WILEY

RESEARCH ARTICLE OPEN ACCESS

Denoising Improves Cross-Scanner and Cross-Protocol Test–Retest Reproducibility of Diffusion Tensor and Kurtosis Imaging

Benjamin Ades-Aron[1,2] 🔵 | Santiago Coelho[1] 🔵 | Gregory Lemberskiy[2] | Jelle Veraart[1] | Steven H. Baete[1] | Timothy M. Shepherd[1] | Dmitry S. Novikov[1] | Els Fieremans[1]

[1]Bernard and Irene Schwartz Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, New York, New York, USA | [2]Microstructure Imaging Inc., Brooklyn, New York, USA

Correspondence: Benjamin Ades-Aron (benjamin.ades-aron@nyulangone.org)

ABSTRACT

The clinical translation of diffusion magnetic resonance imaging (dMRI)-derived quantitative contrasts hinges on robust reproducibility, minimizing both same-scanner and cross-scanner variability. As multi-site data sets, including multi-shell dMRI, expand in scope, enhancing reproducibility across variable MRI systems and MRI protocols becomes crucial. This study evaluates the reproducibility of diffusion kurtosis imaging (DKI) metrics (beyond conventional diffusion tensor imaging (DTI)), at the voxel and region-of-interest (ROI) levels on magnitude and complex-valued dMRI data, using denoising with and without harmonization. We compared same-scanner, cross-scanner, and cross-protocol variability for a multi-shell dMRI protocol (2-mm isotropic resolution, $b = 0$, 1000, 2000 s/mm^2) in 20 subjects. We first evaluated the effectiveness of Marchenko-Pastur Principal Component Analysis (MPPCA) based denoising strategies for both magnitude and complex data to mitigate noise-induced bias and variance, to improve dMRI parametric maps and reproducibility. Next, we examined the impact of denoising under different population analysis approaches, specifically comparing voxel-wise versus region of interest (ROI)-based methods. We also evaluated the role of denoising when harmonizing dMRI across scanners and protocols. The results indicate that DTI and DKI maps visually improve after MPPCA denoising, with noticeably fewer outliers in kurtosis maps. Denoising, either using magnitude or complex dMRI, enhances voxel-wise reproducibility, with test–retest variability of kurtosis indices reduced from 15%–20% without denoising to 5%–10% after denoising. Complex dMRI denoising reduces the noise floor by up to 60%. Denoising not only reduced variability across scans and protocols, but also increased statistical power for low SNR voxel-wise comparisons when comparing cross sectional groups. In conclusion, MPPCA denoising, either over magnitude or complex dMRI data, enhances the reproducibility and precision of higher-order diffusion metrics across same-scanner, cross-scanner, and cross-protocol assessments. The enhancement in data quality and precision facilitates the broader application and acceptance of these advanced imaging techniques in both clinical practice and large-scale neuroimaging studies.

Abbreviations: ABCD, adolescent brain cognitive development; ADNI, Alzheimer's disease neuroimaging initiative; CCC, concordance correlation coefficient; CV, coefficient of variation; DKI, diffusion kurtosis imaging; dMRI, diffusion magnetic resonance imaging; DTI, diffusion tensor imaging; HCP, Human Connectome Project; ICC, intraclass correlation coefficient; MPPCA, Marchenko-Pastur principal component analysis; ROI, region of interest.

---

1 | Introduction

Diffusion magnetic resonance imaging (dMRI) is a non-invasive method enabling clinicians and researchers to characterize tissue microstructure beyond the nominal resolution of MRI. It measures the restricted diffusion of water within tissues (Jones and Diffusion 2010), providing insights into neurological structure and function (Alexander et al. 2019; Jelescu and Budde 2017; Novikov, Kiselev, and Jespersen 2018). Recent large scale data initiatives like the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al. 2008), Adolescent Brain Cognitive Development (ABCD) (Jack Jr et al. 2008), the Human Connectome Project (HCP) (Van Essen et al. 2013), and UK Biobank (Sudlow et al. 2015), aim to utilize multi-shell dMRI to achieve a more comprehensive characterization of brain microstructure than possible with conventional diffusion tensor imaging (DTI). Higher-order diffusion signal representations and biophysical models (Alexander et al. 2019; Jelescu and Budde 2017; Novikov et al. 2019; Novikov, Kiselev, and Jespersen 2018) have been shown to characterize age related tissue changes in healthy populations (Beck et al. 2021; Cox et al. 2016; Novikov et al. 2019) and pathology stemming from disease (Dong et al. 2020; Falangola et al. 2013; de Kouchkovsky et al. 2016; Pogosbekian et al. 2021; Stauffer et al. 2021).

Despite its potential, the variability of dMRI-derived parameters, across different scanners and even on the same scanner, challenges its translation into clinical practice (Vollmar et al. 2010). This variability stems from factors like changing nuclear magnetic resonance (NMR) contrast (e.g., due to operating at different echo time or field strength), imaging artifacts like bias due to motion or Gibbs ringing, and MRI noise which introduces random fluctuations and bias from the non-central-χ MRI noise floor (Aja-Fernández, Tristán-Vega, and Alberola-López 2009; Gudbjartsson and Patz 1995). These artifacts, coupled with poor noise propagation through the high-order representations (such as diffusion kurtosis imaging [DKI] (Jensen et al. 2005)), lead to large outliers in quantitative maps (Sprenger et al. 2016), further reducing the reproducibility of such parameters. These variabilities limit dMRI to a qualitative rather than a quantitative modality, particularly as MRI technology itself rapidly advances, prompting frequent updates to study protocols (such as parallel imaging (Griswold et al. 2002; Sodickson and Manning 1997; Pruessmann et al. 1999) and simultaneous multislice [SMS] imaging (Setsompop et al. 2012)) that affect NMR contrast and noise floors. Low signal-to-noise-ratio (SNR) reduces the precision of dMRI and amplifies scanner- and protocol-dependent biases (Jones and Basser 2004), necessitating effective noise suppression strategies (Does et al. 2019; Manzano Patron et al. 2024) to enhance data reliability, harmonization, and enable more accurate comparisons of diffusion parameters between healthy and pathological tissues.

Denoising is a critical component of diffusion MRI (dMRI) preprocessing, aimed at minimizing noise while preserving the integrity of diffusion measurements. Principal component analysis (PCA)-based techniques, such as Marchenko-Pastur principal component analysis (MPPCA) (Veraart, Fieremans, and Novikov 2016; Veraart et al. 2016) and local PCA (LPCA) (Manjón et al. 2013), are widely employed to separate signal from noise by leveraging redundancy in diffusion data across spatial and *q*-space dimensions. Filter-based methods like BM4D (Maggioni et al. 2012) and non-local means (NLM) (Manjon et al. 2010) utilize non-local patch matching to enhance noise reduction while maintaining high fidelity to the underlying anatomical structures. More recently, machine learning approaches like Patch2Self (Fadnavis, Batson, and Garyfallidis 2020) and convolutional neural network (CNN) methods (Tian et al. 2020, 2022) have been applied to dMRI, leveraging neural networks to perform self-supervised denoising, allowing for more effective noise suppression without requiring external training data.

In this study, we evaluate two noise reduction techniques based on Random Matrix Theory—MPPCA for both magnitude and complex image data—to reduce within-scanner, cross-scanner, and cross-protocol variability. We compare *repeatability* (pure test–retest comparisons without changing protocol or hardware) and *reproducibility* (comparisons of same-subject data while varying protocol/hardware conditions) of three kinds of processing: (i) without denoising; (ii) using state-of-the-art MPPCA denoising (Veraart, Fieremans, and Novikov 2016, Veraart et al. 2016) for magnitude images; and (iii) adopting MPPCA to denoise complex images (Henriques et al. 2023; Lemberskiy et al. 2019), to maximize the precision of conventional DTI and higher-order diffusion metrics. MPPCA identifies an optimal threshold for noise-only singular values using the Marchenko-Pastur (MP) distribution. This technique and its variations have become a widely used component of dMRI pre-processing pipelines (Cordero-Grande et al. 2019; Moeller et al. 2021; Olesen et al. 2023). Complex MPPCA operates on complex-valued data (after coil combination), and is enabled through phase estimation and unwinding prior to applying the MPPCA algorithm (Cordero-Grande et al. 2019; Grussu et al. 2020).

We scanned 20 human subjects using in total five dMRI protocols on two different MRI scanners (same vendor but with different gradient hardware). This unique dataset includes for each subject scan-rescan measurements on both scanners using a dMRI protocol optimized for SNR by minimizing echo time for the specific hardware, as well as cross-scanner measurements using the same dMRI protocol on both scanners, with a dMRI protocol optimized for the scanner with the weaker gradients. By comparing denoising strategies across these measurements with varying SNR, we sought to determine how MPPCA denoising can mitigate SNR-related biases and improve the consistency of diffusion metric quantification, with the ultimate goal of enhancing reproducibility across varying hardware and imaging protocols.

2 | Methods

2.1 | Clinical MRI Protocol

This HIPAA-compliant prospective study was approved by the local institutional review board. After providing informed consent, 20 healthy volunteers (10 male/10 female, age $= 32.2 \pm 9.7$ years) underwent brain dMRI on Siemens Magnetom Prisma and Skyra 3T systems. All protocols included

TABLE 1 | The three dMRI protocols used during this study. Each of the 20 subjects underwent these three protocols to acquire in total five datasets, including a test–retest dataset on the Prisma, test–retest on the Skyra, and a single dataset acquired on the Prisma with protocol matched to the Skyra.

Scanner	Number of datasets	Dataset labels	TR (ms)	TE (ms)	$b = 0\,s/$ mm	$b = 1000\,s/$ mm^2	$b = 2000\,s/$ mm^2	Acq. time (min:s)	Gradient (mT/m)	Software version	Coils
Prisma	2	$P^{(1,2)}_{92}$	5300	92	5	20	40	7:28	80	VEC11	20 channel head
Prisma	1	$P^{(1)}_{127}$	6700	127	5	15	30	7:28	80	VEC11	20 channel head
Skyra	2	$S^{(1,2)}_{127}$	6700	127	5	15	30	7:27	40	VEC11	20 channel head

a diffusion weighted monopolar spin echo EPI sequence with 2-mm isotropic resolution, using a 20-channel head coil for reception and b-values commonly used in multi-shell acquisitions (Sudlow et al. 2015; Van Essen et al. 2013) including $b = 0$, 1000, and 2000 s/mm shells, as detailed in Table 1. Acceleration was performed using 6/8 partial Fourier, in-plane parallel imaging (GRAPPA factor 2), and simultaneous multislice (SMS) acceleration with multiband factor 2. Raw data from all acquisitions were saved in the Siemens TWIX data format and fed into a standard reconstruction pipeline implemented in MATLAB for the purpose of saving phase maps. The pipeline (in order of processing steps) included coil noise decorrelation (Brey and Narayana 1988), Nyquist ghost correction (Ahn and Cho 1987), slice-wise GeneRalized Autocalibrating partially parallel acquisitions (GRAPPA) (Griswold et al. 2002; Larkman et al. 2001), Controlled Aliasing In Parallel Imaging Results IN Higher Acceleration (CAIPIRINHA) shift (Breuer et al. 2005), in-plane GRAPPA, trapezoidal regridding (Bernstein, King, and Zhou 2004), adaptive combine (Walsh, Gmitro, and Marcellin 2000), and Projection onto Complex Sets (POCS) (Amartur and Haacke 1991). During this reconstruction, both magnitude and phase images were saved for use during subsequent denoising and artifact correction steps.

Five test–retest dMRI datasets were acquired for each subject, as listed in Table 1. Two test–retest datasets were acquired on the Prisma system using the dMRI protocol with the shortest possible TE = 92 ms. A third dataset was acquired on the Prisma using a modified protocol (with TE = 127 ms) to study cross-protocol reproducibility. Next, two datasets were acquired on the Skyra system using the dMRI protocol with the shortest possible TE = 127 ms (due to smaller maximal gradient). Scan time for all dMRI protocols (subset used here) was approximately the same at 7.5 min, with respectively 60 and 50 directions on the Prisma and Skyra systems.

Between test–retest acquisitions, subjects were removed and then placed back into the scanner. Each test–retest dataset included a reverse phase encoding $b = 0$ image to correct for EPI-induced distortions. A 1-mm isotropic T1-weighted Magnetization Prepared Rapid Gradient Echo (MP-RAGE) sequence (TR/TE/TI = 2200/3.17/900 ms) was acquired on the Prisma system for co-registration of all five datasets per subject.

These five datasets were used to perform three comparisons.

1. Effect of denoising on *within-scanner variability* (*repeatability*) by comparing Prisma versus rescan Prisma[2] with TR/TE 5300/92 ms ($P^{(1)}_{92}$ vs. $P^{(2)}_{92}$) and comparing Skyra[1] versus rescan Skyra[2] with TR/TE = 6700/127 ms, referred to as $S^{(1)}_{127}$ versus $S^{(2)}_{127}$.

2. Effect of denoising on *cross-protocol variability* (*reproducibility*) by comparing data from the same scanner with unmatched TE and SNR: Prisma[1] TR/TE = 5300/92 ms versus Prisma[1] TR/TE = 6700/127 ms, referred to as $P^{(1)}_{92}$ vs. $P^{(1)}_{127}$.

3. Effect of denoising on *inter-scanner variability* (*reproducibility*) by comparing data from different scanners, but matched TE: Skyra[1] TR/TE = 6700/127 ms versus Prisma TR/TE = 6700/127 ms, referred to as $S^{(1)}_{127}$ versus $P^{(1)}_{127}$.

2.2 | MPPCA-Based Denoising Methods

Denoising was performed using an augmentation of the MPPCA algorithm (Veraart, Fieremans, and Novikov 2016; Veraart et al. 2016). MPPCA exploits data redundancy in the singular value decomposition (SVD)/principal component analysis (PCA) domain using properties of the eigenspectrum of random covariance matrices, which yields an objective number p of signal-carrying components to be kept and all other components removed as purely noise-carrying.

