Section S1. Isolation of GWS from GRACE TWS

At monthly or longer timescales, soil water storage (SWS) and GWS account for the majority of TWS signals in NCP; other components like snow water equivalent (<0.01 km³/yr) are negligible [Feng et al., 2013]. The GWS anomalies (i.e., the deviation from the long-term mean) are derived by subtracting the model-simulated SWS anomalies and the impacts on mass changes from the reservoir regulation (RES) and the south-to-north water diversion (WD) from the GRACE-derived total TWS anomalies:

\[ GWS = TWS - SWS - RES - WD, \]  
\[ (A1) \]

where \( TWS \) is estimated from GRACE data; \( SWS \) from the soil moisture simulated by five Land Surface Models (LSMs): CLM, MOS, NOAH and VIC from GLDAS-1, and the CPC model; \( RES \) and \( WD \) from the estimation in previous study of Tang et al. [2013].

We use the Level-2 RL05 data from GRACE provided by the Center for Space Research (CSR) of the University of Texas. The data from January 2003 to July 2013 are used in this study, with a total of 121 months, excluding the missing data of June 2003, January and June 2011, May and October 2012, and March 2013. We performed a Cubic Spline Interpolation (CSI) [Press et al., 1992] for the monthly...
TWS time series to obtain a conservative estimate of the missing values. (The data can be obtained here: http://www.csr.utexas.edu/grace/, or http://icgem.gfz-potsdam.de/ICGEM/ICGEM.html).

The RL05 data are greatly improved on the basis of RL04, and the noise of the signal is substantially lower [Bettadpur and the CSR Level-2 Team, 2012a]. However, the small-scale basin-average time series data are susceptible to the episodic additional errors, depending on specific place and time [Bettadpur and the CSR Level-2 Team, 2012b]. To guarantee the accuracy of GRACE TWS anomalies, the data sets are further pre-processed here in this study. We remove a 6-year-averaged gravity field from 2005 to 2010, the same period as the in situ GW-level measurements, exclude the Stoke coefficient of degree 0 and degree 1, and replace the C20 terms from the satellite laser ranging [Cheng and Tapley, 2004]. The de-correlation method P3M10 [Chambers, 2006; Chen et al., 2007] and a 200-km Gaussian filter [Swenson and Wahr, 2006] are applied to reduce the errors in GRACE data. Owing to that the estimation is performed at a smaller subregional scale, an exact regional kernel function [Swenson and Wahr, 2002; Longuevergne et al., 2010], which is truncated at degree 60 and filtered using 200-km Gaussian smoother at the 0.25° resolution, is further applied to each catchment respectively to obtain the TWS anomalies in the study area according to the following equation:

\[
\Delta h_{\text{region}} = \frac{a \rho_e}{3 \Omega_{\text{region}} \rho_w} \sum_{l=0}^{60} \sum_{m=-l}^{l} \frac{2l+1}{1+k_l} W_l (v^l_m \Delta C + v^l_m \Delta S),
\]

(A2)

where \(\Delta h_{\text{region}}\) is the regional-averaged equivalent water height; \(a\) and \(\rho_e\) are the mean radius and density (5517 kg/m\(^3\)) of the earth; \(\rho_w\) is the mean density of water (1000 kg/m\(^3\)); \(\Omega_{\text{region}}\) is the singular area of the region (\(\Omega_{\text{region}} = S_{\text{region}} / a^2\)); \(W_l\) is the smoothing coefficients of the Gaussian filter at degree \(l\); \(\Delta C\) and \(\Delta S\) are the stoke coefficients from GRACE; \(k_l\) is the load potential
Love number at degree $l$; $v_{lm}^c$ and $v_{lm}^s$ are the spherical harmonic coefficients describing the exact kernel function:

$$v(\theta, \phi) = \frac{1}{4\pi} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \hat{P}(\cos \theta) \left\{ v_{lm}^c \cos m\phi + v_{lm}^s \sin m\phi \right\},$$  

(A3)

$$\begin{bmatrix} v_{lm}^c \\ v_{lm}^s \end{bmatrix} = \int v(\theta, \phi) \hat{P}(\cos \theta) \begin{bmatrix} \cos m\phi \\ \sin m\phi \end{bmatrix} d\Omega, \quad \text{(A4)}$$

