A Appendix

A.1 Power comparisons with other tests

While the quality index is the closest method to ours, it is not the only other method for comparing the distributions of functions. One particularly powerful method is the Functional Anderson-Darling (FAD) test of (Pomann et al., 2016). In their paper they demonstrated superior power over all other functional distribution tests except for the Rank based Band Depth Test ($BAND$) of (Lopez-Pintado and Romo, 2009), which was not compared against. We compare our method KD against the Quality Index (QI), FAD, and $BAND$ under the same simulation settings as in the main paper.

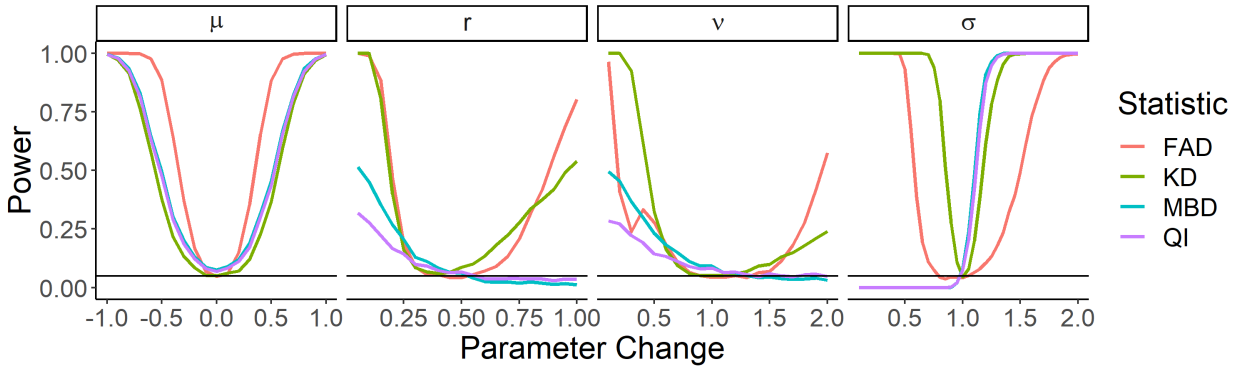

Figure 1: Power of KD, QI, FAD, and $BAND$ in detecting changes in the four parameters in the Gaussian process. Mean, Range and smoothness are presented as shifts of parameters in Y from X. Standard deviation is presented as a multiple of standard deviation in X.

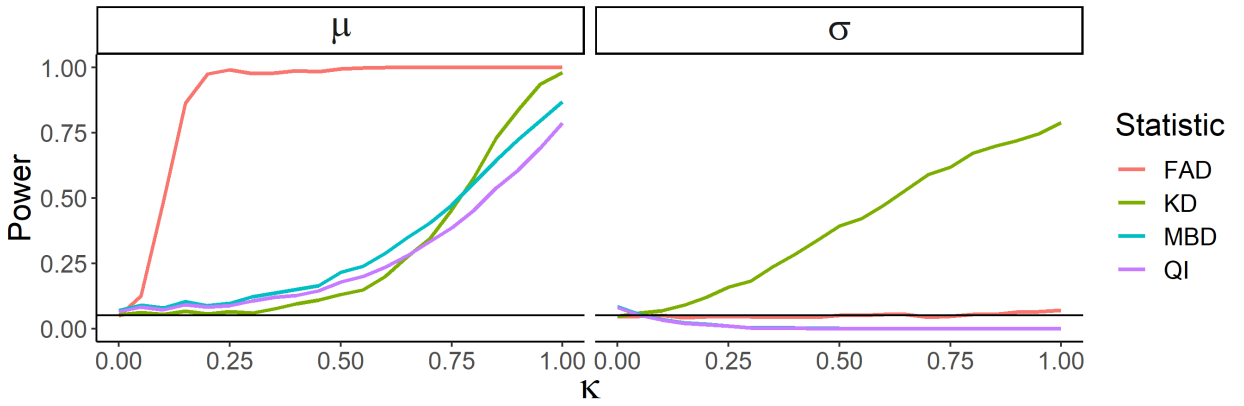

Figure 2: Power of KD, QI, FAD, and BAND in detecting changes in the four parameters in the Gaussian process. Mean, Range and smoothness are presented as shifts of parameters in Y from X. Standard deviation is presented as a multiple of standard deviation in X.