A data matrix $X = \{X_{mn}\} \equiv \{S_m(\boldsymbol{r_n})\}$ of size $M \times N$ is formed by M measurements from N voxels in a patch. Let $M' = \min(M, N)$ and $N' = \max(M, N)$. Low-rank denoising corresponds to keeping p largest components in the SVD of $X = \sum_{i=1}^{M'} s_i u_i \otimes v_i$, where u_i and v_i are the left and right singular vectors. Equivalently, PCA corresponds to keeping p top eigenvalues $x_i = s_i^2$ explaining most of the variance of the (sample) covariance matrix $XX^\top$. MPPCA yields the number p of the components of X to keep, by identifying the MP distribution formed by the remaining $M' - p$ components that correspond to pure noise in the limit $M, N \gg 1$ and $p \ll M, N$ (low-rank condition). MPPCA self-consistently finds p, and noise variance σ^2 as the sum over the $M' - p$ components attributed to the MP distribution.

In recent years there having been several proposed additions to the original MPPCA algorithm, including symmetric thresholding of singular values (Olesen et al. 2023), eigenvalue shrinkage (Gavish and Donoho 2017), advanced patching methods including nonlocal spatial matching (Buades, Coll, and Morel 2005; Maggioni et al. 2012; Manjón, Coupé, and Buades 2015; Zhao et al. 2022), angular matching (St-Jean, Coupé, and Descoteaux 2016), and structural adaptation (Bao et al. 2013). Here, MPPCA is augmented in three stages: (i) Adaptive patching to select the signals forming a data matrix X around a given voxel; (ii) Symmetric SVD threshold selection (Olesen et al. 2023); (iii) Singular value shrinkage (Gavish and Donoho 2017). Here we carefully chose augmentations to MPPCA to balance spatial redundancy, signal similarity, and noise reduction efficiency. For additional information about available tunable parameters in MPPCA, users can view the software documentation available online at https://nyu-diffusionmri.github.io/DESIGNER-v2/docs/designer/background. We here describe each stage:

i. *Adaptive patching*—In contrast to the original MPPCA approach where patches are squares or cubes around a given voxel (e.g., $5 \times 5 \times 5$ voxels), we enhanced the spatial redundancy, in addition to the redundancy in the measurements (here in the diffusion q-space), by pre-selecting voxels that have similar ground truth. Our goal is to minimize the number p of components, and to maximize their contributions s_i, such that they describe most of the signal—this is the assumption of the underlying noise-free signal to be of low-rank. The ultimate best choice of the patch would be having all N voxels with the same ground truth; in this case, $p \equiv 1$ (the rank of noise-free $XX^\top$ is 1), irrespective of the complexity of their signal S_m. The desire to have a few large s_i, $i = 1, \ldots, p$ (as opposed to many smaller ones) comes from the fact that components whose noise-free singular values $s_i^{(0)} < s_* = \sigma(MN)^{1/4}$ are below the threshold s_* become indistinguishable from noise. They fall below

what's referred to as the *phase transition* (Johnstone 2007). By selecting a subset of voxels that have similar underlying tissue priors, we intend to overcome the information loss associated with many components potentially falling under the phase transition threshold s_* if all voxels in a patch were to have different signals.

Hence, for a voxel at $\boldsymbol{r_0}$, we would like to include the signals $S_m(\boldsymbol{r_n})$ such that both S_m are close to each other for different $\boldsymbol{r_n}$, and the voxels $\boldsymbol{r_n}$ are not too far from $\boldsymbol{r_0}$. Formally, we introduce the "distance" between signals

$$w_{\alpha, \beta}[S(\boldsymbol{r}), S(\boldsymbol{r'})] = |\boldsymbol{r} - \boldsymbol{r'}|^\alpha \cdot |S(\boldsymbol{r}) - S(\boldsymbol{r'})|^\beta \tag{1}$$

Here $|\boldsymbol{r} - \boldsymbol{r'}|$ is the Euclidean distance between voxels in three dimensions, and $|S(\boldsymbol{r}) - S(\boldsymbol{r'})| = \sqrt{\sum_m |S_m(\boldsymbol{r}) - S_m(\boldsymbol{r'})|^2}$ is the Euclidean distance between signals (the norm over the measurement index m). The balance of preferring the distance between voxels and between signals is tuned by the choice of exponents α and β. Here we choose $\alpha = 1$ and $\beta = 2$ based on an empirical observation of improved denoising performance when $\beta > \alpha$ (preferring similarity of signals to the distance from $\boldsymbol{r_0}$). When $\alpha \gg \beta$ this method converges to the original MPPCA patching implementation (local signal-independent patch around $\boldsymbol{r_0}$). Based on the above distance function, we choose a patch around voxel at $\boldsymbol{r_0}$ as N voxels (including $\boldsymbol{r_0}$), for which $w_{1,2}[S(\boldsymbol{r_n}), S(\boldsymbol{r_0})]$ is the smallest; here N was fixed to 100, Figure 1, selected from each larger $7 \times 7 \times 7$ cubic patch.

ii. *Symmetric thresholding*—For the pure-noise case of $p = 0$, all components of $XX^\top$ form the MP distribution, which has two independent properties: $\sigma^2 = \sum x_i / (MN)$ and $\sigma^2 = (x_1 - x_{M'}) / 4\sqrt{MN}$. For a general case (noisy signal), MPPCA uses these two properties to define two functions $\sigma_{1,2}(p)$ accounting for the noise variance from the bottom $M' - p$ components, with p being the solution for $\sigma_1(p) = \sigma_2(p)$. Here, we use the following definitions (Cordero-Grande et al. 2019; Olesen et al. 2023):

$$\sigma_1^2(p) = \frac{1}{(N' - p)(M' - p)} \sum_{i=p+1}^{M'} x_i \tag{2.1}$$

$$\sigma_2^2(p) = \frac{x_{p+1} - x_{M'}}{4\sqrt{(N' - p)(M' - p)}} \tag{2.2}$$

These are symmetric in M' and N', whereas the original MPPCA formulation had N' instead of $N' - p$ in Equation (2). This symmetric augmentation of $\sigma_{1,2}(p)$ empirically provides a more robust estimation of MP threshold for not very large N and M, and works well for $M \approx N$. It is also practically important that the patch size is allowed to vary, with possibilities of both $M > N$ and $M < N$. In this study N was fixed at 100, and M was the total number of diffusion measurements for all subjects and scan–rescan datasets. $M = 65$ for $P_{92}^{(1,2)}$, $M = 50$ for $S_{127}^{(1,2)}$, and $M = 50$ for $P_{127}^{(1)}$.

iii. *Singular value shrinkage*—To overcome the eigenvalue repulsion due to noisy components, we reduce the sample singular values s_i when recombining selected principal components into a low-rank matrix $\widehat{X}_\eta = \sqrt{M'}\sigma \sum_{i=1}^{p} \eta\left(s_i / \left(\sqrt{M'}\sigma\right)\right) u_i \otimes v_i$. According to
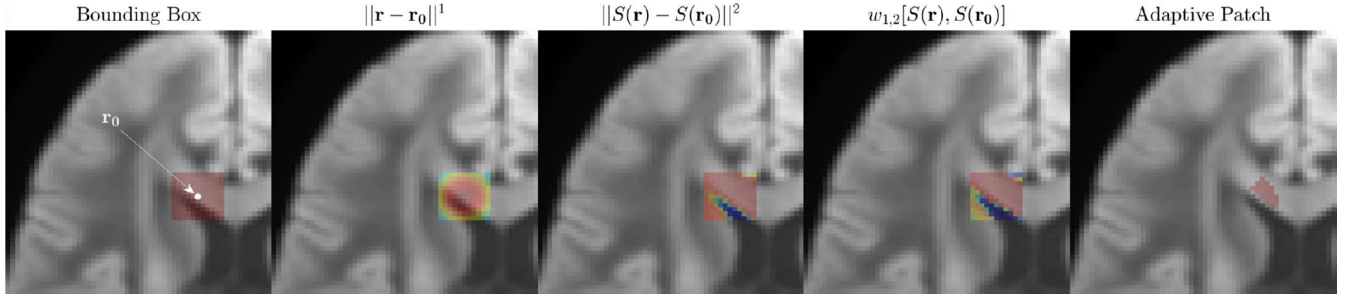
FIGURE 1 | Example of how distances $w_{\alpha,\beta}[S(r_n), S(r_0)]$ Equation (1), are generated for adaptive patching. These distances are used to threshold a cubic patch into the shape of local anatomy. We first choose a sufficiently large bounding box (in this example an $11 \times 11 \times 11$ bounding box was chosen for demonstration purposes, in the study a $7 \times 7 \times 7$ bounding box was used). Next, the Euclidean distances between voxels (2nd panel) and signals (3rd panel) are calculated. The weights (1) are computed based on the product of the two distances, tuned by exponents α and β. The patch selection corresponding to $N = 100$ is shown in the last panel.

Gavish and Donoho (2017), the optimal shrinker function based on minimizing MSE of the Frobenius norm can be expressed as

$$\eta(s) = \begin{cases} \frac{1}{s}\sqrt{\left(s^2 - \gamma - 1\right)^2 - 4\gamma}, & s > 1 + \sqrt{\gamma} \\ 0, & s \leq 1 + \sqrt{\gamma} \end{cases} \quad (3)$$

where $\gamma = M'/N' \leq 1$.

2.3 | Denoising Experiments

dMRI raw data was preprocessed in three different ways for comparison:

i. Magnitude (non-denoised) dMRI data.

ii. MPPCA: Magnitude dMRI data were MPPCA-denoised (including the adaptive patch, symmetric thresholding, eigenvalue shrinkage).

iii. MP-Complex: dMRI phases were first denoised using MPPCA (symmetric thresholding) using a 15×15 2D box-shaped kernel within each slice. The denoised and spatially smoothly varying phase $\phi_{\mathrm{dn}}(\boldsymbol{r})$ is then unwound according to: $S_{\mathrm{real}} = \mathrm{Re}\left(S_{\mathrm{complex}} e^{-i\phi_{\mathrm{dn}}}\right)$. Finally, the noisy phase-unwound signal S_{real} is denoised using an adaptive 3D moving patch, symmetric thresholding, and eigenvalue shrinkage. Phase unwinding helps remove spurious components arising due to shot-to-shot phase variations in dMRI (Grussu et al. 2020; Lemberskiy et al. 2019). In addition to improved denoising performance (due to the two-pass approach), MP-Complex also reduces the Rician noise floor of the data, reducing bias and increasing the precision of downstream parameter estimation.

2.4 | Processing Pipeline and Diffusion Parameter Estimation

All dMRI images were processed using the DESIGNER pipeline (Ades-Aron et al. 2018; Chen et al. 2024), using Partial Fourier induced Gibbs artifact correction (Kellner et al. 2016; Lee, Novikov,

and Fieremans 2021), EPI distortion correction (Smith et al. 2004), and eddy-current and motion correction (including slice-wise outlier replacement) (Andersson et al. 2016; Andersson and Sotiropoulos 2016). The signal's normalized rotational invariants of zeroth (S_0, also known as spherical mean) order were linearly estimated for each diffusion shell (Novikov, Veraart, et al. 2018) to provide a convenient basis for population-wise registration.

Diffusion and kurtosis tensors were estimated through unconstrained weighted linear least squares (Veraart et al. 2013) from which we derived mean diffusivity (MD), axial diffusivity (AD), radial diffusivity (RD), fractional anisotropy (FA), mean kurtosis (MK), axial kurtosis (AK), and radial kurtosis (RK). DTI parameters (e.g., FA and MD) were derived specifically from a DTI fit, utilizing only the $b = 0$ and $b = 1000\,\mathrm{s/mm}$ (Jelescu and Budde 2017) diffusion shells. These parameters are not derived from the DKI fit, where additional higher b values are included. This distinction is important, as the reproducibility of DTI parameters derived from a DTI-specific fit may differ from those derived from a DKI fit. For completeness we also show ROI means for DTI parameters derived from the DKI fit in Figure S4.

The minimal protocol used in this study uses kurtosis estimation from the cumulant expansion up to $\mathcal{O}(b^2)$:

$$\ln \frac{S(b, \hat{\boldsymbol{n}})}{S|_{b=0}} = -bD(\hat{\boldsymbol{n}}) + \frac{1}{6}b^2 \overline{D}^2 W(\hat{\boldsymbol{n}}) + \ldots \quad (4)$$

where $D(\hat{\boldsymbol{n}}) = D_{ij} n_i n_j$, $W(\hat{\boldsymbol{n}}) = W_{ijkl} n_i n_j n_k n_l$, and

$$MD \equiv \overline{D} = \int\limits_{|\hat{\boldsymbol{n}}|=1} \frac{d\hat{\boldsymbol{n}}}{4\pi} D(\hat{\boldsymbol{n}}) = \frac{1}{3}D_{ii} \quad (5)$$

(the Einstein's convention of summation over pairs of repeated indices is assumed throughout).

The definition of mean kurtosis is ambiguous in the dMRI literature. The original paper of Jensen et al. (2005) suggested that MK is an angular average of the directional kurtosis $K(\hat{\boldsymbol{n}}) = \overline{D}^2 W(\hat{\boldsymbol{n}})/D^2(\hat{\boldsymbol{n}})$:

$$MK = \overline{D}^2 \int\limits_{|\hat{\boldsymbol{n}}|=1} \frac{d\hat{\boldsymbol{n}}}{4\pi} \frac{W(\hat{\boldsymbol{n}})}{D^2(\hat{\boldsymbol{n}})} \quad (6)$$

This definition emphasizes the directional character of the kurtosis as opposed to that of the cumulant, However, this definition has two notable drawbacks. First, fundamentally, the result cannot be compactly represented as a trace of a certain tensor (e.g., analogously to Equation 5)—due to a non-trivial directional dependence of the denominator. Second, practically, this definition leads to a relatively low precision and is strongly affected by outliers, which come from directions where diffusivity $D(\hat{n}) \ll \overline{D}$ is small.

