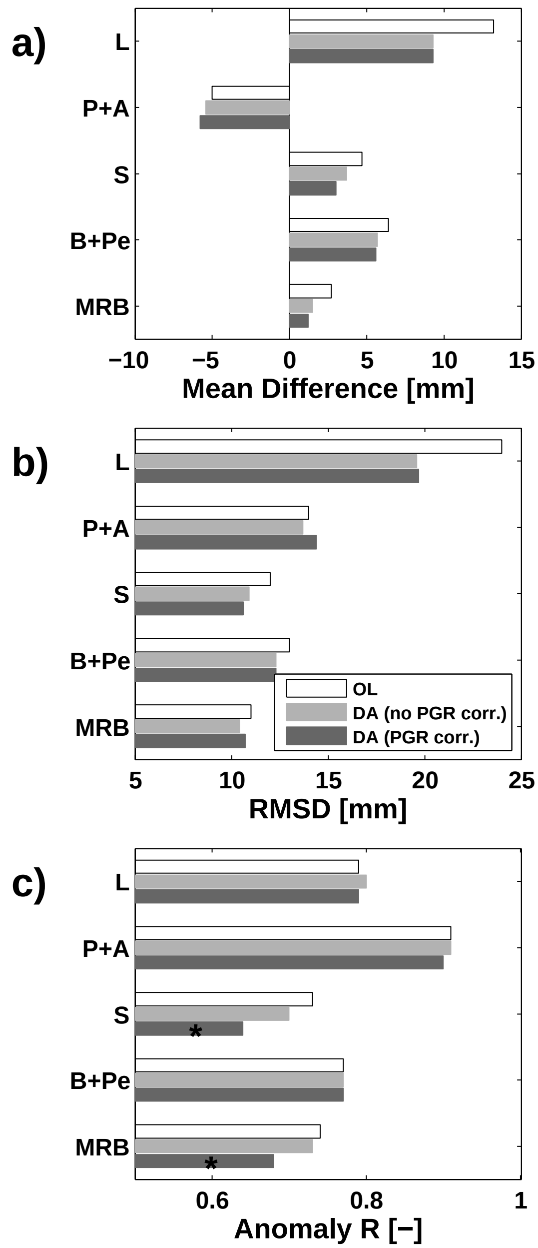


Figure 6. SWE statistics of a) MD, b) RMSD, and c) anomaly R for open-loop (white), DA without PGR correction (light gray), and DA with PGR correction (dark gray) results relative to CMC-derived SWE estimates via *Sturm et al.* [2010]. For anomaly R values, asterisks indicate statistically significant differences between the OL and DA.