The FAD method is extremely powerful against changes in the mean of the data, however compared with the depth based methods its noticeably less powerful against variance changes (Figure 1). Under the heterogeneous changes (Figure 2) our test is still the only test to maintain its power in detecting heterogeneous variance changes.

A.2 Convergence under a non-Gaussian Process

Because our test does not depend on any parametric assumptions of the data we wanted to see how convergence, size, and power were maintained when the data came from a markedly Non-Gaussian process. For these simulations we used the same settings as in the main paper's simulations except that the functions were generated with a multivariate t distribution instead of a multivariate Gaussian distribution. We analogously denote these functions as coming from a *t-process*.

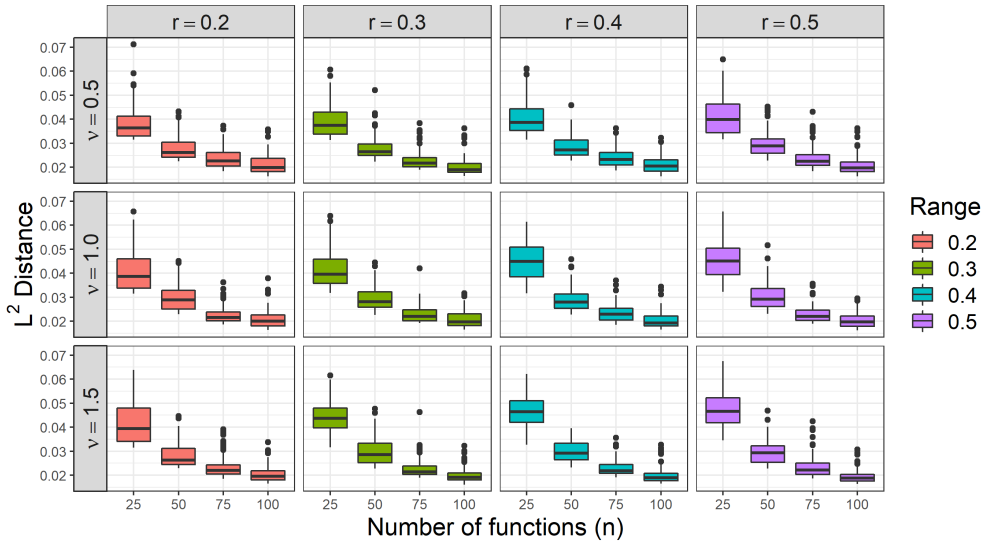

Figure 3: $\mathbb{L}^2$ distance between the permutation distribution and the Kolmogorov distribution under 12 different range and smoothness settings. Non Gaussian Process.