In this paper, our assessment is based on the RL-05 data sets provided by UTCSR. Up to now, the latest release of GRACE data provided by CSR has a maximum degree of 96 for the spherical harmonic coefficients. It is reported that there are no regionally-specific error pattern differences associated with solution computation to d/o 96 or 60 [Sakumura et al., 2014]. In this study, we choose to use the previous release of the RL05 data at degree 60. The results from other data products (GeoForschungsZentrum (GFZ) Potsdam, Groupe de Recherches de Géodésie Spatiale (GRGS), and Jet Propulsion Laboratory (JPL)) are also shown in this auxiliary material to show the impact of the uncertainties in GRACE data. GFZ and JPL provide data (spherical harmonic coefficients) at the monthly intervals with a maximum degree of 90. The latest monthly data sets provided by GRGS have expanded the maximum degree from 50 to 80. The time span of the data from GFZ and JPL we use is the same as that of the data from CSR, i.e. from January 2003 to July 2013, with 6 months missing data (mentioned before). The data provided by GRGS from January 2003 to December 2012 are used, excluding the missing data of 3 months (June 2003, January 2011 and October 2012). In order to compare with the results from CSR, all of the above three data products (GFZ, JPL and GRGS) are selected at the same time interval (monthly) and the same degree (60) as CSR data, and they are further processed in the same way as for the CSR data.

The SWS is estimated by using the simulation data from the following five LSMs: the four LSMs
driven by GLDAS-1 (Mosaic, Noah, the Community Land Model (CLM), the Variable Infiltration Capacity (VIC)) [Rodell et al., 2004], and the modeled soil moisture (SM) data provided by the Climate Prediction Center (CPC) [Fan and van den Dool, 2004]. We use the monthly data from GLDAS-1 and CPC at a full time span from January 2003 to July 2013. The data can be obtained from here: http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings (GLDAS), http://www.esrl.noaa.gov/psd/data/gridded/tables/monthly.html (CPC).

The SM data from LSMs are simulated at different depths with a maximum depth of 1.9-3.5 m for four LSMs in GLDAS, and 1.6 m for the CPC model. Instead of using the SM from a single model, we use the averaged times series from the five LSMs for the GWS disaggregation from TWS. As it was implied that the water storage at depths exceeding the soil moisture depth is interpreted as GWS [Huang et al., 2012], the GWS anomalies estimated in this study then can represent the vertically integrated water storage variations from the aquifers below 3.5 m, the maximum depth of Mosaic. To obtain the anomalies of SM, the 2005-2010 average (i.e., the same time period with the available in situ GW-level measurements) is removed.

Section S2: Forward modeling for bias correction in GRACE data

The terminology “bias” in this study means the reduction of signals inside of the subregion (or “leakage-out”), and the term “leakage” means the effects of signal from the outside regions (or “leakage-in”). This study follow the updated method of Scanlon et al. [2012] to reduce the bias and leakage errors for the small-scale hydrologic application using GRACE data.

In the updated processing procedures, the global gridded SWS data obtained from LSMs (unconstrained simulations) need to be expanded into the spherical harmonic coefficients and processed in the same way as GRACE data (i.e. truncated at the degree 60, using a 200-km Gaussian
smoother, but no de-correlation filter is applied [Chen et al., 2014]). Then, the subtraction of processed SWS data from the GRACE TWS (i.e. TWS-SWS) represents the GWS with bias and leakage errors, which can be evaluated and corrected by the forward modeling using the “true” GWS information. Generally, the “true” GWS information is derived from either models or in situ measurements [Scanlon et al., 2012; Feng et al., 2013]. In this study, we use the in situ GW-level measurements for both shallow and deep aquifers to represent the true data inside of the subregions, because the regional hydrologic model in and out of the NCP is not available. We do not choose the global LSMS because most of recent LSMSs do not include the GWS component, and the simulation would have larger uncertainties at smaller subregional scale. Even the sophisticated global hydrological model, the WGHM, failed to simulate reasonable groundwater storage changes in the NCP [Döll et al., 2014], due to failures to accounting for lateral flow and reasonable recharge for the piedmont region. Besides, the bulletin statistical data from the Groundwater Bulletin of China Northern Plains [Ministry of Water Resources of China (MWR), 2010] are used to represent the true data outside of the NCP, mainly in Shanxi, Henan and Shandong. Our experiments are focused on modeling the regional-averaged GWS depletion rate. Although the time span of these “true” data (2005-2010) is shorter than the GRACE data (2003-2013), it is among the period that the continuous depletion occurred, and these “true” data can well represent the actual differences on GWS exploitation between PP and ECP.