In this work, we adopt the definition that involves angular average of the cumulant (Hansen et al. 2013):

$$\text{MW} \equiv \overline{W} = \int\limits_{|\hat{n}|=1} \frac{d\hat{n}}{4\pi} W(\hat{n}) = \frac{1}{5} W_{iikk} \qquad (7)$$

The last equation follows from the equivalence between symmetric trace-free tensors and spherical harmonics, cf., e.g., appendix C of Novikov, Veraart, et al. (2018) $\mathcal{O}(b^2)$, and has been used in a number of earlier works (Hansen et al. 2013; Jespersen et al. 2018; Lu et al. 2006). Hence, instead of performing an integral over a limited number of directions, we can calculate $\overline{W}$ very fast and exactly from the estimated W tensor by simply taking a full trace. This quantity is dimensionless (as is MK), and has qualitatively similar contrast to MK. The axial and radial kurtoses, AW and RW, are calculated as projections onto the principal fiber direction and onto the plane transverse to it, respectively:

$$\text{AD} = D(\hat{v}_1)$$
$$\text{RD} = \int\limits_{|\hat{n}|=1} \frac{d\hat{n}}{2\pi} D(\hat{n}) \delta(\hat{n} \cdot \hat{v}_1)$$
$$\text{AW} = W(\hat{v}_1) \qquad (8)$$
$$\text{RW} = \int\limits_{|\hat{n}|=1} \frac{d\hat{n}}{2\pi} W(\hat{n}) \delta(\hat{n} \cdot \hat{v}_1)$$

While MK, AK, and RK are more commonly used in literature than MW, AW, and RW, both parameter definitions will be generated in what follows, owing to the lower probability for outliers when using W compared with K (as shown in Figure S3).

2.5 | Data Analysis and Statistics

To evaluate the effect size of the denoising step in preprocessing on any dMRI parameter x used in this study, we assess the variability across pairs of scans by the coefficient of variation $\text{CV}_x = \sigma_x / \mu_x$, where the estimate of the mean is $\hat{\mu_x} = \frac{1}{2}(x_1 + x_2)$, and the estimate of absolute error $\hat{\sigma_x} = \frac{\sqrt{\pi}}{2}|x_1 - x_2|$. Here $x_{1,2}$ refer to the parameter estimates from the two scans being compared. Test–retest variability was compared within scanner, across scanners, and across echo times, on a region of interest (ROI) level, and using a voxel-wise approach. In addition, concordance analysis and the concordance correlation coefficient (CCC) (Lawrence and Lin 1989) were used to assess reproducibility and agreement between scans.

Voxel-wise analysis: For each subject, a multimodal template was generated based on S_0 rotational invariant maps and FA maps from each of the five repeated acquisitions. The subject-wise

template was generated with Mrtrix3 (Tournier et al. 2019), using rigid registrations to a midpoint space, resulting in template maps for each of the 20 subjects. Next, a population template was generated based on all 20-subject template S_0 and FA pairs using nonlinear deformations (again using Mrtrix3). Rigid transformations from original space and warps to population space were concatenated, and then parametric maps for each subject, scan, and denoising level were transformed to the population template using cubic spline interpolation. For voxel-wise analyses, CV were computed in the space of subject templates (after transforming parameters using rigid transforms) and for Figure 3 only, in the space of the unbiased population midpoint space (by transforming parameters using nonlinear transforms). Interpolation is applied during both rigid and nonlinear transformations, as cubic spline interpolation is applied when transforming parametric maps to the population template. However, the main distinction lies in the fact that nonlinear deformations introduce small additional distortions due to the nature of the warps, which may slightly confound interpretations in comparison to rigid transformations. We performed a concordance analysis showing the degree of concordance correlation between test–retest pairs in pooled voxels over all subjects and show Bland-Altmann plots to measure the error between the same pairs.

Harmonization comparison: The effect of harmonization was compared to denoising for data with bias due to varying echo times (and noise floors). Harmonization was performed using the state-of-the-art linear-RISH (Mirzaalian et al. 2016) method. Linear-RISH uses mean rotationally invariant spherical harmonic (RISH) features over a population to normalize those of each individual, and after normalization, comparisons were performed in subject's native space. All transformations were computed once and applied to all datasets to eliminate registration induced error. Denoising was performed on both magnitude and complex data (Figure 7).

We used the A-1 (case 2) intraclass correlation coefficient (ICC; McGraw and Wong 1996) over all brain voxels to assess test–retest reliability in this study. This form of ICC measures the absolute agreement among measurements taken on randomly selected subjects, making it well-suited for comparing results across different scanners and protocols where variability in acquisition settings is expected. The choice of ICC (A-1) allows us to account for random effects and evaluate how consistently the same subjects' diffusion metrics are measured under varying conditions, ensuring that observed differences are not driven by systematic biases in hardware or protocols. We compared harmonization without denoising, denoising alone, and using both harmonization and denoising.

ROI analyses were performed in the original space of each scan (to minimize registration bias in CV measurements) by computing a nonlinear registration between the JHU white matter (WM) atlas (Hua et al. 2008) and the overall population template, and applying the inverse warps used to transform each subject map to the population average. Analyzed WM ROIs include genu, body, and splenium of the corpus callosum, internal capsule, and corona radiata. Since the original JHU ROIs (Hua et al. 2008) are quite large, we thresholded them in template space (not in subject space) by removing voxels in lowest 5th percentile of FA to

TABLE 2 | Mean coefficients of variation (CV) for each DTI and DKI parameter and denoising type evaluated in four regions: Splenium of the corpus callosum, posterior limb of the internal capsule (PLIC), anterior corona radiata (ACR), and thalamus. We show CVs for each of the three comparisons: Within-scan variability ($P_{92}^{(1)}$ vs. $P_{92}^{(2)}$ and $S_{127}^{(1)}$ vs. $S_{127}^{(2)}$), cross-protocol variability ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), and cross-scanner variability ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$).

| | | PRISMA$^{(1)}_{92}$ PRISMA$^{(2)}_{92}$ | | | SKYRA$^{(1)}_{127}$ SKYRA$^{(2)}_{127}$ | | | PRISMA$^{(1)}_{92}$ PRISMA$^{(1)}_{127}$ | | | PRISMA$^{(1)}_{127}$ SKYRA$^{(1)}_{127}$ | | |
		Complex	Magnitude	None	Complex	Magnitude	None	Complex	Magnitude	None	Complex	Magnitude	None
SPLENIUM (CC)	MD	0.056	0.060	0.089	0.103	0.102	0.158	0.174	0.148	0.180	0.224	0.197	0.241
	AD	0.045	0.048	0.097	0.078	0.075	0.149	0.115	0.102	0.165	0.143	0.130	0.199
	RD	0.149	0.157	0.218	0.291	0.289	0.374	0.382	0.327	0.392	0.515	0.451	0.495
	FA	0.051	0.052	0.079	0.075	0.076	0.102	0.130	0.110	0.138	0.159	0.148	0.167
	MW	0.042	0.041	0.093	0.105	0.084	0.196	0.104	0.149	0.173	0.121	0.116	0.218
	AW	0.065	0.061	0.106	0.094	0.071	0.141	0.102	0.193	0.187	0.111	0.112	0.172
	RW	0.123	0.148	0.429	0.308	0.254	0.909	0.248	0.298	1.240	0.282	0.303	1.084
PLIC	MD	0.034	0.035	0.069	0.051	0.048	0.102	0.051	0.053	0.089	0.061	0.058	0.105
	AD	0.039	0.039	0.092	0.051	0.053	0.132	0.058	0.062	0.117	0.058	0.060	0.135
	RD	0.066	0.069	0.136	0.115	0.115	0.214	0.102	0.104	0.183	0.142	0.144	0.227
	FA	0.045	0.046	0.092	0.068	0.071	0.116	0.068	0.070	0.120	0.082	0.086	0.127
	MW	0.026	0.027	0.070	0.045	0.039	0.116	0.042	0.091	0.110	0.051	0.042	0.115
	AW	0.054	0.047	0.111	0.069	0.057	0.139	0.072	0.130	0.162	0.076	0.065	0.140
	RW	0.058	0.060	0.221	0.101	0.091	0.519	0.081	0.092	0.405	0.113	0.102	0.547
ACR	MD	0.039	0.041	0.074	0.069	0.067	0.121	0.063	0.064	0.100	0.079	0.074	0.127
	AD	0.045	0.045	0.098	0.072	0.065	0.145	0.072	0.071	0.126	0.076	0.070	0.147
	RD	0.062	0.064	0.112	0.107	0.104	0.177	0.089	0.091	0.162	0.122	0.116	0.191
	FA	0.081	0.083	0.152	0.116	0.115	0.198	0.113	0.109	0.223	0.130	0.130	0.197
	MW	0.031	0.029	0.074	0.065	0.050	0.128	0.058	0.113	0.133	0.070	0.059	0.131
	AW	0.053	0.045	0.107	0.081	0.058	0.136	0.079	0.158	0.167	0.089	0.064	0.141
	RW	0.059	0.057	0.188	0.115	0.092	0.421	0.103	0.086	0.364	0.119	0.100	0.450

(Continues)

TABLE 2 | (Continued)

		PRISMA$^{(1)}_{92}$ PRISMA$^{(2)}_{92}$			SKYRA$^{(2)}_{127}$ SKYRA$^{(1)}_{127}$			PRISMA$^{(1)}_{92}$ PRISMA$^{(1)}_{127}$			PRISMA$^{(1)}_{127}$ SKYRA$^{(1)}_{127}$		
		Complex	Magnitude	None	Complex	Magnitude	None	Complex	Magnitude	None	Complex	Magnitude	None
Thalamus	MD	0.058	0.060	0.092	0.088	0.084	0.146	0.116	0.117	0.161	0.129	0.123	0.182
	AD	0.059	0.060	0.116	0.083	0.076	0.167	0.110	0.105	0.207	0.116	0.107	0.196
	RD	0.074	0.077	0.117	0.121	0.114	0.191	0.149	0.150	0.226	0.176	0.167	0.239
	FA	0.113	0.119	0.236	0.151	0.144	0.260	0.161	0.163	0.389	0.175	0.168	0.300
	MW	0.055	0.052	0.132	0.100	0.086	0.221	0.112	0.182	0.218	0.115	0.117	0.238
	AW	0.073	0.064	0.128	0.109	0.088	0.151	0.108	0.207	0.185	0.123	0.117	0.178
	RW	0.092	0.092	0.296	0.142	0.124	1.116	0.186	0.157	0.622	0.179	0.158	0.884

Note: The color shadings represent the minimum CV value (green), median CV (white), and maximum CV (red) over all comparisons.

minimize misregistration and partial-volume effects. All these ROIs were combined into a single large WM region when tabulating full WM scale statistics. Gray matter (GM) ROIs were derived using Freesurfer (Desikan et al. 2006), and CVs were computed in the thalamus, as a representative gray matter region due to its key role in sensorimotor integration and its structural connections to various white matter tracts, which are critical in diffusion MRI studies. Additionally, the thalamus is a large central region and less prone to artifacts due to partial volume effects and EPI induced distortion. Freesurfer was performed using the MP-RAGE acquired for each subject as an input, and ROIs were rigidly registered to diffusion space using the preprocessed b0 image as a registration reference. When taking ROI means, outliers in kurtosis metrics were thresholded prior to aggregating statistical metrics by excluding pixels with kurtosis < -1 or kurtosis > 10, and when quantifying the prevalence of outliers, the same range was used to determine the percentage of outlier voxels in a given white matter region (Figure S3).

Noise floor analysis was performed in diffusion weighted images to measure the level of bias reduction that comes from unwinding the denoised phase in complex data. We measured the noise floor by computing the spherical mean signal over directions $\overline{S}|_{b=2000}$ in the ventricles at $b = 2000 \, \text{s/mm}$ (Jelescu and Budde 2017). Due to diffusion attenuation, there is negligible signal ($\sim e^{-6} \approx 0.002$) left in CSF at this gradient strength, therefore $\overline{S}|_{b=2000} / \sqrt{\pi/2}$ can serve as a noise level estimator for a Rayleigh-distributed random variable. Ventricular segmentation was performed using Freesurfer as described above. The noise floor was further normalized by the signal $S_0 \equiv S|_{b=0}$ at $b = 0$:

$$\text{Noisefloor} \equiv \frac{\sigma_{noise}}{S_0} = \frac{1}{\sqrt{\pi/2}} \frac{\overline{S}|_{b=2000}}{S|_{b=0}} \quad (9)$$

Power analysis was performed to test the number of subjects necessary to detect a 5% effect size based on group means and standard deviations observed in this healthy control population. A 5% effect size was chosen based on Cohen's d values reported in literature (Kochunov et al. 2022), where for various psychiatric disorders grou pwise differences in mean FA could range from 0% to 42%. To compute sample size, we used the formula:

$$n_1 = \frac{2(Z_{\alpha/2} + Z_\beta)^2 \sigma_p^2}{(\mu_{x1} - \mu_{x2})^2} \quad (10)$$

Here n_1 is the sample size for a single group, $\mu_{1,2}$ refers to the means of the two groups being compared, we assume that the variance in measurements for the test and retest data are the same, and σ_p^2 is the sum of test–retest variance and intersubject variance,

$$\sigma_p^2 = \text{var}(x_1) + \overline{\sigma}^2 \quad (11)$$

where $\overline{\sigma}$ is the population averaged σ_x (true variance without the effect of measurement noise), and the variance of test measurements over the population (for N subjects) is $\text{var}(x_1) = \frac{1}{N} \sum_{i=1}^{N-1} (x_{i1} - \mu_{x1})^2$. Z_α is the z score corresponding to

significance level $\alpha/2$, and Z_β the critical value corresponding to the power of the test β (Kim 2016). This simulated the sample size required to reach voxel-wise statistical significance in a two-sided t-test with 80% power and $\alpha = 0.05$. We created two groups whose diffusion and kurtosis parameters differed by 5% and measured the number of subjects required to reach statistical significance. This analysis was performed for low-SNR voxel-wise data using test–retest CVs from Table 2 and high-SNR ROI-wise data using test–retest CVs from Table S1. For both comparisons intersubject variance was computed from ROI-wise data and had a range from 0.03 to 0.04.