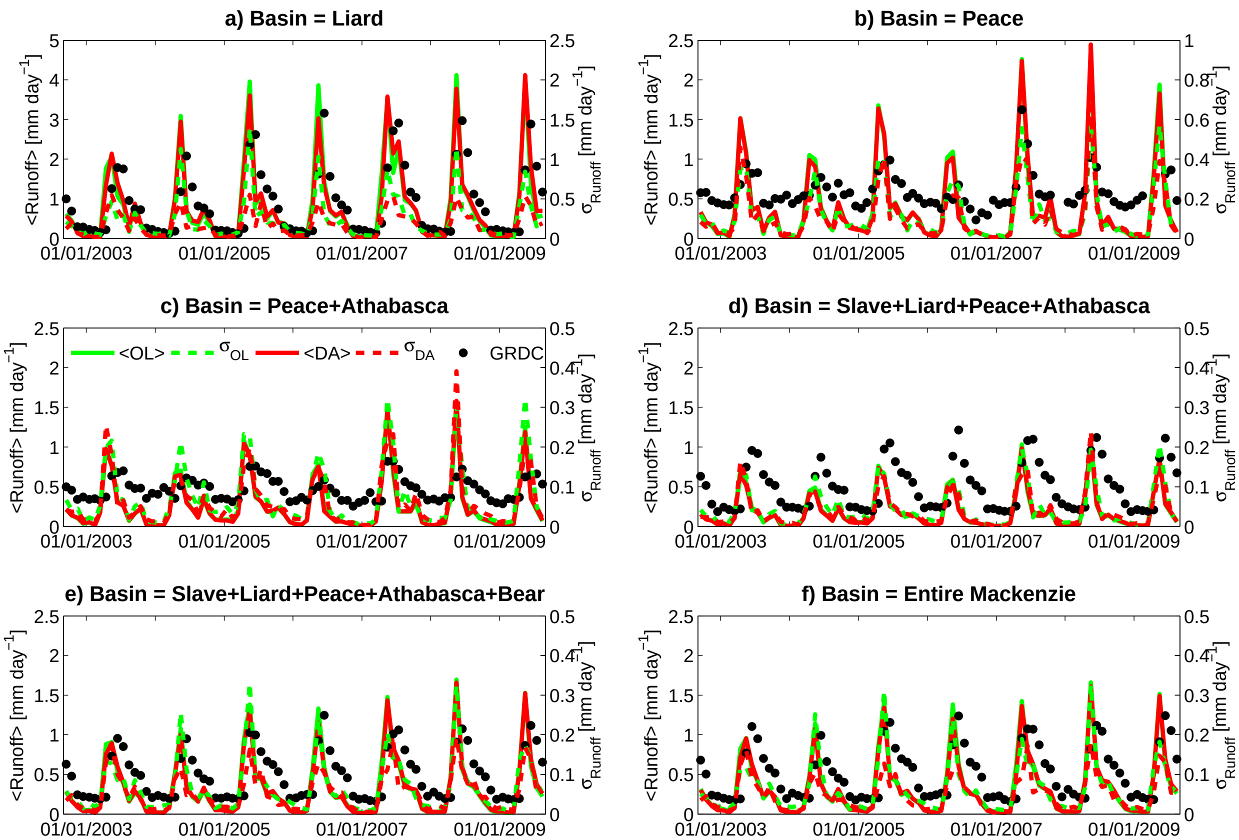


Figure 7. Runoff from OL (green), DA (red), and GRDC observations (black dots) at 6 different locations for the GRGS product without PGR correction. Upland drainage area increases from the upper-left subplot through the lower-right subplot (see Table 3 for definitions). Solid lines represent the ensemble means (left axis) and dashed lines represent the ensemble standard deviations (right axis).

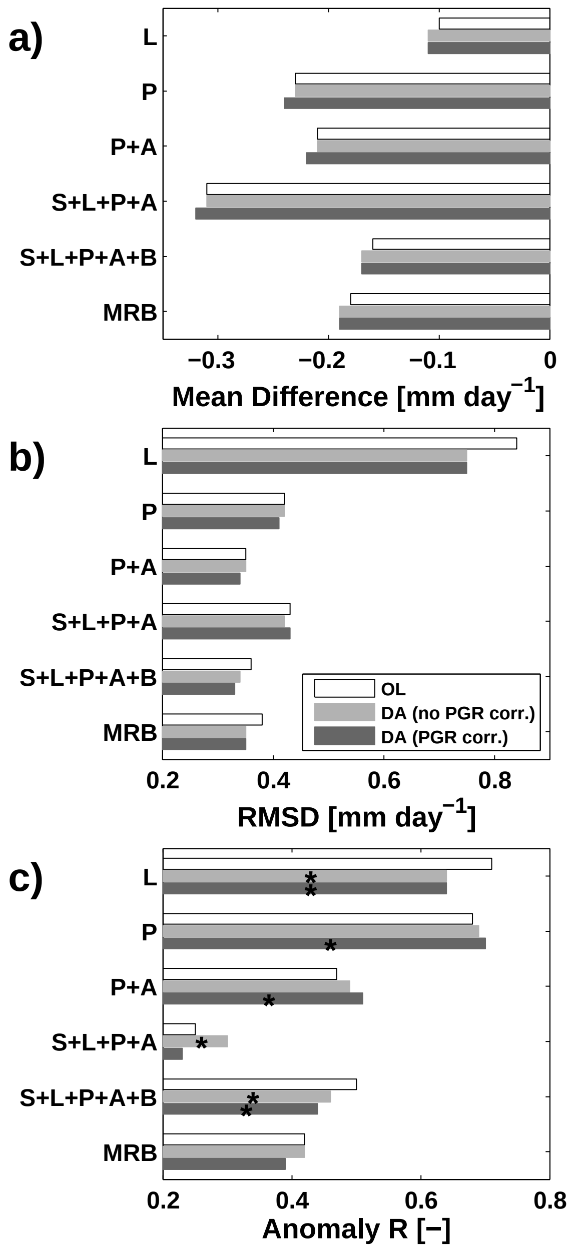


Figure 8. Runoff statistics of a) MD, b) RMSD, and c) anomaly R for open-loop (white), DA without PGR correction (light gray), and DA with PGR correction (dark gray) results relative to GRDC runoff estimates for the GRGS product without PGR correction. For anomaly R values, asterisks indicate statistically significant differences between the OL and DA.