Firstly, we perform a forward modeling exercise to see how the 1-mm mass distribution over the subregions is represented at the GRACE solution (i.e. 1°× 1° grids, truncated at degree 60, filtered using 200-km Gaussian smoother, but no de-correlation filter is applied [Chen et al., 2014]). This hypothetic modeling task only considers a single region in PP or ECP which can be used to explore the leakage-out effects. As shown in Figure S1, the leakage-out effects are significant for the two
subregions. Nearly 80% signals leak out from PP, and ~70% leak out from ECP. The leakage-out from PP (ECP) is likely to partially become the leakage-in for ECP (PP). It is found that each mm of mass within the PP (ECP) region induces a 14% (22%) of leakage into the ECP (PP). It can be concluded from this task of 1-mm mass hypothetic modeling that the effects from the bias are significantly larger than that from the leakage for both PP and ECP.

Next, we deal with the real situation using the “true” data. As shown in Figure S2, we use the regional mean depletion rate (2005-2010) of GWS both from the interior and exterior to the subregions. The GWS depletion rates estimated from in situ GW-level measurements in PP and ECP are 50.8 and 13.8 mm/yr, respectively (Figure S2a). According to the information in the Groundwater Bulletin Statistics, the GWS depletion exterior to the entire NCP region mainly includes the Shanxi basins, Henan plains, and part of the Shandong province, of which the GWS depletion rate is 11.9, 7.6 and 2.0 mm/yr, respectively, mainly over the unconfined aquifers (Figure S2a). After the harmonic expansion using the grids with the GRACE solution, truncation at degree 60, and filtering using the 200-km Gaussian smoother, the “true” GWS depletion rate (regional average) in the PP is reduced to 12.6 mm/yr (by 75.2%) and that in ECP reduces to 11.8 mm/yr (by 14.5%) (Figure S2b). Most of the signals in the PP are leaked out mainly because of its smaller size and the more irregular boundary shape. The signal in ECP is less reduced since most of the biased signals in ECP are offset by the leakage from PP. Moreover, the difference on the GWS depletion rate between PP and ECP becomes insignificant after truncation and smoothing.

In this study, we further performed a “hypothetic” forward modeling using the area-weighted average of the GWS depletion rate in the two subregions. Figure S2f depicts the spatial difference of GWS depletion rate between the “true” (Figure S2a, S2b) and “hypothetic” (Figure S2c, S2d) forward
modeling approach. The difference for PP (ECP) is ~2.4 (1.0) mm/yr, ~11% (7%) of the actual total difference between Figure S2a and S2c (PP: ~22.7 mm/yr, ECP: ~14.3 mm/yr). The weak difference between the two forward modeling approaches indicates different spatial mass distribution may result in rather similar spatial patterns in the GRACE data.

Furthermore, the spatial distribution of GWS depletion derived from forward modeling (Figure S2b) differs from that from GRACE (Figure S2e). As seen, the spatial location of the maximum GWS loss does not agree well with each other. That is because the GRACE satellites see the mean over the region several times larger than the area of NCP. But the forward modeling is focused on the subregional scale in NCP with irregular shape. Besides, the prior information in the surrounding regions of NCP used for the forward modeling only includes some local shallow GWS depletion. But GRACE detects more heterogeneous mass variation, e.g. the mass loss in Shanxi due to the coal mining, and some deep GWS depletion in the surrounding provinces. Moreover, different GRACE products processed using different processing techniques would also lead to various spatial patterns of mass changes, as was shown by Feng et al. [2013]. Consequently, the spatial location of the maximum GRACE signals does not necessarily fit with the maximum water mass change on the earth.