3 | Results

Figure 2 shows representative diffusion and kurtosis maps for a single subject (26-year-old female) for test–retest data acquired on the Prisma scanner for MD, MK, MW, and FA. Two effects are notable from these maps: (1) The improved qualitative denoising effect from no denoising, to magnitude, to complex approaches. (2) The large decrease in kurtosis outliers observable as "black voxels" in MK maps both due to denoising, and by moving to the MW representation. This effect is also shown in Figure S3, where we show the reduction in outliers in white matter by switching from K to W representations, here the total percentage of outliers was reduced from over 70% on the Skyra system from MK to under 10% for MW.

3.1 | Voxel-Wise Variability

Voxel-wise test–retest error maps averaged over subjects are shown in Figure 3, for MD, FA, MW, AW, and RW (corresponding voxel-wise maps for kurtosis are provided in Figure S1). Voxel-wise CV is increased in regions near tissue boundaries because of noise limiting the precision of registration across subjects. However, this effect is reduced when using denoised dMRI, as visible in the MW-map around the genu (Figure 3). We show σ_x in Figure 3. rather than CV (σ_x / μ_x) due to the potential instability when normalizing by small mean values. In regions where the mean (μ_x) approaches zero, the CV can become disproportionately large, leading to misleading interpretations of variability. By focusing on σ_x, we provide a more stable and interpretable measure of dispersion that avoids these issues, particularly in low-signal voxels.

To minimize the effect of misregistration, Table 2 lists the mean voxel-wise CV pooled across different subjects within the splenium, posterior limb of the internal capsule (PLIC), anterior corona radiata (ACR), and thalamus, respectively. Regional CVs for kurtosis (K instead of W) including outliers are also shown in supplemental Figure S2. W-parameters offer better robustness to outliers (see Figure S3), such that without denoising, MW in the internal capsule has voxel-wise CV on the order of 9%–12% and denoising lowers test–retest variability on the Prisma system down to 3%–4%. FA has variability on the order of 7%–10% without denoising in highly aligned WM regions, and denoising lowers variability to 4%–6% for both Prisma and Skyra data.

The largest benefit of denoising and lowering the noise floor presents when comparing different echo times (Table 2). In AW we observe decreases in CV from 16.2% to 7.2% in the PLIC with complex denoising, along with corresponding large reductions in CV in other remaining WM areas. In line with the maps shown in Figure 3, we observe that CVs are consistently the lowest in areas such as the internal capsule, that are less prone to partial volume effects and Gibbs ringing. Indeed, regions near the ventricles such as the corpus callosum often have relatively higher variability.
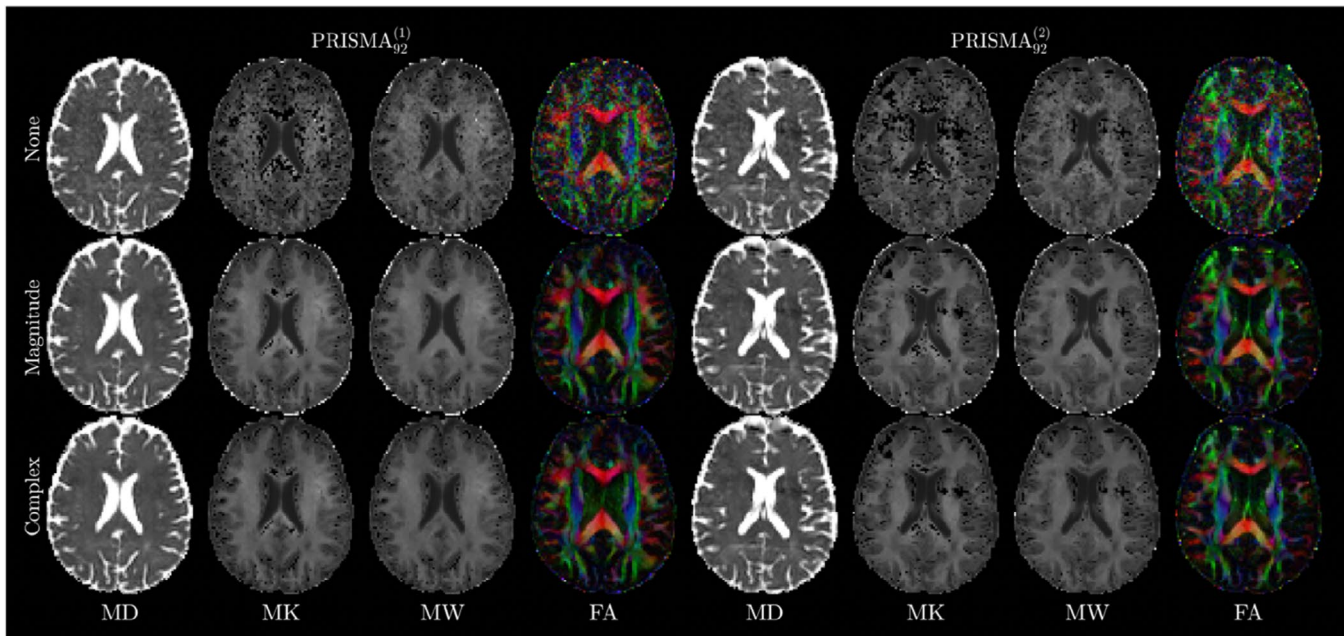

FIGURE 2 | Illustration of different denoising methods on DTI and DKI measures (MD, MK, MW, and FA) across scan-rescan data on the Prisma for a representative subject (26-year-old female). None: Non-denoised data, Magnitude: MPPCA denoised magnitude data, Complex: MPPCA denoised complex data. Rician bias reduction is evident in the increased contrast of FA images in complex denoised data. Outliers were reduced in kurtosis maps both through denoising and the use of W representation.
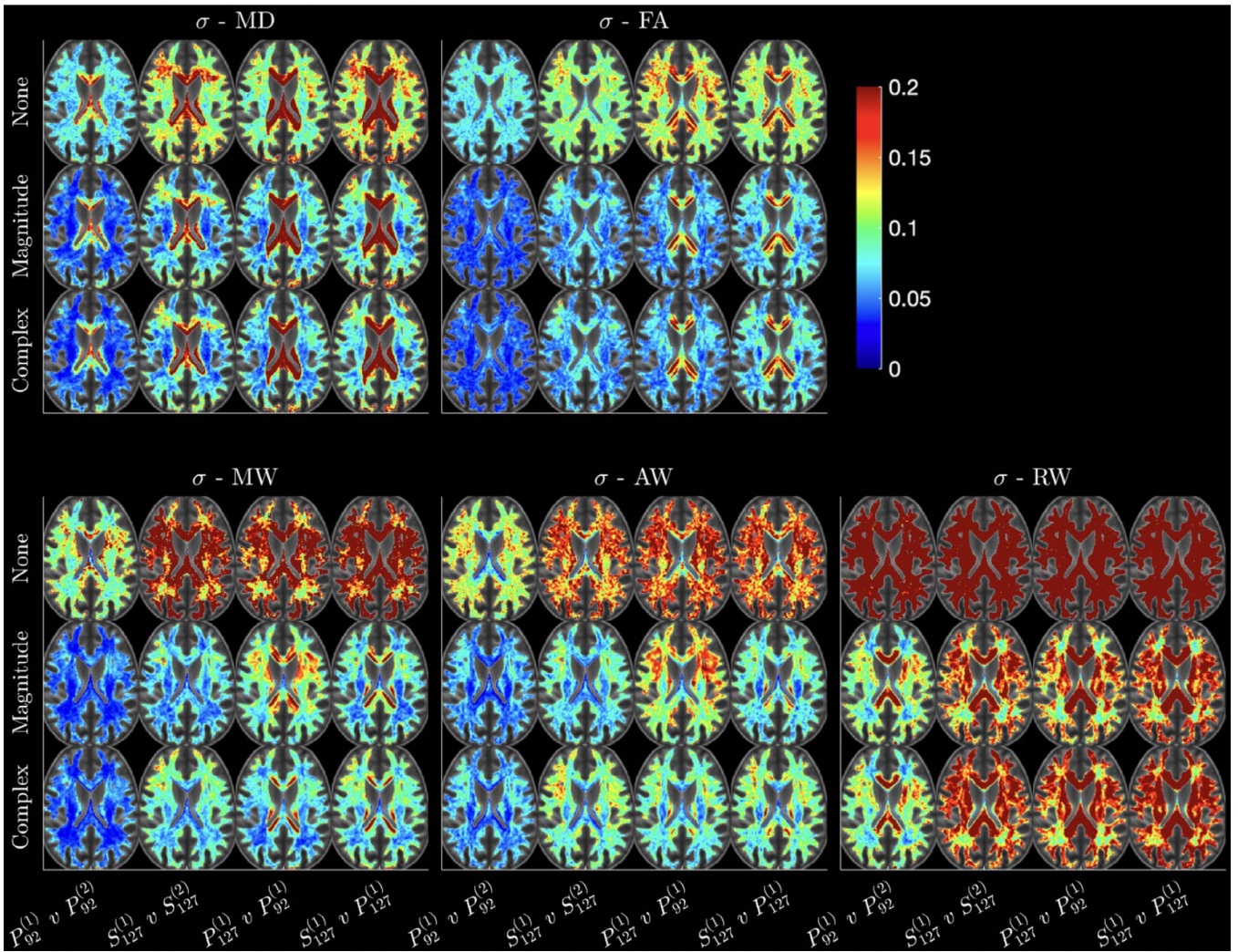
FIGURE 3 | Maps of the absolute error σ_x, created by computing σ_x in each subject's space, followed by warping into a common space and averaging over all 20 subjects. White matter masks were computed based on an FA threshold > 0.2 in the overall population template FA image. While interpolation affects both rigid and nonlinear transformations, the effect is more pronounced in the case of nonlinear warps, and thus, the error computed after nonlinear transformations are only used in this figure to avoid potential confounding in the interpretation of voxel-wise results. For each parameter left two columns show within-scan repeatability ($P_{92}^{(1)}$ vs. $P_{92}^{(2)}$ and $S_{127}^{(1)}$ vs. $S_{127}^{(2)}$), third column shows cross-protocol variability ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), and fourth column shows cross-scanner variability ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$).

Box plots of mean voxel-wise CVs in PLIC are shown in Figure 4 over all subjects (corresponding boxplots for kurtosis CVs are provided in Supporting Information Material). We observe strongest effects of denoising in W between data with differing echo times ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), where denoising complex data leads to a 12% decrease in CV in AW, along with marked improvements in CV in AD and FA from $\sim9\%$ down to 6%. Denoising (both magnitude and complex) give strong improvements in voxel-wise CV in kurtosis parameters because denoising helps minimize outliers in both K and W maps. While complex and magnitude denoising often give very similar changes in CV, MP complex gives the greatest improvement in cross-echo time reproducibility (3%–7% in MD, MW, and FA) because of the reduction in noise floor.

Figure 5a,b show the results of concordance analysis on voxel-wise data. It is visible from concordance correlation coefficients (shown in Figure 5b) that denoising leads to stronger correlations and lower variance in all test–retest datasets. In fact,

without denoising, voxel-wise correlations drop to as low as 0.13 across scanners and protocols in kurtosis values. Correlation and Bland–Altman analyses demonstrate the strongest correlations and narrowest voxel-wise parameter distributions between scans on the same scanner with the same echo-time after denoising. The correlations are lower for data acquired on different scanners (same echo time), or with different echo time (same scanner). Bland–Altman plots (Figure 5b) show the shift in parameter values resulting from decreased noise floor with complex denoising, particularly in MW. The middle row of Figure 5b shows how denoising complex data decreases WM kurtosis values while also minimizing test–retest error across all comparisons.

3.2 | ROI-Wise Variability

Figure 6a shows boxplots of ROI-means DTI and DKI parameters, respectively, in WM over all 20 subjects, for all five scans, and for
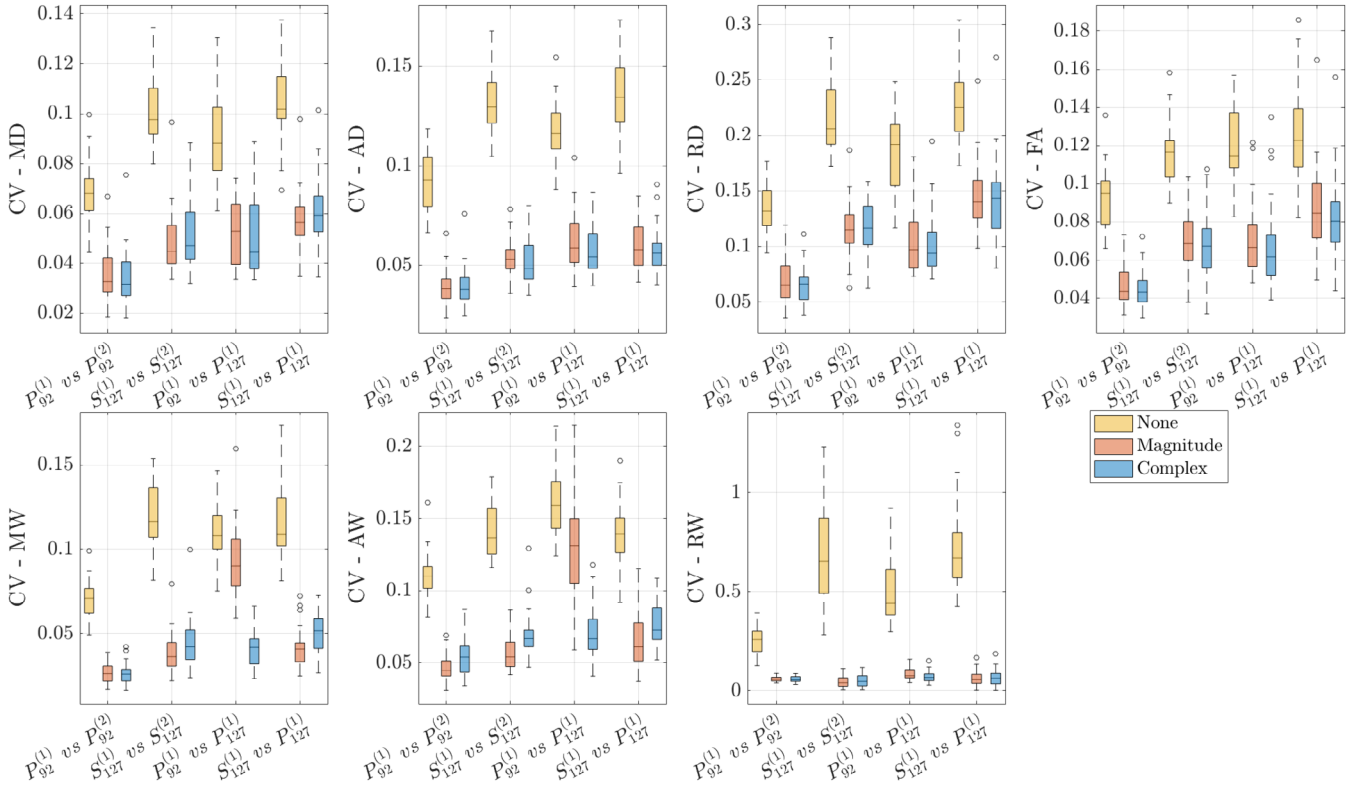
FIGURE 4 | Box plots of voxelwise CV, σ_x / μ_x, for each DTI and DKI parameter from pooling all voxels in whiter matter (PLIC and ALIC) over all 20 subjects. X-axes show each of the four comparisons: Within-scan variability ($P_{92}^{(1)}$ vs. $P_{92}^{(2)}$ and $S_{127}^{(1)}$ vs. $S_{127}^{(2)}$), cross-protocol variability ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), and cross-scanner variability ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$). Note that CVs here are not subject to non-linear warps, as they are pooled from subject-wise templates rather than the population template.