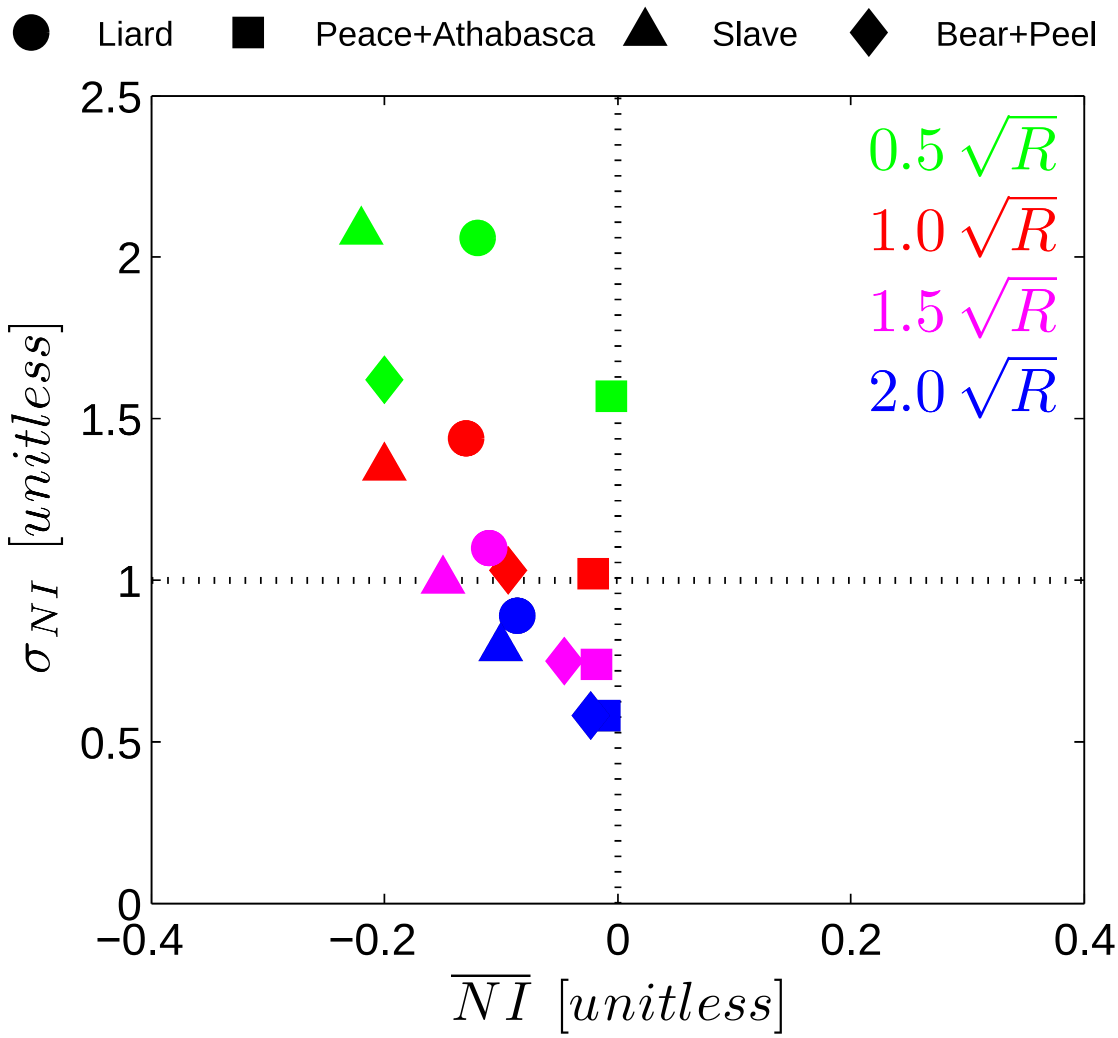


Figure 9. Innovation statistics for the GRGS product without PGR correction for the 4 sub-basins shown as different marker shapes. The different marker colors represent varying amounts of GRACE measurement error standard deviation relative to the nominal values shown in Table 1.