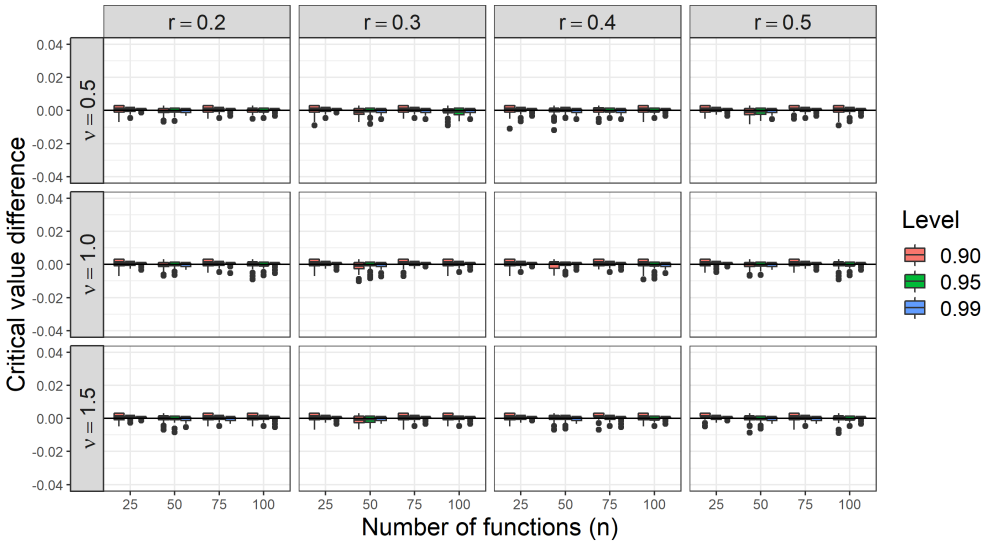

Figure 4: Kolmogorov critical values minus permutation critical values at three common test levels: 0.90, 0.95. 0.99 under 12 different range and smoothness settings. Non Gaussian Process.

Under a t-process, convergence in $Ł^2$ is observed to be slower than the corresponding Gaussian process. Critical values, however, are almost immediately unbiased verses their asymptotic counterparts. Together these indicate that the distribution of KD is harder to

approximate when the data is heavy tailed, but that this is relatively unimpactful since decisions regarding significance are unaffected by using the asymptotic distribution.

A.3 Size under a Non-Gaussian Process

We next looked at the size under t-process data. Size is controlled at relatively the same levels as when Gaussian process data was used. This is due to the critical values of the permutation distribution and the asymptotic distribution being in near agrement, even at small sample sizes. The same pattern of needed sufficient range or smoothness to achieve the nominal level is still observed.

		$\nu = 0.5$				$\nu = 1.0$				$\nu = 1.5$			
n	m	r = 0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
50	50	0.16	0.11	0.07	0.07	0.08	0.06	0.06	0.05	0.06	0.05	0.05	0.04
		(0.25)	(0.19)	(0.14)	(0.14)	(0.14)	(0.12)	(0.11)	(0.10)	(0.12)	(0.11)	(0.10)	(0.08)
50	100	0.13	0.09	0.08	0.06	0.08	0.07	0.05	0.05	0.08	0.05	0.05	0.05
		(0.29)	(0.19)	(0.18)	(0.14)	(0.18)	(0.14)	(0.11)	(0.09)	(0.15)	(0.10)	(0.10)	(0.09)
50	200	0.14	0.08	0.08	0.06	0.08	0.06	0.07	0.05	0.07	0.06	0.05	0.06
		(0.34)	(0.22)	(0.19)	(0.16)	(0.18)	(0.12)	(0.12)	(0.10)	(0.15)	(0.11)	(0.11)	(0.10)
50	300	0.14	0.09	0.09	0.06	0.08	0.06	0.06	0.05	0.06	0.06	0.06	0.05
		(0.36)	(0.26)	(0.22)	(0.17)	(0.20)	(0.16)	(0.13)	(0.12)	(0.17)	(0.12)	(0.10)	(0.10)
100	50	0.15	0.09	0.07	0.06	0.08	0.05	0.06	0.05	0.05	0.04	0.04	0.05