all denoising methods, and Figure 6b shows corresponding ROI-wise CVs (rather than voxel-wise CVs shown in Figure 4). DTI parameters were obtained from DTI-fitting including $b = 0$ and $b = 1000\,\text{s/mm}^2$ (DTI parameters obtained from DKI-fitting are shown in Figure S4). We observe variation in all parameter values according to TE. which can be related to corresponding SNR, which was derived as 13.1, 12.6, 9.3, 9.0, and 9.4 at $b = 0$-image for $P_{92}^{(1)}$, $P_{92}^{(2)}$, $P_{127}^{(1)}$, $S_{127}^{(1)}$, and $S_{127}^{(2)}$, respectively. While magnitude denoising increased variability most pronouncedly in AD, both magnitude and complex denoising reduced the variability in FA, and complex denoising clearly reduced AW and MW, likely due to reduced effect of the noise floor.

There is less advantage to denoising when averaging over hundreds of voxels within a region as compared to voxel-wise analyses (as SNR increases with the square root of the number of voxels in the ROI), but CVs still improve by 1%–9% (Table S1) for kurtosis parameters in the ACR when comparing data with differing echo times. We attribute this specific improved agreement in cross-echo time data to the reduction in noise floor provided by complex denoising, which is more effective at mitigating noise-related biases from longer echo times.

Figure 6b shows boxplots of ROI-wise CV DTI parameters and DKI parameters, respectively in the PLIC over all 20 subjects, for the four comparisons, and for all three processing methods. Test–retest variability on Prisma and Skyra hover between 1% and 4% for diffusion and kurtosis values, and the lowest test–retest variability is observed on the Prisma due to its high

SNR because of the shorter TE on this scanner. Cross-protocol and cross-scanner variability are higher due to inconsistency in echo time and hardware, respectively, resulting in greater differences in SNR between compared datasets. A notable decrease in CV is observed in AW and MW when examining cross-protocol variability because of bias reduction through complex denoising—resulting in decreased eigenvalue repulsion and noise floor.

Table S1 shows mean values of ROI-averaged CV over all 20 subjects. We found average test–retest CV for MD on the Prisma scanner to be about 1%–2% in WM, on the Skyra we found slightly larger CVs, but also on the order of 1%–2% regardless of whether denoising was used. Cross-protocol CV was higher, in the regime of 3%–4%, and cross-scan CV in the range of 2%–3%. Complex denoising consistently lowers variability in diffusivity, however this benefit was quite small (0.5%–1.0% improvements) at the level of the ROI. The strongest reduction in variability is observed for AW and complex denoising thanks to noise floor reduction prior to ROI averaging.

3.3 | Comparison Between Harmonization and Denoising

RISH harmonization was applied to voxel-wise data. Figure 7 shows ICCs across protocol or across scanner for data processed with magnitude or complex denoising and/or harmonization. ICC based on harmonization alone is lowest for FA and MW, both
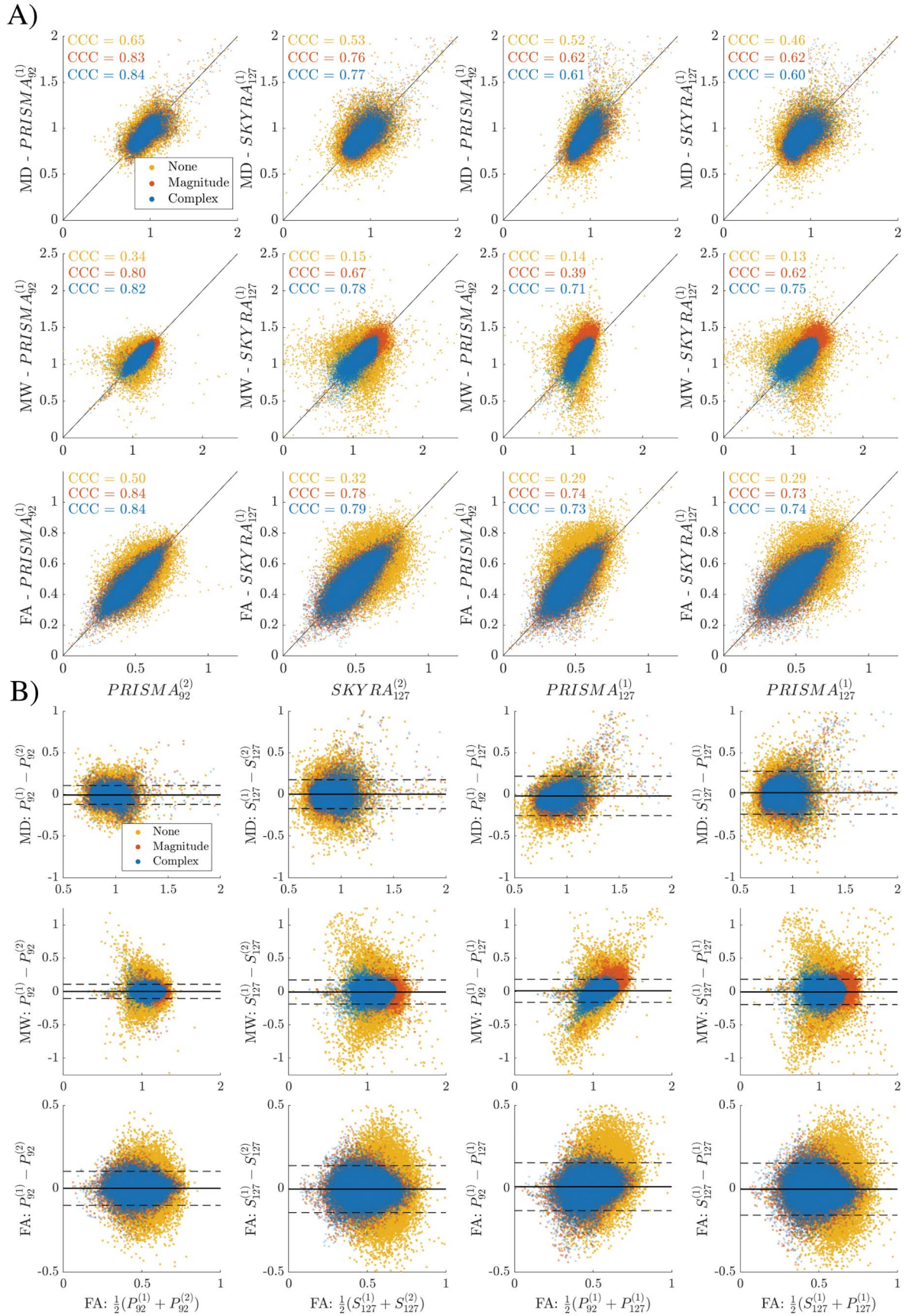
FIGURE 5 | (a) Voxel-wise scatter plots over white matter for all 20 subjects. Concordance correlation coefficients (CCC) are shown on the upper left corner of each plot for DTI and DKI parameters over white matter (internal capsule). (b) Bland–Altman plots over WM for all 20 subjects. Y axes show the error between test–retest measurements and X axes show the mean. Solid black line shows an error of 0, and dashed lines indicate error greater than 3 standard deviations away from the mean in complex denoised data.
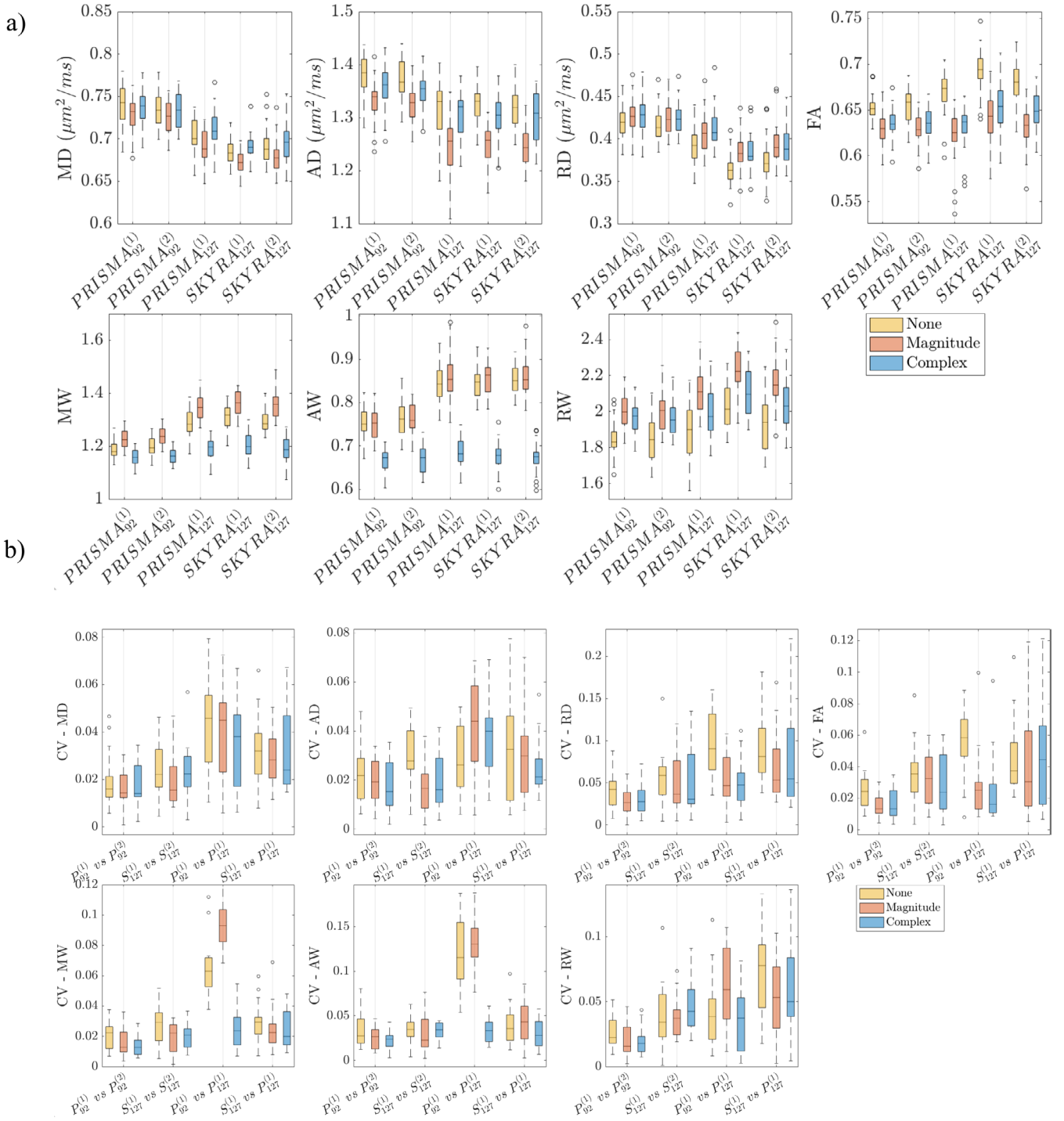
FIGURE 6 | (a) Boxplots of average ROI-values in white matter for each DTI and DKI parameters over all 20 subjects. (b) Boxplots of ROI-wise CV for each DTI and DKI parameters over all 20 subjects, where σ_x / μ_x is from within the white matter ROI. The white matter ROI included PLIC, ALIC. DTI parameters are shown in the top row and DKI parameters in the bottom row. Box plots show each of the four comparisons: Cross-protocol variability ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), within-scan variability ($P_{92}^{(1)}$ vs. $P_{92}^{(2)}$ and $S_{127}^{(1)}$ vs. $S_{127}^{(2)}$), and cross-scanner variability ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$).

metrics that are heavily affected by noise propagation. MD, which is more robust to noise, has high ICC both with harmonization alone and with denoising alone. In both cross-protocol and cross-scanner comparisons MW had the greatest benefit from combining denoising with linear-RISH, where cross-protocol, magnitude denoising improved the ICC from 0.09 to 0.32, and magnitude denoising+RISH further improved the ICC to 0.85. The same trend was observed in the cross-scanner comparison over MW, but to a

lesser extent. In both comparisons, FA showed a decrease in ICC with denoising+RISH compared to denoising alone.