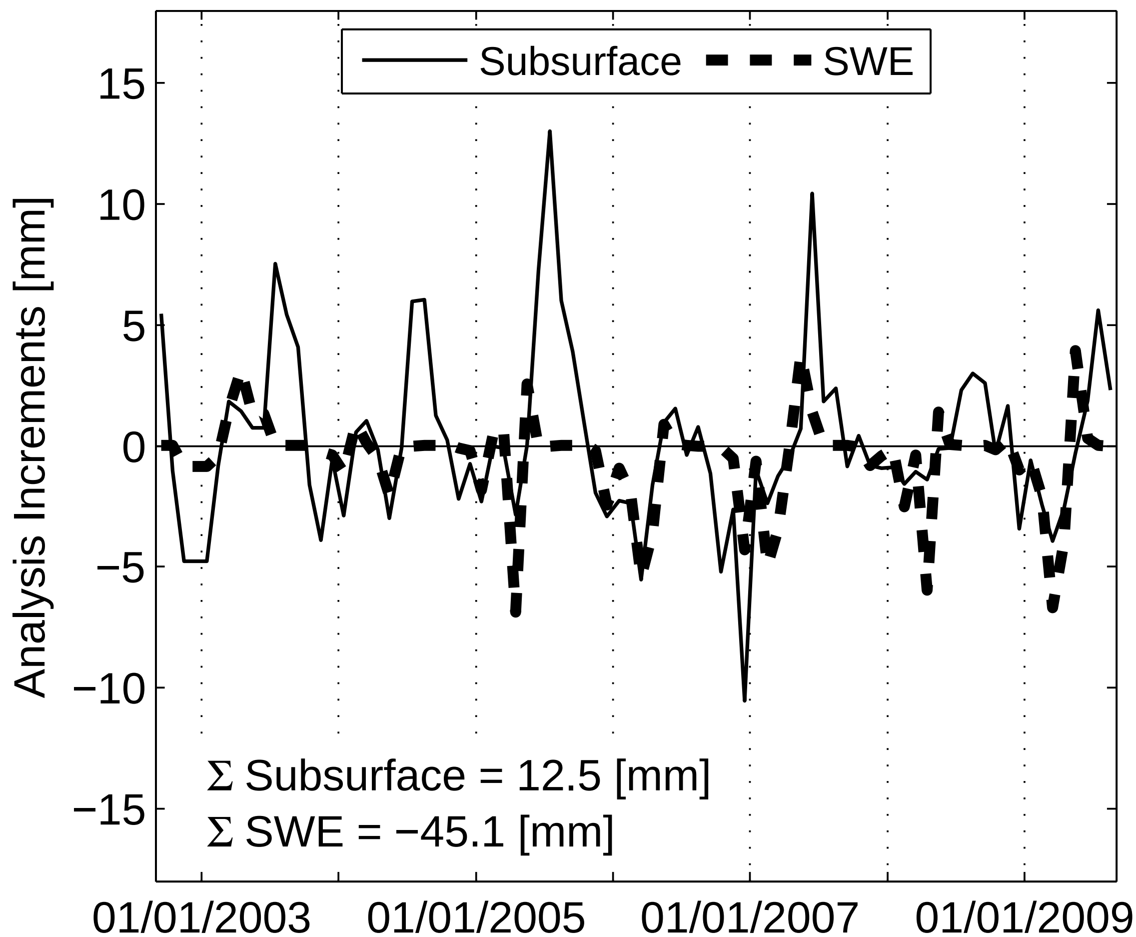


Figure 10. Analysis increments for the entire MRB using the GRGS product without PGR correction. The thin, solid line represents the subsurface increments (displayed as the negative of the catchment deficit increments) whereas the thick, dashed line represents the increments from the summation of the three individual SWE layers.