		(0.15)	(0.12)	(0.10)	(0.09)	(0.10)	(0.08)	(0.08)	(0.06)	(0.08)	(0.08)	(0.07)	(0.07)
100	100	0.08	0.07	0.05	0.06	0.05	0.05	0.05	0.04	0.06	0.05	0.04	0.04
		(0.16)	(0.12)	(0.10)	(0.09)	(0.12)	(0.10)	(0.08)	(0.07)	(0.09)	(0.09)	(0.08)	(0.08)
100	200	0.10	0.06	0.06	0.05	0.07	0.05	0.05	0.05	0.06	0.06	0.04	0.05
		(0.21)	(0.14)	(0.12)	(0.11)	(0.12)	(0.11)	(0.09)	(0.08)	(0.10)	(0.09)	(0.08)	(0.07)
100	300	0.10	0.07	0.05	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
		(0.23)	(0.16)	(0.13)	(0.12)	(0.13)	(0.10)	(0.10)	(0.09)	(0.11)	(0.08)	(0.09)	(0.08)
200	50	0.15	0.10	0.08	0.07	0.08	0.07	0.05	0.06	0.07	0.06	0.05	0.06
		(0.09)	(0.07)	(0.08)	(0.08)	(0.07)	(0.08)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
200	100	0.10	0.08	0.06	0.05	0.06	0.05	0.05	0.04	0.05	0.05	0.05	0.05
		(0.10)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.06)	(0.07)	(0.08)	(0.07)	(0.06)	(0.07)
200	200	0.08	0.06	0.06	0.05	0.06	0.05	0.05	0.04	0.06	0.06	0.05	0.05
		(0.14)	(0.10)	(0.09)	(0.09)	(0.10)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.06)
200	300	0.09	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.06
		(0.14)	(0.11)	(0.11)	(0.10)	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)	(0.06)	(0.07)
300	50	0.14	0.09	0.07	0.07	0.07	0.06	0.06	0.05	0.06	0.06	0.05	0.06
		(0.08)	(0.07)	(0.06)	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)	(0.07)	(0.05)	(0.06)	(0.06)
300	100	0.10	0.07	0.06	0.06	0.07	0.05	0.05	0.06	0.06	0.06	0.06	0.05
		(0.09)	(0.08)	(0.06)	(0.07)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
300	200	0.08	0.07	0.07	0.05	0.06	0.06	0.06	0.05	0.06	0.07	0.06	0.06
		(0.11)	(0.09)	(0.08)	(0.07)	(0.08)	(0.07)	(0.06)	(0.07)	(0.07)	(0.08)	(0.06)	(0.06)
300	300	0.08	0.07	0.05	0.05	0.06	0.06	0.05	0.05	0.06	0.05	0.05	0.05
		(0.10)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.06)	(0.07)	(0.05)	(0.07)

Table 1: Size of KD and QI (in parenthesis) under 12 combinations of range, r, and smoothness, ν, and 16 combinations of sample sizes, n and m, for X and Y respectively. Data was generated from a Non-Gaussian process.

A.4 Power comparisons under a Non-Gaussian Process

Finally we considered power under homogeneous and heterogeneous parameter changes under Non-Gaussian data (t-process). The same settings to test power in the main paper's simulations were against used to generate data. As in the convergence and size simulation the sampled functions were generated from a t-process with 3 degrees of freedom.. The power curves (Figures 5 and 6) are generally flatter than the corresponding power curves under a Gaussian process, however the relationship between methods remains the same. FAD still dominates detecting changes in the mean and KD, MBD, and QI dominate detecting changes in the standard deviation. All methods lose considerable power in detecting range and smoothness changes. Notably the FAD test ran into computational issues trying to estimate the functional principal components due the t-process frequently generating very outlying curves.

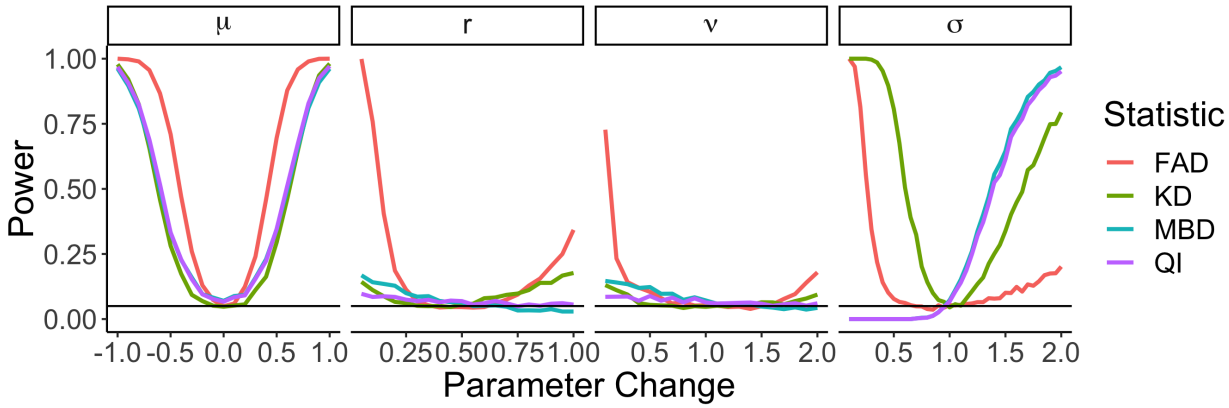

Figure 5: Power of KD, QI, FAD, and $BAND$ in detecting changes in the four parameters in the Gaussian process. Mean, Range and smoothness are presented as shifts of parameters in Y from X. Standard deviation is presented as a multiple of standard deviation in X.

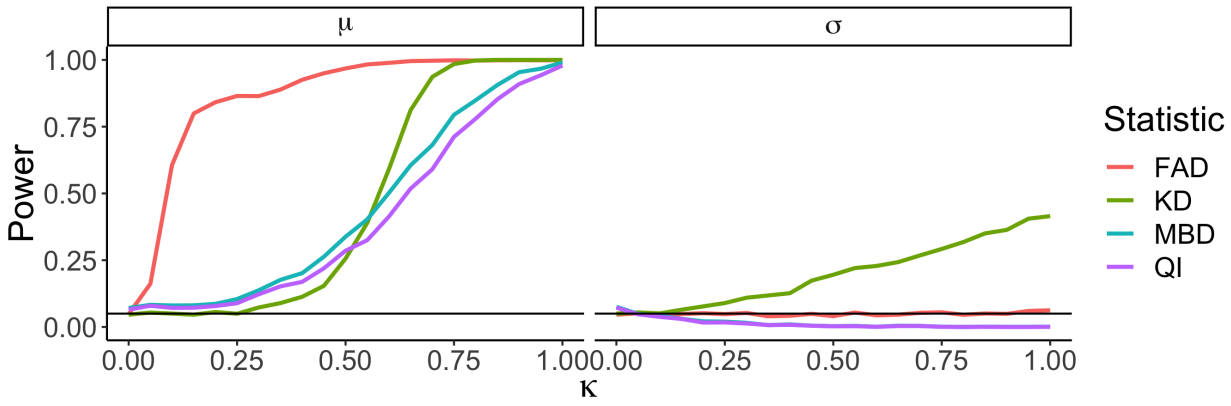

Figure 6: Power of KD, QI, FAD, and $BAND$ in detecting changes in the four parameters in the Gaussian process. Mean, Range and smoothness are presented as shifts of parameters in Y from X. Standard deviation is presented as a multiple of standard deviation in X.

A.5 FAD v.s. KD on PHYDA

The preceding power plots show that their is no clear dominating method, between FAD and KD across all of the parameters in the Gaussian and Non-Gaussian simulations. FAD clearly detects mean differences better and KD clearly detects standard deviation differences better. This is particularly true in the case of heterogeneous mean and variance changes under a t-process (Figure 6), i.e. the more realistic setting. We argue that because FAD fails to detect heterogeneous changes in the variance, it misses out on the crucial finding in our data analysis, namely that the analysis ensembles become more distinct from the background over time. These changes appear to be primarily driven by a downward trend in the variance of the analysis state (see Figure 7).

As can be seen in Figure 7, the average difference between the background and analysis remains relatively constant over time. Because the averages differences are even slightly different from 0, FAD has no issue with detecting a significant difference. The real differentiator is how the ratio of the variances changes over time. With the exception of the very

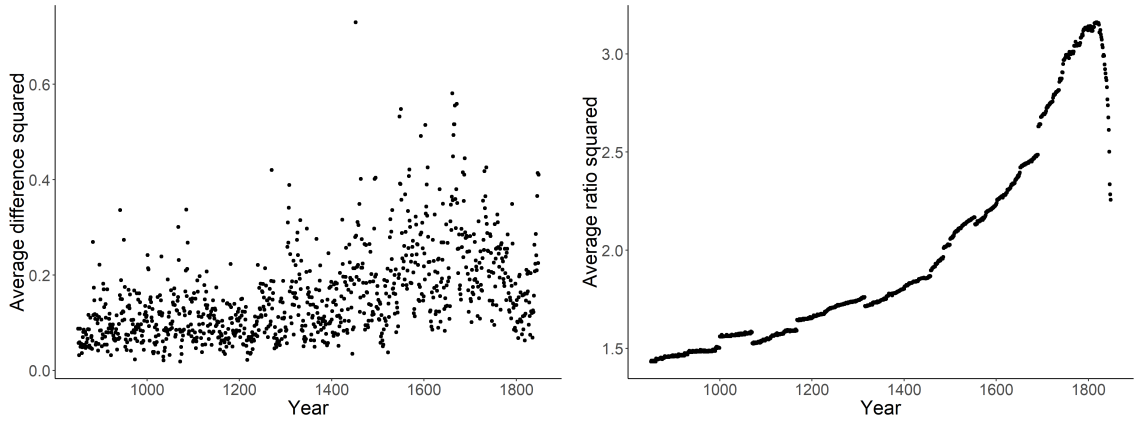

Figure 7: **Left:** Average squared pointwise mean differences between the background and analysis ensembles for each year in the reconstruction. **Right:** Average squared pointwise ratio of the background and analysis ensemble standard deviations for each year in the reconstruction.

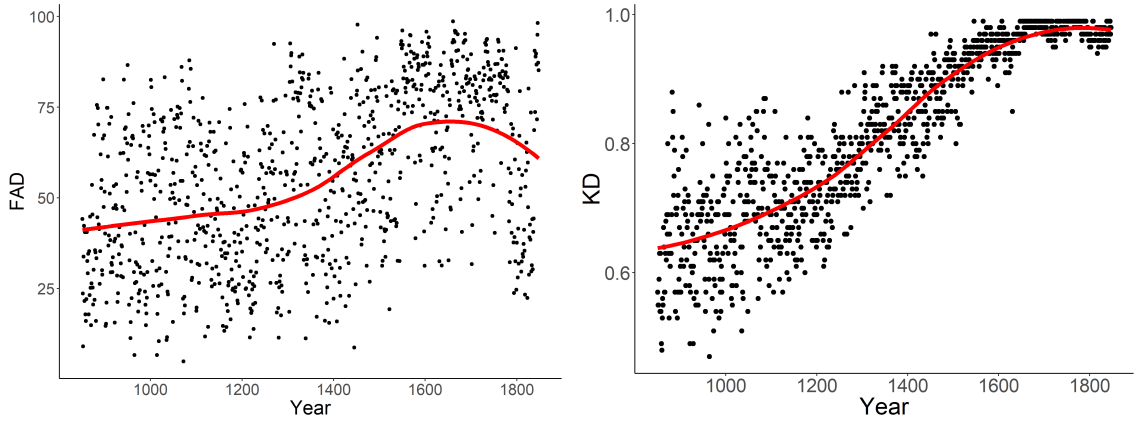

Figure 8: FAD vs KD values on the PHYDA climate data over the reconstruction period 850CE to 1850CE. Both tests detect significant distribution changes, but FAD is primarily driven by the mean differences. KD derives its value from the mean changes, the increase standard deviation changes, and higher moment changes not displayed here.

end of the reconstruction, the average variance ratio increases almost monotonically. This pattern reveals that one of the primary effects of including additional proxies is a reduction in uncertainty. This near monotonic increase in uncertainty reduction is largely reflected in the associated time series of K values (Figure 8). If we compare against the values of FAD over time (Figure 8) we can see that it does not register this aspect of the distribution change. FAD generally only follows the trend of the mean differences, while KD follows both.

References

Lopez-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association 104*, 718–734.

Pomann, G., A. Staicu, and S. Ghosh (2016). A two sample distribution free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C Applied Statistics 65*(3), 395–414.