3.4 | Noise Floor Estimation

Figure 8 shows the noise floors for dMRI data denoised using MP-complex, MP-magnitude, and no denoising for each of
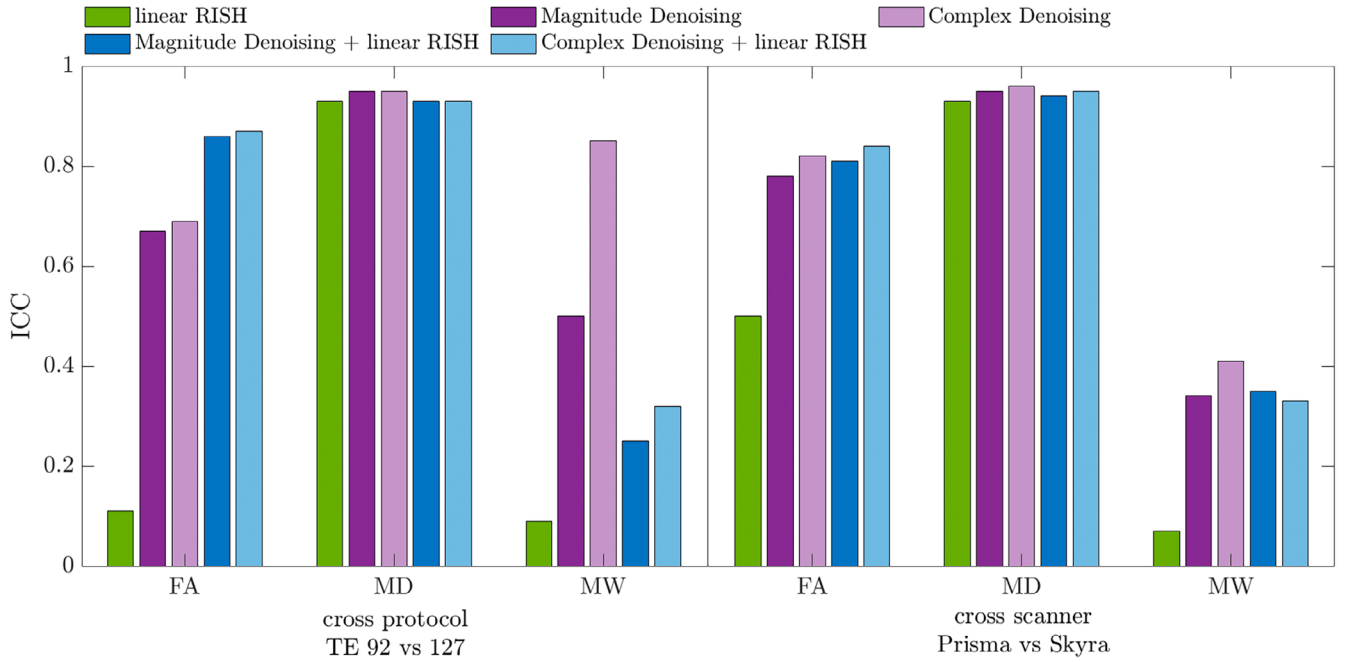
FIGURE 7 | Intraclass correlation coefficients for two comparisons: Cross-protocol variability, and cross-scanner variability. Denoising of both complex and magnitude data consistently improves the ICC.
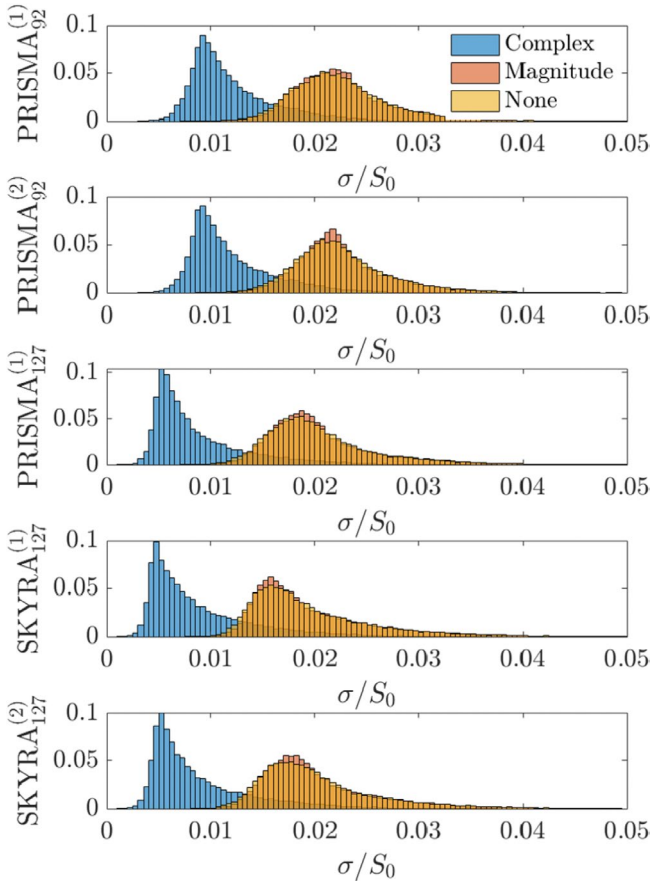

FIGURE 8 | Noise floor estimation for each of the five scans according to Equation (9). Distributions shown come from pooled voxels in the ventricles over all subjects. As expected, MP magnitude shows the same noise floor as no-denoising (as it does not remove the Rician bias). MP complex decreases the noise floor by a factor of 2–4.

the five scans. We found that the baseline noise level for non-denoised (or magnitude denoised) data is about 1.8%–2.5% of the dynamic range of the DWI dataset. Complex denoising lowers the noise floor to 0.4%–1%, on the Prisma system constituted a 2.5× reduction in noise floor. This reduction in noise floor propagates through tensor estimation and leads to the decreases in parameter bias present in Figure 6a. We note that the noise floor is higher in the low echo-time Prisma data compared to the data acquired with TE = 127 ms, which may be attributable to contributions from physiological noise or increased B0 inhomogeneities present in lower TE data, even though this data has higher overall SNR.

3.5 | Statistical Power Estimation

Figure 9 shows the results of a statistical estimate of the sample size required to detect 5% difference in group means, for groups with equal variance. This analysis was performed for a voxel located in a common WM region (PLIC). Here we found that denoising universally lowers the number of subjects required to reach statistical significance ($p = 0.05$ in a two-sided t-test). For voxel-wise data (Figure 9a), on the Prisma scanner, denoising allowed for sample size decreases of 51.2%, 64.9%, and 50% for MD, MW, and FA, respectively. On the Skyra we found that denoising allowed for corresponding decreases of 53.4%, 73.4%, and 35.9%. Both magnitude and complex denoising reduced the required sample size from 40–80 subjects (MD and MW) down to 20–40 subjects for MD and 10–40 subjects for MW, and from 70–100 subjects down to 40–70 subjects for FA. Both complex and magnitude denoising produced similar effects, with the notable exception that when comparing groups who underwent scans with differing TE, where for MW, complex denoising provided an addition 66.6% reduction
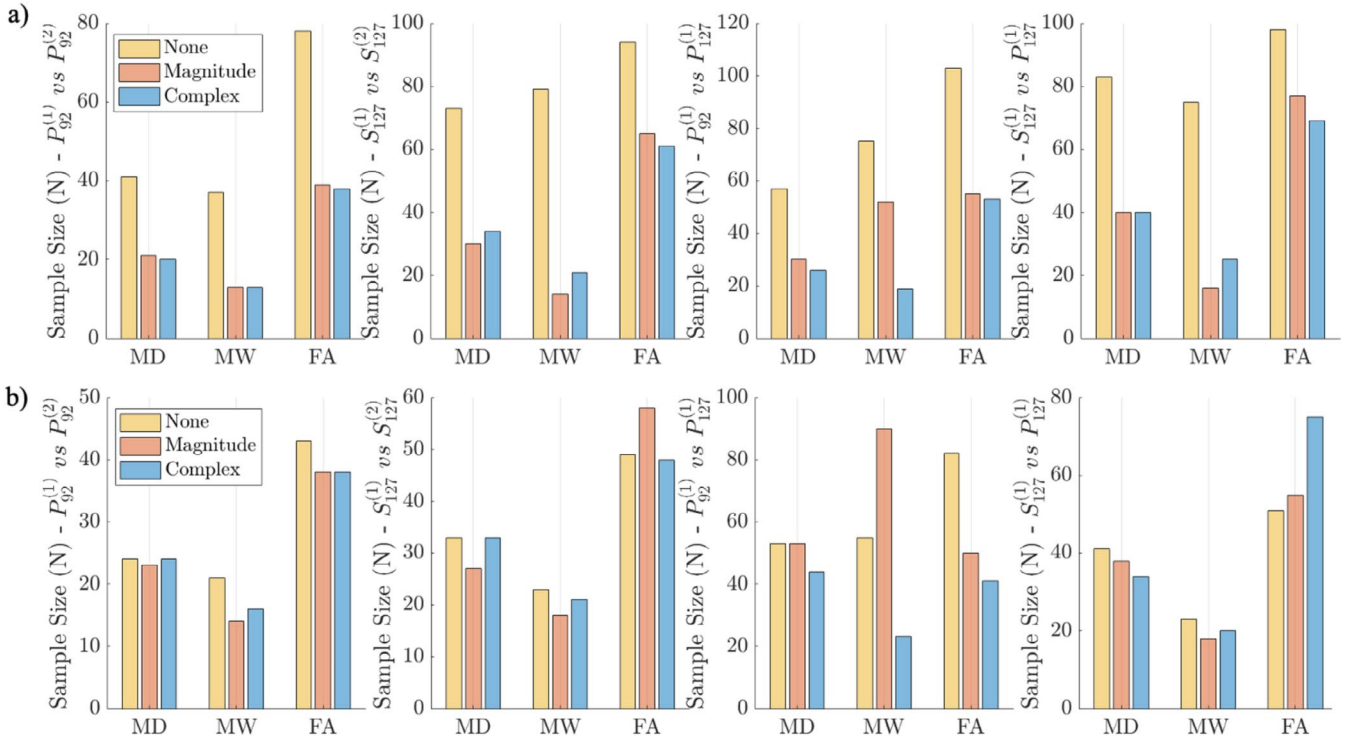
FIGURE 9 | (a) Required sample size needed to significantly distinguish groups separated by an effect size of 5% with $\alpha = 0.05$ and a power of 0.8 when performing voxel-wise analysis, based on parameter ROI means and voxel-wise test–retest standard deviations over all 20 subjects. (b) Required sample size needed when performing ROI-based analysis, based on parameter ROI means and ROI-wise test–retest standard deviations. Rows show sample sizes needed for four scenarios, with either cross-scanner variability ($P_{92}^{(1)}$ vs. $P_{92}^{(2)}$ and $S_{127}^{(1)}$ vs. $S_{127}^{(2)}$), cross-protocol variability ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), and cross-scanner variability ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$).

in sample size compared to magnitude denoising. For ROI-wise data (Figure 9b), the high SNR enabled by ROI averaging led to minimal sample size changes between denoised and non-denoised data.

For data acquired using protocols with different TE ($P_{92}^{(1)}$ vs. $P_{127}^{(1)}$), complex denoising had the largest impact on sample size improvement because it offers a correction for differing noise floors. Complex denoising allowed for a sample size decrease of 53.3%, 73.9%, and 60% in MD, MW, and FA. On different scanners with the same TE ($S_{127}^{(1)}$ vs. $P_{127}^{(1)}$), both magnitude and complex denoising allows for a decrease in sample size of similar magnitude, hence we can conclude that these datasets contain similar noise floors.

4 | Discussion

The goal of this study was to measure the reproducibility of higher-order diffusion MR metrics. Denoising (Veraart, Fieremans, and Novikov 2016) of magnitude or complex diffusion MRI data, together with targeted artifact removal (Ades-Aron et al. 2018; Chen et al. 2024), improves reproducibility of higher-order diffusion parameters across scanners and protocols. Denoising reduces variations across test–retest datasets from ∼15%–20% to ∼5%–10% in kurtosis indices at the level of individual voxels. Subsequent measurements on different scanners and different TE (92 ms, 127 ms) were found to have voxel-wise precision varying from 3% to 15% for both DTI and DKI

metrics after denoising. Notably, we found the greatest improvement in reproducibility across data with differing signal-to-noise ratios (from varying echo times) when applying MPPCA to complex-valued data, likely because of the ∼2.5× noise floor reduction offered by complex denoising.

When comparing denoising directly to harmonization using linear-RISH, denoising outperformed harmonization in reducing bias from varying noise floors. Moreover, combining denoising with harmonization in voxel-wise assessments improved intra-scanner test–retest ICCs by 55% for FA, over harmonization alone. This highlights denoising's critical role in maximizing data precision and reliability in multi-site dMRI studies.

This works stands out compared to previous works due to the comprehensive nature of the multi-shell diffusion MRI dataset of 20 subjects, which allows for evaluating three separate types of reproducibility, along with analysis of how noise varies in repeated voxel-wise measurements. Previous studies aimed to evaluate harmonization schemes include denoising as an aspect of their harmonization (Ning et al. 2020) either directly, or because of neural network harmonization where denoising is implicit to the network due to clean supervised training data. The values reported here are largely in line with prior results of DTI repeatability found in the literature (Kristo et al. 2013; Laguna et al. 2020; Liu et al. 2014; Vollmar et al. 2010) where for ROIs in the corpus callosum, FA has been shown to have ∼2%–3% within-scan and between-scan variability in aligned WM regions. Our reported values also agree with a previous report

of intra-scanner ROI mean kurtosis repeatability of 1%–4% (Kamiya et al. 2020). Similarly, we also show correspondence with reported reproducibility (Coelho et al. 2022) for tissue microstructure parameters. Such reproducibility may improve our ability to measure subtle changes in nervous tissue functional organization and complexity.

For this work, we employed an established though perhaps less commonly used definition for kurtosis (Hansen et al. 2013; Jespersen et al. 2018; Lu et al. 2006), (Equation 7), as we observed that kurtosis maps using the traditional definition, (Equation 6) feature more outliers due to unphysical kurtosis values or "black voxels" (Figure 2) as compared to the alternative definition (Equation 7), as also quantified in Figure S3. Alternatively, image correction methods, for example, based on the MK-curve (Zhang et al. 2019), can also reduce the occurrence of "black voxels" and thereby improve the stability of kurtosis estimates, and subsequently increase the robustness and reproducibility of DKI measurements (Christiaanse et al. 2023). Such correction methods are complementary to our proposed denoising methods and could be combined for optimal robustness and reproducibility.

We found that MPPCA denoising has a strong impact on voxel-wise data, lowering CVs in kurtosis from ~40% down to 5%. As shown in Figure 4, the effect is most noticeable in homogeneous brain regions. However, voxel-wise CV is nonuniform across WM due to different levels of tissue homogeneity in different regions, and CV is typically higher in regions with tissue boundaries such as the genu and splenium of the corpus callosum, owing largely to the effect of registration-based spatial interpolations along with small misregistration effects. Similarly, voxel-wise CV in GM (thalami) is higher than in homogenous WM owing to heterogeneous tissue composition.

There is less of a need for denoising when analyzing large groups of voxels (Figure 6b and Table S1). Indeed, ROI analyses are more reproducible because of averaging many voxels compared to voxel-wise analyses (Kornaropoulos et al. 2022), but may be also less specific and obscured due to spatially local effects. Conversely, voxel-wise analyses better capture biological variability, in which case denoising profoundly improves reproducibility. Indeed, voxel-wise, cross-protocol, and cross-scanner CVs for non-denoised diffusion parameters reach up to 20%, and after denoising variability drops to 3%–8% in most areas (Table 2). DTI differences between groups for varying psychiatric disorders can be as low as 1% and up to 50% effect size (Hansen et al. 2013), therefore our observed reduction in CV (Figure 4) gained through denoising may be clinically useful.

In line with the observed reduction in voxel-wise CV after denoising, we also found that denoising decreases the sample size needed to reach statistical significance between groups separated by a 5% effect size by about 50% in MD, MW, and FA when performing a voxel-wise analysis (Figure 9a). The degree of sample size reduction is determined by the underlying SNR of the dataset. In the voxel-wise data, the intrinsic low SNR ($SNR = 13$ at $TE = 92$ ms and $SNR = 9$ at TE 127) of our diffusion protocol (due to small voxel size, short scan time, and minimal reconstruction), led to the most pronounced sample size reductions after denoising.

Conversely, and in line with results in Figure 6b, after increasing the SNR through ROI averaging, denoising had minimal (5%) to no impact on the required sample size for denoising (Figure 9b). At the level of ROIs 24 patients are required to detect a 5% difference in mean MD using a 2 sided t-test regardless of denoising strategy, which is in-line with predicted sample sizes from literature (Laguna et al. 2020; Szczepankiewicz et al. 2013).

Figure 6a shows the effect of varying TE, and corresponding SNR, on DTI and DKI parameters. While there is no ground truth for those in vivo values, we know that repulsion due to noise causes artificial increases to FA, AD, and decreases in RD (Basser and Pajevic 2000; Sprenger et al. 2016), an effect that is SNR dependent, and is reduced after denoising. Furthermore, the eigenvalue shrinkage step during denoising has been shown (Chen et al. 2024) to bias DTI-parameters, another SNR-dependent effect that is induced by denoising, with decreased AD values (Chen et al. 2024) by −4% at $SNR = 10$. Finally, the effect of bias due to noise floor is also SNR-dependent, and can be reduced by complex denoising. All these effects may impact the outcome DTI/DKI parameters, as illustrated in Figure 6b, where shrinkage may explain the increased variability in AD after magnitude denoising, and reduced eigenvalue repulsion may explain the reduced variability in FA after both magnitude and complex denoising. In addition, complex denoising reduced the effect of the noise floor, resulting in lower and more uniform AW and MW across TE.

The additional benefit of MPPCA denoising of complex dMRI data (magnitude and phase) compared to denoising magnitude dMRI data is most pronounced in reducing parameter bias of higher-order dMRI metrics such as AW and to a lesser effect MW, as shown in Figure 4 and Figure 5a, particularly lowering the variability between protocols with different TE. Interestingly, the deviation of the regression slope from 1 in cross echo time (92 vs. 127 ms) CV of MW data (Figure 5b) is reduced most after complex denoising due to lowering the noise floor caused by higher noise floor in the long-TE scan.

Complex MP-denoising has the potential for further improvement since the algorithm uses a "two-pass" process, where we run the MP algorithm twice, first to estimate the noise free signal phase and then to perform the actual denoising. For this method to work effectively, these two steps must be sufficiently independent, that is, the first step should not introduce noise correlations into the second step. If the patches used in each step are too similar, the denoising will not perform as robustly in the second denoising step. In this study we tried to make the patches as different as possible (15×15 2D patches during phase estimation, 3D adaptive patches during the actual denoising), since the noise correlations introduced by the phase unwinding in the first step should be made as small as possible to maximize the performance of this approach. This is the reason why in some cases ($P_{127}^{(1)}$ vs. $S_{127}^{(1)}$ AW in Figure 5a) we observe slightly better denoising performance in MP-magnitude compared with MP-complex.

Here we show that the MPPCA denoising approach provides not only a notable and quantifiable improvement in reproducibility of cross-protocol dMRI parameters, but also provides a

powerful tool for data harmonization (by means of reducing the noise floor). Indeed, due to the different echo times and numbers of gradient directions used in $P_{92}^{(1)}$ and $P_{127}^{(1)}$ protocols, the resulting dMRI data have differing SNR and noise floor (see Figure 8). Without denoising, this results in parametric maps with differing levels of bias, and different numbers of outliers in kurtosis parameters owing to the different levels of precision in the raw data. Hence, reducing the noise floor by denoising complex dMRI data has the potential to improve parameter accuracy and *harmonize* data from sources with different SNR and thus, different noise floor. This is exemplified by the results of the harmonization analysis (Figure 7). We found that the impact of both complex denoising and harmonization was greatest when adjusting for the noise floor induced bias in FA in data with differing echo times. Since complex denoising can accurately reduce noise floor bias, it should be included as an essential first step to harmonizing data from separate sources.

Limitations of this study include the lack of a ground truth inhibiting us from knowing the exact degree of bias and variance induced by protocol specific effects. Recent work (Chen et al. 2024) showed that eigenvalue shrinkage can bias DTI and DKI parameters, particularly at low SNR, illustrated in Figure 6a, which warrants further investigation into its usefulness when combining dMRI with different noise levels. In addition, there are also additional sources of test–retest variability that were not measured here, including differing vendors, field strengths, *q*-space sampling regimes, and number of head coils, but could be the subject of future work. The custom reconstruction pipeline used in this study allowed for saving of both magnitude and phase images, which facilitated both magnitude and complex denoising. When using vendor and sequence-specific online reconstruction pipelines, phase map preservation, needed for complex denoising, may not be as readily accessible, and potential unknown reconstruction steps (e.g., filtering, interpolation) may interfere with the MPPCA denoising algorithm. Nevertheless, the core principles of our analysis, particularly the use of complex data for denoising, remain applicable and can be adapted to various reconstruction pipelines. We assessed only young adult healthy controls, and it should be noted that age or pathological changes to tissue microstructure may also increase diffusion parameter variability.

MP complex code is available as part of the DESIGNER-v2 diffusion MRI processing pipeline (https://github.com/NYU-DiffusionMRI/DESIGNER-v2) and can be run either through python or through a dedicated Docker container.

5 | Conclusion

MPPCA denoising reduces variability across scanners and echo times to 5%–10% voxel-wise and 3%–5% at the ROI-level for DTI and DKI metrics and has the potential to minimize the sample size required for voxel-wise population-wise statistics by 40%–70% depending on SNR. Denoising of complex dMRI enables noise floor reduction by up to 60%. Such improvement—that measurements from *individual voxels* become quantitative and reproducible—is an essential step towards bringing quantitative MRI and tissue microstructure imaging with MRI to clinical practice.

Conflicts of Interest

The authors, including B.A.A., G.L., T.M.S., D.S.N., E.F., and N.Y.U. School of Medicine, hold stock in Microstructure Imaging Inc., which develops post-processing tools for advanced MRI methods. Additionally, D.S.N., E.F., and J.V. are co-inventors of a patent co-owned by the University of Antwerp and NYU School of Medicine, describing technology evaluated in this study.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Ades-Aron, B., J. Veraart, P. Kochunov, et al. 2018. "Evaluation of the Accuracy and Precision of the Diffusion Parameter EStImation With Gibbs and NoisE Removal Pipeline." *NeuroImage* 183: 532–543.

Ahn, C., and Z. Cho. 1987. "A New Phase Correction Method in NMR Imaging Based on Autocorrelation and Histogram Analysis." *IEEE Transactions on Medical Imaging* 6, no. 1: 32–36.

Aja-Fernández, S., A. Tristán-Vega, and C. Alberola-López. 2009. "Noise Estimation in Single- and Multiple-Coil Magnetic Resonance Data Based on Statistical Models." *Magnetic Resonance Imaging* 27, no. 10: 1397–1409.

Alexander, D. C., T. B. Dyrby, M. Nilsson, and H. Zhang. 2019. "Imaging Brain Microstructure With Diffusion MRI: Practicality and Applications." *NMR in Biomedicine* 32, no. 4: e3841.

Amartur, S., and E. M. Haacke. 1991. "Modified Iterative Model Based on Data Extrapolation Method to Reduce Gibbs Ringing." *Journal of Magnetic Resonance Imaging* 1, no. 3: 307–317.

Andersson, J. L., and S. N. Sotiropoulos. 2016. "An Integrated Approach to Correction for Off-Resonance Effects and Subject Movement in Diffusion MR Imaging." *NeuroImage* 125: 1063–1078.

Andersson, J. L., M. S. Graham, E. Zsoldos, and S. N. Sotiropoulos. 2016. "Incorporating Outlier Detection and Replacement Into a Non-Parametric Framework for Movement and Distortion Correction of Diffusion MR Images." *NeuroImage* 141: 556–572.

Bao, L., M. Robini, W. Liu, and Y. Zhu. 2013. "Structure-Adaptive Sparse Denoising for Diffusion-Tensor MRI." *Medical Image Analysis* 17, no. 4: 442–457.

Basser, P. J., and S. Pajevic. 2000. "Statistical Artifacts in Diffusion Tensor MRI (DT-MRI) Caused by Background Noise." *Magnetic Resonance in Medicine* 44, no. 1: 41–50.

Beck, D., A.-M. G. de Lange, I. I. Maximov, et al. 2021. "White Matter Microstructure Across the Adult Lifespan: A Mixed Longitudinal and Cross-Sectional Study Using Advanced Diffusion Models and Brain-Age Prediction." *NeuroImage* 224: 117441.

Bernstein, M. A., K. F. King, and X. J. Zhou. 2004. *Handbook of MRI Pulse Sequences*. Amsterdam; Boston: Elsevier Academic Press.

Breuer, F. A., M. Blaimer, R. M. Heidemann, M. F. Mueller, M. A. Griswold, and P. M. Jakob. 2005. "Controlled Aliasing in Parallel Imaging Results in Higher Acceleration (CAIPIRINHA) for Multi-Slice Imaging." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 53, no. 3: 684–691.

Brey, W. W., and P. A. Narayana. 1988. "Correction for Intensity Falloff in Surface Coil Magnetic Resonance Imaging." *Medical Physics* 15, no. 2: 241–245.

Buades, A., B. Coll, and J.-M. Morel. 2005. "A Review of Image Denoising Algorithms, With a New One." *Multiscale Modeling and Simulation* 4, no. 2: 490–530.

Chen, J., B. Ades-Aron, H.-H. Lee, et al. 2024. "Optimization and Validation of the DESIGNER Preprocessing Pipeline for Clinical Diffusion MRI in White Matter Aging." *Imaging Neuroscience* 2: 1–17.

Christiaanse, E., P. O. Wyss, A. Scheel-Sailer, et al. 2023. "Mean Kurtosis-Curve (MK-Curve) Correction Improves the Test–Retest Reproducibility of Diffusion Kurtosis Imaging at 3 T." *NMR in Biomedicine* 36, no. 3: e4856.

Coelho, S., S. H. Baete, G. Lemberskiy, et al. 2022. "Reproducibility of the Standard Model of Diffusion in White Matter on Clinical MRI Systems." *NeuroImage* 257: 119290.

Cordero-Grande, L., D. Christiaens, J. Hutter, A. N. Price, and J. V. Hajnal. 2019. "Complex Diffusion-Weighted Image Estimation via Matrix Recovery Under General Noise Models." *NeuroImage* 200: 391–404.

Cox, S. R., S. J. Ritchie, E. M. Tucker-Drob, et al. 2016. "Ageing and Brain White Matter Structure in 3,513 UK Biobank Participants." *Nature Communications* 7, no. 1: 13629.

de Kouchkovsky, I., E. Fieremans, L. Fleysher, J. Herbert, R. I. Grossman, and M. Inglese. 2016. "Quantification of Normal-Appearing White Matter Tract Integrity in Multiple Sclerosis: A Diffusion Kurtosis Imaging Study." *Journal of Neurology* 263, no. 6: 1146–1155.

Desikan, R. S., F. Ségonne, B. Fischl, et al. 2006. "An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans Into Gyral Based Regions of Interest." *NeuroImage* 31, no. 3: 968–980.

Does, M. D., J. L. Olesen, K. D. Harkins, et al. 2019. "Evaluation of Principal Component Analysis Image Denoising on Multi-Exponential MRI Relaxometry." *Magnetic Resonance in Medicine* 81, no. 6: 3503–3514.

Dong, J. W., I. O. Jelescu, B. Ades-Aron, et al. 2020. "Diffusion MRI Biomarkers of White Matter Microstructure Vary Nonmonotonically With Increasing Cerebral Amyloid Deposition." *Neurobiology of Aging* 89: 118–128.

Fadnavis, S., J. Batson, and E. Garyfallidis. 2020. "Patch2Self: Denoising Diffusion MRI With Self-Supervised Learning." *Advances in Neural Information Processing Systems* 33: 16293–16303.

Falangola, M. F., J. H. Jensen, A. Tabesh, et al. 2013. "Non-Gaussian Diffusion MRI Assessment of Brain Microstructure in Mild Cognitive Impairment and Alzheimer's Disease." *Magnetic Resonance Imaging* 31, no. 6: 840–846.

Gavish, M., and D. L. Donoho. 2017. "Optimal Shrinkage of Singular Values." *IEEE Transactions on Information Theory* 63, no. 4: 2137–2152.

Griswold, M. A., P. M. Jakob, R. M. Heidemann, et al. 2002. "Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA)." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 47, no. 6: 1202–1210.

Grussu, F., M. Battiston, J. Veraart, et al. 2020. "Multi-Parametric Quantitative In Vivo Spinal Cord MRI With Unified Signal Readout and Image Denoising." *NeuroImage* 116884: 116884.

Gudbjartsson, H., and S. Patz. 1995. "The Rician Distribution of Noisy Mri Data." *Magnetic Resonance in Medicine* 34, no. 6: 910–914.

Hansen, B., T. E. Lund, R. Sangill, and S. N. Jespersen. 2013. "Experimentally and Computationally Fast Method for Estimation of a Mean Kurtosis." *Magnetic Resonance in Medicine* 69, no. 6: 1754–1760.

Henriques, R. N., A. Ianuş, L. Novello, J. Jovicich, S. N. Jespersen, and N. Shemesh. 2023. "Efficient PCA Denoising of Spatially Correlated Redundant MRI Data." *Imaging Neuroscience* 1: 1–26.

Hua, K., J. Zhang, S. Wakana, et al. 2008. "Tract Probability Maps in Stereotaxic Spaces: Analyses of White Matter Anatomy and Tract-Specific Quantification." *NeuroImage* 39, no. 1: 336–347.

Jack, C. R., Jr., M. A. Bernstein, N. C. Fox, et al. 2008. "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for." *Magnetic Resonance in Medicine* 27, no. 4: 685–691.

Jelescu, I. O., and M. D. Budde. 2017. "Design and Validation of Diffusion MRI Models of White Matter." *Frontiers of Physics* 5: 61.

Jensen, J. H., J. A. Helpern, A. Ramani, H. Lu, and K. Kaczynski. 2005. "Diffusional Kurtosis Imaging: The Quantification of Non-Gaussian Water Diffusion by Means of Magnetic Resonance Imaging." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 53, no. 6: 1432–1440.

Jespersen, S. N., J. L. Olesen, B. Hansen, and N. Shemesh. 2018. "Diffusion Time Dependence of Microstructural Parameters in Fixed Spinal Cord." *NeuroImage* 182: 329–342.

Johnstone, I. M. 2007. "High Dimensional Statistical Inference and Random Matrices." In *Proceedings of the International Congress of Mathematicians Madrid*, 307–333.

Jones, D. K., and M. R. I. Diffusion. 2010. *Theory, Methods, and Application*, xvi, 767. Oxford; New York: Oxford University Press.

Jones, D. K., and P. J. Basser. 2004. "'Squashing Peanuts and Smashing Pumpkins': How Noise Distorts Diffusion-Weighted MR Data." *Magnetic Resonance in Medicine* 52, no. 5: 979–993.

Kamiya, K., K. Kamagata, K. Ogaki, et al. 2020. "Brain White-Matter Degeneration due to Aging and Parkinson Disease as Revealed by Double Diffusion Encoding." *Frontiers in Neuroscience—Switzerland* 14: 584510.

Kellner, E., B. Dhital, V. G. Kiselev, and M. Reisert. 2016. "Gibbs-Ringing Artifact Removal Based on Local Subvoxel-Shifts." *Magnetic Resonance in Medicine* 76, no. 5: 1574–1581.

Kim, H.-Y. 2016. "Statistical Notes for Clinical Researchers: Sample Size Calculation 1. Comparison of Two Independent Sample Means." *Restorative Dentistry & Endodontics* 41, no. 1: 74–78.

Kochunov, P., L. E. Hong, E. L. Dennis, et al. 2022. "ENIGMA-DTI: Translating Reproducible White Matter Deficits Into Personalized Vulnerability Metrics in Cross-Diagnostic Psychiatric Research." *Human Brain Mapping* 43, no. 1: 194–206.

Kornaropoulos, E. N., S. Winzeck, T. Rumetshofer, et al. 2022. "Sensitivity of Diffusion MRI to White Matter Pathology: Influence of Diffusion Protocol, Magnetic Field Strength, and Processing Pipeline in Systemic Lupus Erythematosus." *Frontiers in Neurology* 13: 837385.

Kristo, G., A. Leemans, B. de Gelder, M. Raemaekers, G.-J. Rutten, and N. Ramsey. 2013. "Reliability of the Corticospinal Tract and Arcuate Fasciculus Reconstructed With DTI-Based Tractography: Implications for Clinical Practice." *European Radiology* 23: 28–36.

Laguna, P. A. L., A. J. Combes, J. Streffer, et al. 2020. "Reproducibility, Reliability and Variability of FA and MD in the Older Healthy Population: A Test-Retest Multiparametric Analysis." *NeuroImage: Clinical* 26: 102168.

Larkman, D. J., J. V. Hajnal, A. H. Herlihy, G. A. Coutts, I. R. Young, and G. Ehnholm. 2001. "Use of Multicoil Arrays for Separation of Signal From Multiple Slices Simultaneously Excited." *Journal of Magnetic*

Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 13, no. 2: 313–317.

Lawrence, I., and K. Lin. 1989. "A Concordance Correlation Coefficient To Evaluate Reproducibility." *Biometrics* 45: 255–268.

Lee, H. H., D. S. Novikov, and E. Fieremans. 2021. "Removal of Partial Fourier-Induced Gibbs (RPG) Ringing Artifacts in MRI." *Magnetic Resonance in Medicine* 86: 2733–2750.

Lemberskiy, G., S. Baete, J. Veraart, T. M. Shepherd, E. Fieremans, and D. S. Novikov. 2019. "Achieving Sub-mm Clinical Diffusion MRI Resolution By Removing Noise During Reconstruction Using Random Matrix Theory."

Liu, X., Y. Yang, J. Sun, et al. 2014. "Reproducibility of Diffusion Tensor Imaging in Normal Subjects: An Evaluation of Different Gradient Sampling Schemes and Registration Algorithm." *Neuroradiology* 56: 497–510.

Lu, H., J. H. Jensen, A. Ramani, and J. A. Helpern. 2006. "Three-Dimensional Characterization of Non-Gaussian Water Diffusion in Humans Using Diffusion Kurtosis Imaging." *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance in Vivo* 19, no. 2: 236–247.

Maggioni, M., V. Katkovnik, K. Egiazarian, and A. Foi. 2012. "Nonlocal Transform-Domain Filter for Volumetric Data Denoising and Reconstruction." *IEEE Transactions on Image Processing* 22, no. 1: 119–133.

Manjón, J. V., P. Coupé, and A. Buades. 2015. "MRI Noise Estimation and Denoising Using Non-Local PCA." *Medical Image Analysis* 22, no. 1: 35–47.

Manjón, J. V., P. Coupé, L. Concha, A. Buades, D. L. Collins, and M. Robles. 2013. "Diffusion Weighted Image Denoising Using Overcomplete Local PCA." *PLoS One* 8, no. 9: e73021.

Manjon, J. V., P. Coupe, L. Marti-Bonmati, D. L. Collins, and M. Robles. 2010. "Adaptive Non-Local Means Denoising of MR Images With Spatially Varying Noise Levels." *Journal of Magnetic Resonance Imaging* 31, no. 1: 192–203.

Manzano Patron, J. P., S. Moeller, J. L. Andersson, K. Ugurbil, E. Yacoub, and S. N. Sotiropoulos. 2024. "Denoising Diffusion MRI: Considerations and Implications for Analysis." *Imaging Neuroscience* 2: 1–29.

McGraw, K. O., and S. P. Wong. 1996. "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods* 1, no. 1: 30–46.

Mirzaalian, H., L. Ning, P. Savadjiev, et al. 2016. "Inter-Site and Inter-Scanner Diffusion MRI Data Harmonization." *NeuroImage* 135: 311–323.

Moeller, S., P. K. Pisharady, S. Ramanna, et al. 2021. "NOise Reduction With DIstribution Corrected (NORDIC) PCA in dMRI With Complex-Valued Parameter-Free Locally Low-Rank Processing." *NeuroImage* 226: 117539.

Ning, L., E. Bonet-Carne, F. Grussu, et al. 2020. "Cross-Scanner and Cross-Protocol Multi-Shell Diffusion MRI Data Harmonization: Algorithms and Results." *NeuroImage* 221: 117128.

Novikov, D. S., E. Fieremans, S. N. Jespersen, and V. G. Kiselev. 2019. "Quantifying Brain Microstructure With Diffusion MRI: Theory and Parameter Estimation." *NMR in Biomedicine* 32: e3998.

Novikov, D. S., J. Veraart, I. O. Jelescu, and E. Fieremans. 2018. "Rotationally-Invariant Mapping of Scalar and Orientational Metrics of Neuronal Microstructure With Diffusion MRI." *NeuroImage* 174: 518–538.

Novikov, D. S., V. G. Kiselev, and S. N. Jespersen. 2018. "On modeling." *Magnetic Resonance in Medicine* 79, no. 6: 3172–3193.

Olesen, J. L., A. Ianus, L. Østergaard, N. Shemesh, and S. N. Jespersen. 2023. "Tensor Denoising of Multidimensional MRI Data." *Magnetic Resonance in Medicine* 89, no. 3: 1160–1172.

Pogosbekian, E., I. N. Pronin, N. Zakharova, et al. 2021. "Feasibility of Generalised Diffusion Kurtosis Imaging Approach for Brain Glioma Grading." *Neuroradiology* 63: 1241–1251.

Pruessmann, K. P., M. Weiger, M. B. Scheidegger, and P. Boesiger. 1999. "SENSE: Sensitivity Encoding for Fast MRI." *Magnetic Resonance in Medicine* 42, no. 5: 952–962.

Setsompop, K., B. A. Gagoski, J. R. Polimeni, T. Witzel, V. J. Wedeen, and L. L. Wald. 2012. "Blipped-Controlled Aliasing in Parallel Imaging for Simultaneous Multislice Echo Planar Imaging With Reduced g-Factor Penalty." *Magnetic Resonance in Medicine* 67, no. 5: 1210–1224.

Smith, S. M., M. Jenkinson, M. W. Woolrich, et al. 2004. "Advances in Functional and Structural MR Image Analysis and Implementation as FSL." *NeuroImage* 23: S208–S219.

Sodickson, D. K., and W. J. Manning. 1997. "Simultaneous Acquisition of Spatial Harmonics (SMASH): Fast Imaging With Radiofrequency Coil Arrays." *Magnetic Resonance in Medicine* 38, no. 4: 591–603.

Sprenger, T., J. I. Sperl, B. Fernandez, et al. 2016. "Bias and Precision Analysis of Diffusional Kurtosis Imaging for Different Acquisition Schemes." *Magnetic Resonance in Medicine* 76, no. 6: 1684–1696.

St-Jean, S., P. Coupé, and M. Descoteaux. 2016. "Non Local Spatial and Angular Matching: Enabling Higher Spatial Resolution Diffusion MRI Datasets Through Adaptive Denoising." *Medical Image Analysis* 32: 115–130.

Stauffer, E.-M., R. A. Bethlehem, V. Warrier, et al. 2021. "Grey and White Matter Microstructure Is Associated With Polygenic Risk for Schizophrenia." *Molecular Psychiatry* 26, no. 12: 7709–7718.

Sudlow, C., J. Gallacher, N. Allen, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLoS Medicine* 12, no. 3: e1001779.

Szczepankiewicz, F., J. Lätt, R. Wirestam, et al. 2013. "Variability in Diffusion Kurtosis Imaging: Impact on Study Design, Statistical Power and Interpretation." *NeuroImage* 76: 145–154.

Tian, C., Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu. 2020. "Attention-Guided CNN for Image Denoising." *Neural Networks* 124: 117–129.

Tian, Q., Z. Li, Q. Fan, et al. 2022. "SDnDTI: Self-Supervised Deep Learning-Based Denoising for Diffusion Tensor MRI." *NeuroImage* 253: 119033.

Tournier, J.-D., R. Smith, D. Raffelt, et al. 2019. "MRtrix3: A Fast, Flexible and Open Software Framework for Medical Image Processing and Visualisation." *NeuroImage* 202: 116137.

Van Essen, D. C., S. M. Smith, D. M. Barch, et al. 2013. "The WU-Minn Human Connectome Project: An Overview." *NeuroImage* 80: 62–79.

Veraart, J., D. S. Novikov, D. Christiaens, B. Ades-Aron, J. Sijbers, and E. Fieremans. 2016. "Denoising of Diffusion MRI Using Random Matrix Theory." *NeuroImage* 142: 384–396.

Veraart, J., E. Fieremans, and D. S. Novikov. 2016. "Diffusion MRI Noise Mapping Using Random Matrix Theory." *Magnetic Resonance in Medicine* 76, no. 5: 1582–1593.

Veraart, J., J. Sijbers, S. Sunaert, A. Leemans, and B. Jeurissen. 2013. "Weighted Linear Least Squares Estimation of Diffusion MRI Parameters: Strengths, Limitations, and Pitfalls." *NeuroImage* 81: 335–346.

Vollmar, C., J. O'Muircheartaigh, G. J. Barker, et al. 2010. "Identical, but Not the Same: Intra-Site and Inter-Site Reproducibility of Fractional Anisotropy Measures on Two 3.0 T Scanners." *NeuroImage* 51, no. 4: 1384–1394.

Walsh, D. O., A. F. Gmitro, and M. W. Marcellin. 2000. "Adaptive Reconstruction of Phased Array MR Imagery." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 43, no. 5: 682–690.

Zhang, F., L. Ning, L. J. O'Donnell, and O. Pasternak. 2019. "MK-Curve-Characterizing the Relation Between Mean Kurtosis and Alterations in the Diffusion MRI Signal." *NeuroImage* 196: 68–80.

Zhao, Y., Z. Yi, L. Xiao, et al. 2022. "Joint Denoising of Diffusion-Weighted Images via Structured Low-Rank Patch Matrix Approximation." *Magnetic Resonance in Medicine* 88, no. 6: 2461–2474.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.