To restore the amplitude of GRACE-derived GWS (i.e. TWS-SWS) in the subregions, a multiplicative factor aiming at correcting the combined effects of both bias and leakage is estimated as 4.04 for PP and 1.17 for ECP by calculating the ratio between the filtered and the unfiltered depletion rate for each subregion. Some caveats should be noted here. Generally, the bias error can be corrected using a multiplicative factor [Chen et al., 2005]. While a better choice for correcting the leakage error could be using the time series rather than a single factor since the human-induced GWS depletion can be time-dependent. This requires a full time span of either continuous model simulations or routine
ground-based observations. However, the model-simulated GWS data are not available around the NCP region at present and the time spans of available in situ measurements are limited, therefore the use of a single multiplicative factor is selected here, meaning that the “true” information taken from the in situ GW-level measurements and the bulletin statistics are fully trusted. More importantly, as shown in Table 1 in the manuscript, the GWS in PP and ECP shows similar annual phase. The leakage error from the adjacent subregion could be partly offset by the bias in the target subregion for their similar annual phase [Kless et al., 2007]. Therefore bias error accounts for the majority of the total error at subregional scale. That is to say the multiplicative factor would be mainly used to correct the large bias error. Besides, the multiplicative factors reveal the relative differences in two subregions, and it is worthwhile to explore how effective the multiplicative factors can be used to restore the signals in the subregions at least from the perspectives of regional average and long-term variability. It has already been turned out in this study that, with the uncertainties concerned, the corrected GRACE-derived GWS changes by using the multiplicative factor can be well compared with the in-situ-based GWS changes.

It should be noted that forward modeling is not trying to adjust the GRACE signal to match directly with in situ measurements. Instead, forward modeling takes into account the spatial and temporal variability of different water storage components from “true” prior information (generally LSMs or in situ measurements) by mimicking the processing of GRACE signal. Admittedly, there are some better methods. For instance, a latest method proposed by Chen et al. [2014] allows the leakage correction expressed as a time series. In the study of Chen et al. [2014], the terminology “leakage” error is referred to as the combined effects of both the “bias” and “leakage” we mentioned before. As for the method of Chen et al. [2014], the “true” GWS variation can be obtained by performing a fully
unconstrained global forward modeling for the monthly GRACE-derived GWS (i.e. GRACE TWS minus the modeled SWS that are truncated and filtered as the GRACE data were processed) based on successive iterations without using additional prior information.

The GRACE-derived GWS from using different data products are intercompared and are also compared with the in situ GWS measurements from 2005 to 2010 as shown in Figure S3 with the corresponding statistics given in Table S1. The results derived by using the CSR GRACE data show the highest correlation with the observations. The correlation in PP is generally larger than that in ECP. These differences due to different GRACE data indicate that the sensitivity of the selection of GRACE data products may not be a trivial issue for its applicability in different regions of the world. Further researches are warranted in the related significant issues.

Figure S1. The impact of a 1-mm water layer distributed on the two subregions of interest, the (up) PP and (down) ECP. Left: The full resolution. Right: The filtered mass distribution at the GRACE resolution (truncated at degree 60 plus 200-km Gaussian smoother). Please note the value of the color panels is different.

Figure S2. Trend maps of GWS depletion in the NCP derived from in situ GW-level measurements and GW bulletin statistics (or prior information, 2005-2010), forward modeling (2005-2010) and GRACE (2003-2013). (a) Regional-averaged “true” GWS depletion rate in PP (50.8 mm/yr) and ECP (13.8 mm/yr) from the prior information. (b) Spatial distribution of the “true” depletion rate on GRACE after performing the forward modeling (i.e. truncated at degree 60 and 200-km Gaussian smoother). (c) “Hypothetic” GWS depletion rate in the NCP (28.1 mm/yr, no spatial difference) using
the area-weighted average of the two subregional depletion rate from the prior information. (d) Spatial distribution of the “hypothetic” depletion rate on GRACE after performing the forward modeling. (e) The apparent GWS depletion rate in the NC region from GRACE. (f) The difference of the mass distribution on GRACE between the two forward modeling approaches using the “true” and “hypothetic” data, obtained by (b) minus (d). Please note the different value of the color panels.

Table S1. Comparisons of GWS depletion rate estimated based on different GRACE data products. The rates here are the linear fitted trends of the smoothed time series shown in Figure S2. The GRACE-derived rate here of has been adjusted by subtracting the impacts from reservoir regulation and water diversion, with a total increase rate of 5.7 mm/yr [Tang et al., 2013]. The GRACE-derived rates of GWS depletion (2003-2013) are compared with the in-situ-measured ones from 2005 to 2010, with the R-square correlation shown below.

Figure S3. Smoothed time series of GRACE-derived GWS changes using different GRACE data products in the PP and ECP from 2003 to 2013, with a comparison with the in-situ-based GWS changes from 2005 to 2010.

Section S3. Uncertainty estimation

The uncertainty in GWS is estimated from propagating various error components into the GWS time series. The following main error components [Swenson et al., 2003; Scanlon et al., 2012] are estimated in this study: the rescaled GRACE measurement error, reprocessing error (e.g., leakage error, filter error), hydrological model error, and the rescaling error. A conservative estimation of GRACE measurement errors is made following the method of Chen et al. [2009]. As the barotropic ocean mass
changes have been removed in GRACE de-aliasing process [Bettadpur, 2007], the true mass variability
(GRACE signal) over the ocean is probably near zero. The GRACE measurement error in this study
can be calculated from the RMS of the residual of the mass variation over the Pacific Ocean (35°-40°
N, the same latitude of the NCP). The estimated GRACE measurement error is 21.3mm from 2003 to
2013. After the rescaling, the GRACE measurement errors for PP and ECP are 85.9 and 24.9 mm/mo.,
respectively. As shown in Table 1 and Table S1, the GRACE measurement error is the largest
component of the estimated uncertainties.

Another major error source is from the hydrologic models. Previous studies (e.g. Long et al.
[2013]) have shown significant uncertainties in the SWS estimation using the GLDAS-1 model
simulations due to the differences in the forcing data, the version of model, and the spatial resolution
of LSMs. In this study, the stand deviation between five LSMs is estimated as the uncertainty of the
hydrologic models in PP and ECP. The rescaling error is estimated from the variability of the
multiplicative factors (~10%), which includes the uncertainty from the exact kernel function (<1% at
0.25° [Longuevergne et al., 2010]) and the uncertainty from the mass distribution of the “true” GWS
by using the spatial mean from the in situ GWS-level data and bulletin statistics. The rescaling error
for the PP is significantly larger because the bias effect in PP is more outstanding. The reprocessing
error in this study is estimated by calculating the standard deviation between the estimates of GRACE
TWS using different de-correlation methods (e.g. P3M10 and P5M6) and the Gaussian filters with
different radius ranging from 100 km to 1000 km. The final trend error for the GRACE-derived GWS
is calculated by propagating the different error components into GWS changes using the Monte Carlo
method, under the assumption that different water components are independent. The errors of TWS (i.e.
SWS+GWS) can be estimated from the mean square root of the sum of the square errors in SWS and
In this study, we also estimate the uncertainty in the in situ GW-level measurements, considering 10% uncertainty in the specific yield value. A conservative estimate of the error of in situ GWS is made by calculating the RMS of the residuals between the least-squares fitted and the original GWS changes. The errors of GRACE-derived GWS changes (2005-2010) are also estimated in this study so as to be better compared with the in situ GWS changes. As shown in Table S2, the monthly error for the 2005-2010 periods is much smaller than that estimated at the whole study time period from 2003 to 2013, but the trend error (2005-2010) is more significant in the PP after propagating the different error components into the GWS changes. This is normal, because different periods of time series are of different sensitivity to the error propagation.

**Table S2.** The monthly errors and trend errors of the GRACE-derived GWS variation in PP and ECP for the 2003-2013 periods and the 2005-2010 periods. The errors of GWS variation derived from in situ well measurements (2005-2010) is also shown below.

**Section S4. The least-Squares Harmonic Analysis (LSHA)**

There are some seasonal signals for the time series of TWS, SWS and GWS, e.g. the linear trend, the annual and semi-annual amplitudes and phases. We use the LSHA [Pawlowicz et al., 2002; Jin et al., 2013] to obtain these signals:

\[
\Delta H(t) = a + bt + \sum_{i=1}^{2} A_i \cos \left[ \frac{2\pi}{T_i} (t - \phi_i) \right] + \varepsilon(t),
\]

where \( \Delta H(t) \) is the fitted time series for the equivalent water height; \( t \) is the time; \( a \) and \( b \) are the constant, and the linear trend, respectively, \( A_i, T_i, \) and \( \phi_i \) represent amplitude, frequency and phase, respectively (\( i = 1 \): the annual, \( i = 2 \), the semi-annual); \( \varepsilon(t) \) is the residual.
Section S5. Published bulletin statistics of water consumption in the Hai River Basin (HRB) from 2003 to 2012

Table S3. Annual statistics of total water consumption (TWC), groundwater consumption (GWC) and the water consumption for irrigation in the HRB (319,651 km$^2$) from the Bulletin of Water Resources (2003~2012) published by the Water Resources Protection Bureau of HRB (km$^3$) [WRPB, 2012] (Volume: km$^3$).

Table S4. Total groundwater consumption (GWC) in the eight provinces within the HRB [WRPB, 2012] (Volume: km$^3$).

References:


