Picking the "best" noise model; evaluating the results from est_noise

John Langbein

Revised; May 2021

The program, est_noise, provides the user a variety of colored noise functions that models and measures the background noise in various time-series. Amongst the noise models potentially tested with est_noise, it is desired to have a method to quantitively select one "best" noise model over competing noise models. The following describes the results of multiple experiments of simulations of colored noise where the coefficients of different noise models are estimated. With each simulation, there is an associated Max. Likelihood estimate (MLE) (actually, log likelihood). Can the differences in MLE, dMLE between two competing models, be used to statistically determine whether the simplest noise model, termed the null model, can be rejected in favor of the more complex? If so, then what is that threshold and its associated confidence? Does that threshold depend upon the length of the time series and does that threshold depend upon the "strength" of colored noise relative to white noise? In part, this study replicates Figure 4 in Langbein (2004). Noted then, and replicated here is the anomalous behavior of Gauss-Markov noise.

Five conclusions are:
1) Threshold dMLE has a some dependence upon the length of the time series.
2) Threshold dMLE depends upon both the candidate noise model and the null model.
3) Threshold dMLE has only a small dependence upon the size of the colored portion of the noise in the time-series.
4) Two other metrics, Akaike and Bayesian Information Criterion (AIC and BIC), that are related to MLE are also evaluated as a means of selecting the better model, but neither of these metrics are recommended.

The basic test simulates multiple time series that are a combination of white noise plus random walk noise and this is identified as the "Null" noise model. For the experiment described here, I've created 5000 such time series. The underlying model consists of 0.5 mm of white noise and random walk consisting of amplitudes of either 0.1, 0.3 1.0, 3.0 or 5.0 mm/yr$^{0.5}$. Along with estimating the parameters of the noise model(s), est_noise estimates the rate, and the amplitude/phase of annual and semi-annual period sinusoids. Neither the rate nor the sinusoids are prescribed to be present in the simulated time-series. Figure 1 shows the power spectra of these simulated time series. With each simulated time-series, I use est_noise to compute the parameters of various noise models, (all include white noise), random walk (RW), Power Law (PL), Flicker plus Random Walk (FLRW), Band-passed filtered plus RW (RWBP), combination of BP, FL, and RW (FLRWBP), FOGM and GM (or generalized Gauss-Markov G-GM). For each of these noise models, MLE is saved and compared with the null RW model yielding dMLE for FLRW – RW, PL – RW, and so on.

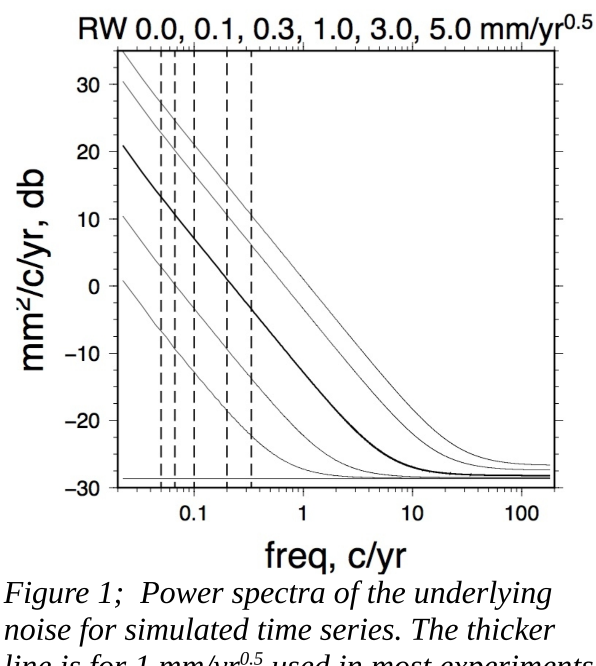

Figure 1; Power spectra of the underlying noise for simulated time series. The thicker line is for 1 mm/yr$^{0.5}$ used in most experiments. The thin, dashed lines represent the 5 lengths, 20, 15 (1/15 c/yr), 10, 5, and 3 years.

One set of experiments simulated time series of different lengths, 3, 5, 10, 15 and 20 years. These all used the 1 mm/yr$^{0.5}$ value of random walk. A second set of experiments uses 10 years of simulated data and varied the size of the simulated random walk. Simulations were done 5000 times with the expectation that dMLE can be evaluated as a threshold to reject the null model. A third set of experiments tests the impact of dMLE of estimating rate when the rate parameter input to est_noise is set to zero. With each of these experiments, the dMLEs are tabulated and sorted such that probability plots are constructed, Figure 2. For the examples that follow, I've selected the dMLEs, one that corresponds to a 99% confidence in rejecting the null, RW model. (This corresponds to 0.01 in the probability plot in Figure 2). For 5000 simulations, the threshold with 99% confidence-level in dMLE would have 50 values of dMLE > the threshold.,

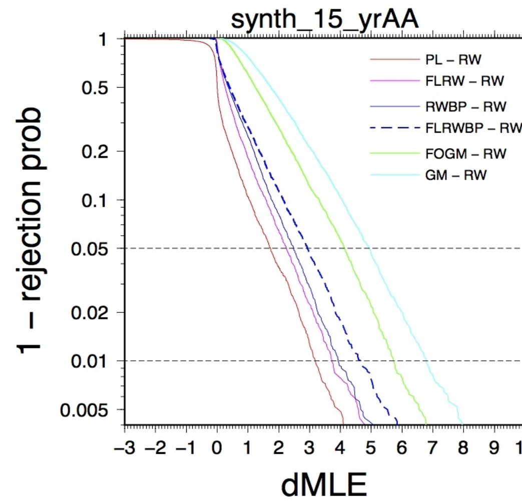

Figure 2, Example of a cumulative histogram of dMLE of 6 noise models relative to RW noise. For reference, the 0.05 and 0.01 levels are indicated.

1) dMLE vs Time series length: Figure 3a show the results of determining the threshold of dMLE for a variety of noise models compared against the null for different lengths of time series, from 3 to 20 years. Each simulated time series has a background rate of zero and 1 mm/yr$^{0.5}$ of random walk and 0.5 mm of white noise and spans a prescribed amount of time, 3, 5, 10, 15 or 20 years. Then for each simulated time-series, est_noise simultaneously estimates the rate, the amplitude of two sinusoids (annual and semi-annual periodicity), the white noise and the random-walk amplitude (the rate and periodicities are typically estimated for most GPS time-series). It was assumed that the noise model is additive as described by Langbein (2017) such that est_noise runs quickly. (There are no gaps in the simulated time-series). For each simulated time-series, the value of MLE is saved. The first "run" with est_noise assumes the RW model and is identified as the null-model. Then for each time series, more complex noise models are evaluated as they have at least one more free parameter than RW. Those models are PL, FLRW, FOGM, GM, RWBP, and FLRWBP. These simulations were repeated 5000 times providing the ability to determine the distribution of dMLE in terms of a probability plot based upon numerically sorting dMLE. Since the underlying model of noise is RW, the change in MLE associated with the more complex model, which includes a RW term, should be 0. In reality, the value of dMLE will be "close to zero" but there will be a few dMLEs that exceed 0. With each noise model, the difference in MLE is computed and saved. After a numerical sort of the dMLEs, the 99% level of dMLE is identified; that is for 5000 simulations, the dMLE for which there are 50 values greater is identified as the 99% level to reject the null hypothesis. The results of these simulations are shown in Figure 3a. Note, for the 10-year time span, this experiment was run four times to test the ability of determining dMLE with 5000 simulations. Several features are apparent: 1) dMLE depends upon the pair of models used for comparison. For example, when comparing PL against RW with 10 years of data, any dMLE greater than ~4.0 suggests that the null model can be rejected at the 99% confidence level. On the other hand, dMLE only needs to exceed 3.3 to reject the null RW-model in favor of FLRW model. 2) The threshold dMLEs for any noise model using GM noise is roughly twice the other dMLEs that use BP, FLRW, or PL noise. (This is discussed below.) 3) dMLE has an inverse dependence upon the length of time of the time series. 4) The robustness of these threshold have been tested by running the 10 year simulation 3 more times each with 5000 synthetic time-series. The range in 99% C-I for dMLE to reject RW in favor of PL is from 3.97 to 4.31. The robustness is illustrated in Figure 3a which for the 10 year intervals, show the threshold dMLEs that determined to reject the null, RW model.

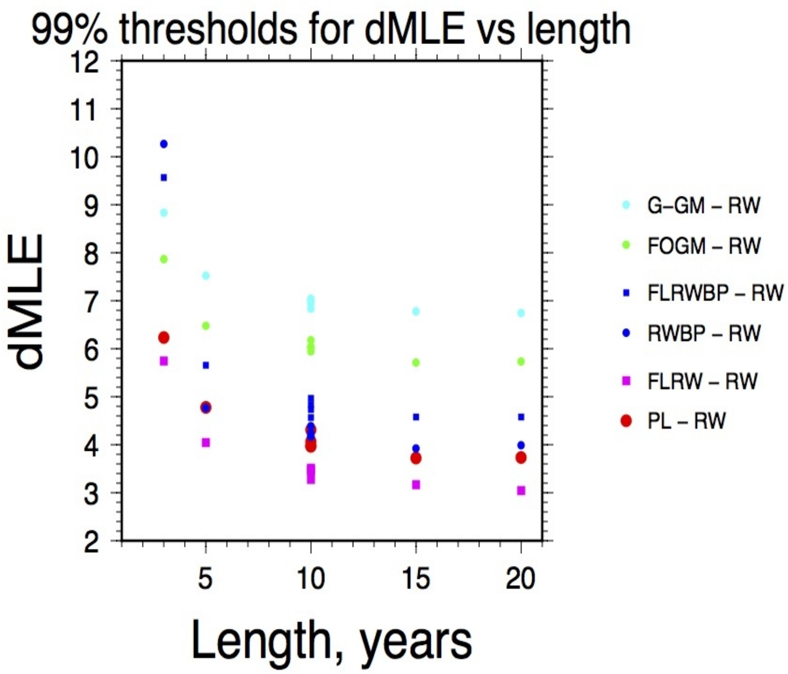

Figure 3a. Differences in MLE from a variety of noise models relative to the NULL random walk model of noise for a variety of lengths of time series. In this set of experiments, RW is 1 mm/yr$^{0.5}$. The values are listed in Appendix 2.

The experiments in Figure 3a were repeated using different amplitudes of RW noise as the null model. These are shown in Figures 3b and 3c where the RW is 0.1 and 5.0 mm/yr$^{0.5}$.

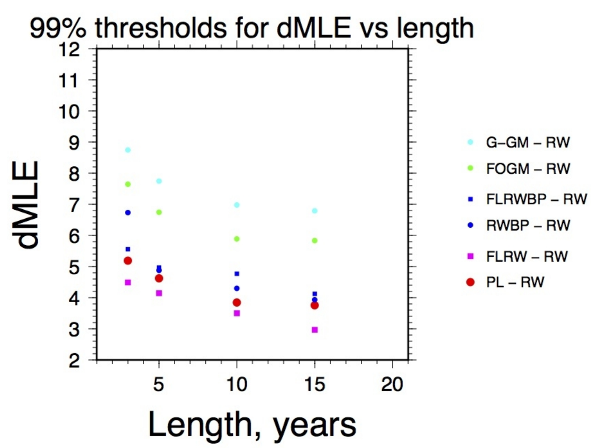

Figure 3b. Same experiments as Figure 3a but the null RW model is 5.0 mm/yr$^{0.5}$

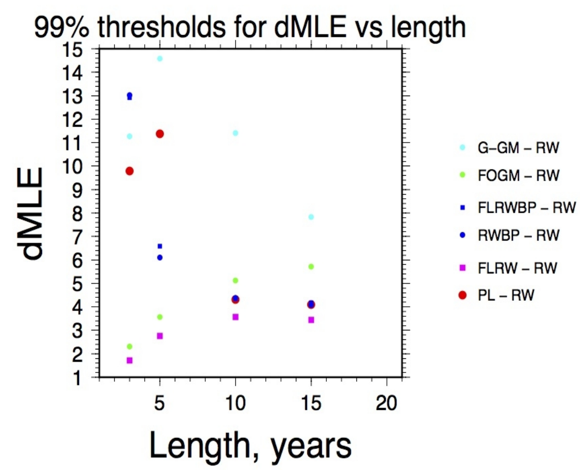

Figure 3c. Same experiments but the null RW model is 0.1 mm/yr$^{0.5}$

The results shown in Figure 3b, where the input RW noise is 5 mm/yr$^{0.5}$ essentially replicates the results in Figure 3a which used and input RW of 1 mm/yr$^{0.5}$. On the other hand, if the input RW noise is reduced to 0.1 mm/yr$^{0.5}$, the results for the shorter, 3 and 5 year lengths, are anomalous for dMLE relative to time series with larger values of RW. Basically, with low amplitude RW noise, as shown in the power spectra in Figure 1, RW amplitude is only marginally detectable relative to white noise at these shorter intervals. Consequently, trying to resolve other low amplitude components of a more complex noise model is also difficult.

A further set of test is described in Appendix 1 where, instead of using RW as the null model, FLRW is the null model and it is tested with the more complex FLRWBP noise-model.

2. dMLE vs Amplitude of Random Walk noise: Figure 4 shows the results of threshold dMLE for different amplitudes of random-walk noise in simulated time series. For these sets of experiments, simulated noise were created using 0.1, 0.3, 1.0, 3.0 and 5.0 mm/yr$^{0.5}$ random walk added to 0.5 mm of white noise. However, the experiments used only 10 years of simulated data.

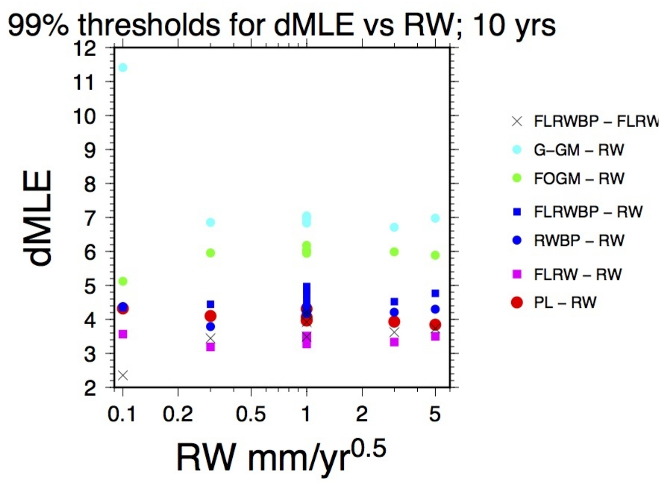

Figure 4. Differences in MLE from a variety of noise models relative to the NULL random walk model of noise for a variety of levels of random-walk noise.

Unlike the dependence of dMLE on time series length discussed in the previous section, dMLE does not show obvious dependence on the amplitude of the underlying noise. For instance, the 99% threshold of dMLE to reject the RW model in favor of FLRW is 3.6 if the underlying RW amplitude is 0.1 mm/yr$^{0.5}$ and 3.5 if the underlying RW amplitude is 5.0 mm/yr$^{0.5}$. However, this observation breaks down for trying to distinguish any GM noise from noise source with a low level of RW noise (0.1 mm/yr$^{0.5}$,green and cyan colored symbols).

Also tabulated and displayed is likely scenario where one might assume that the null model is FLRW and wish to test whether the addition of BP noise could better characterize the background rate. The threshold dMLEs are shown with 'x' in Figure 4.

3. Impact of estimating rate in evaluating dMLE; correlation with FOGM/GM noise models. This test consisted of 600 simulations of 5 year long time-series with 1 mm/yr$^{0.5}$ RW plus 0.5 mm white noise. For each simulated time series, est_noise is run first with estimating the rate plus the parameters of the seven noise models, then a second time where the rate is assumed to be zero. The probability plots for both of these “runs” are shown in Figure 5 and 6.

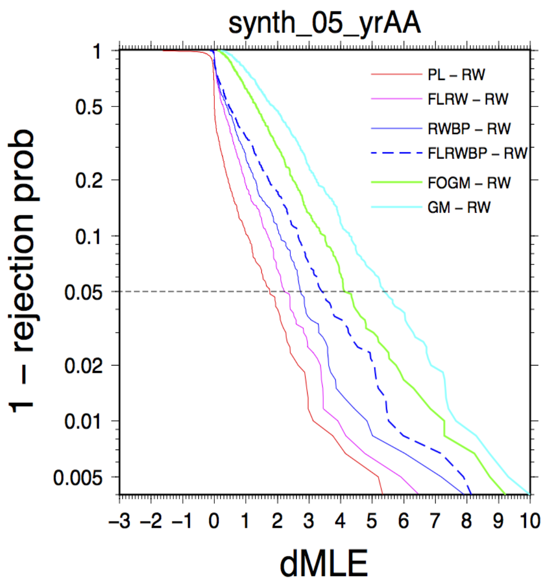

Figure 5. Probability plot of dMLE from 600 simulations of 5 years of RW noise when rate is estimated

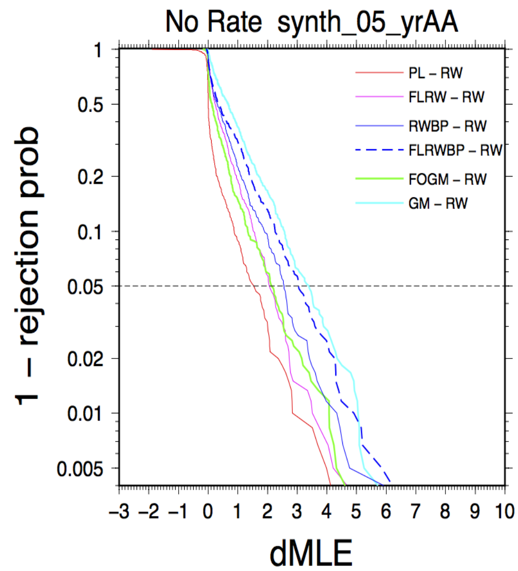

Figure 6. Probability plot of dMLE from 600 simulations of 5 years of RW noise when rate is NOT estimated

The key difference between the two probability plots is the lower dispersion of the rejection criteria of dMLE when rate is not estimated. This is particularity evident with the FOGM and the GM models when compared to the RW model, green and cyan lines, respectively. The differences in the 95% C-I are shown in Table 1 where the FOGM and GM models comparisons are highlighted.

Table 1: dMLE thresholds from 600 simulations testing the impact of estimate rate, 5 year time-series

Model comparison	**dMLE, 95% C-I rejection threshold**	
	Rate estimated	**Rate NOT estimated**
PL – RW	2.19	2.04
FLRW - RW	1.75	1.47
FOGM - RW	***4.09***	***2.10***
GM - RW	***5.34***	***3.30***
BP - RW	2.71	2.54
FLRWBP - RW	3.35	3.04

For the PL, FLRW, and both BP models, the impact estimating rate on dMLE is fairly small, roughly 6 to 20%. On the other hand, the impact of rate-estimation for the GM models is significant, between 60 and 95% contrast.

An alternative view of the impact on rate with the various noise models is plot the histograms of the ratio of estimated rate to the estimated standard-error in rate. The results for all 7 types of noise models are shown in Figure 7. For simulations for which the underlying rate is zero, the histograms of rate should cluster about zero, which does occur for all of the noise models. However, the rate is normalized by its computed standard error; the standard error is derived directly from both the assumed function that represent the background noise and its estimated values. For RW noise, the range for normalized rate is roughly +/- 3, consistent with expectations that 99% of the rate falls within 3-sigma of the “error bar” in rate. Broadly speaking, the +/-3 range for normalized rate is also seen for PL, FLRW, and both BP noise models. On the other hand, visually, all 6 of these histogram slightly exceed the +/-3 range and include a few outliers, too.

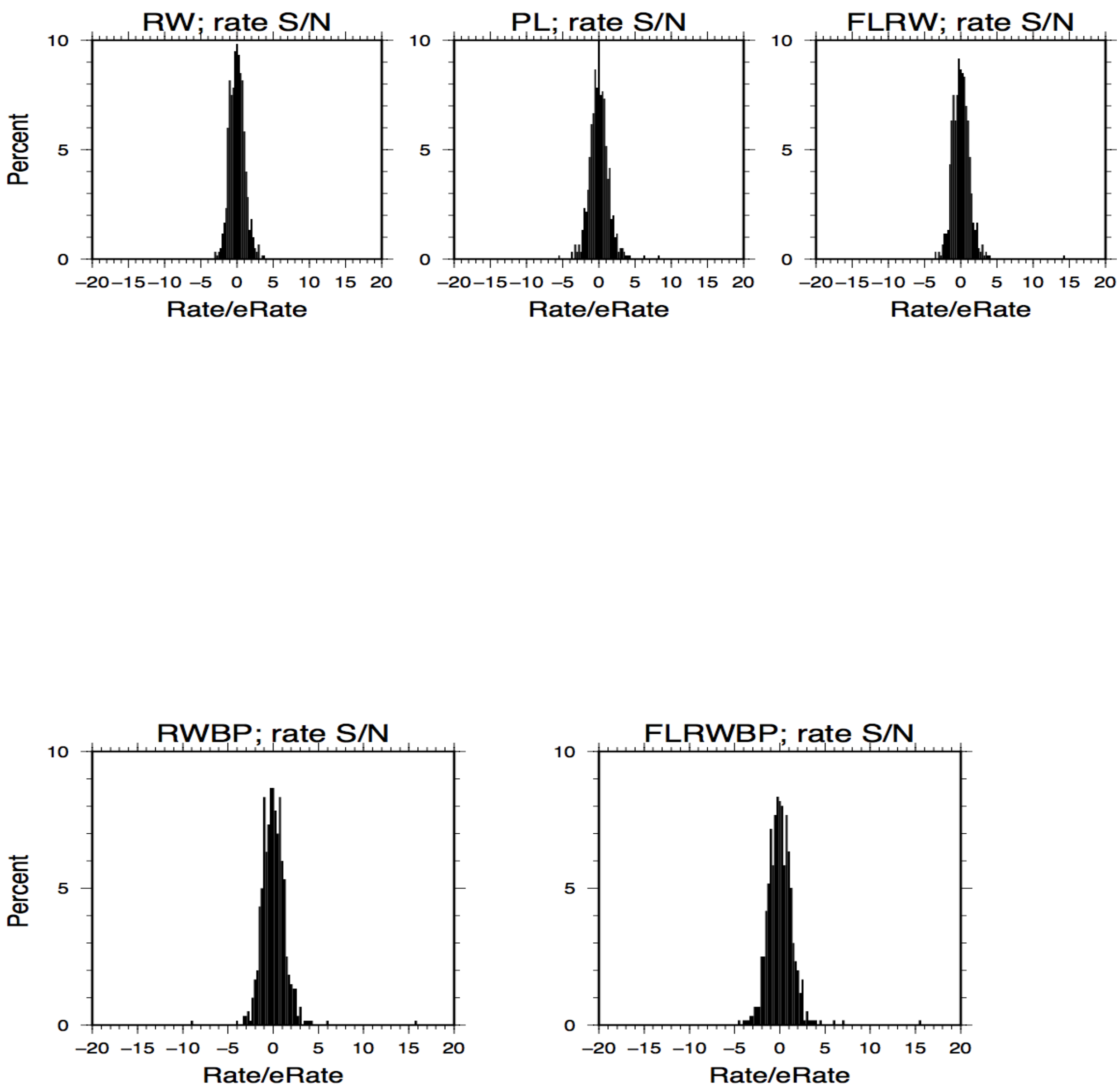

Figure 7 Histograms of estimates of normalized rate (rate divided by the error in rate) using 5 different noise models from 600 simulations of 5 years of data for which underlying rate is zero.

In contrast, the histograms for both GM noise models show that the nominal range for normalized rates vastly exceeds +/-3 and is closer to +/-10, shown in Figure 7. **NOTE – Version 7.30 fixes this problem,** so ignore the two histograms that involve GM noise**.**

4. Are either AIC and/or BIC better than MLE to discriminate between models?. Two other metrics have been used in the "literature" that might help with deciding which of the two competing models is best, Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both of these metrics are output by est_noise, and importantly, they are calculated directly from MLE. However, unlike MLE, AIC also includes a term that represents the number of unknown parameters, which for est_noise, include the time-dependent parameters (ie, rate and sinusoidal terms) and the parameters of the noise model. Therefore, the number of unknowns will increase as a more complex noise model is compared with the null model.

AIC = 2*m - 2*MLE

where m is the number of unknowns.

Like AIC, BIC incorporates the number of unknowns and, in addition, the number of observations into a form:

BIC= m*ln(n)-2*MLE

where n is number of observations.

With some simple arithmetic, dAIC and dBIC are computed as:

dAIC = $2(m_b – m_n)$ -2*dMLE

dBIC = $(m_b – m_n)$*ln(n) – 2*dMLE

where m_n are the number of parameters in the null model and m_b are the number of parameters in the more complex model. For instance, $(m_b – m_n)$ is 1.0 for comparing FLRW, PL, and FOGM against the null RW model, but becomes 3.0 when comparing the FLRWBP with the null RW model (note that the BP model has two additional parameters, the number of poles and the amplitude of BP noise). Note that sense for thresholds for rejecting the null hypothesis using either AIC or BIC is opposite from MLE; the dAIC/dBIC must be less than the threshold to reject the null model.

Assuming that AIC and BIC accounts correctly for the number of unknowns and the length of the data (BIC), I would anticipate that one or both of these metrics would show invariance with the number of model parameters and/or the length of the data set.

Figure 8 and 9 reproduce the results shown in Figure 3a but use dAIC (Figure 8) and dBIC (Figure 9). For better comparison with the dMLE plot in Figure 3a, I have chosen to divide by two both dAIC and dBIC.

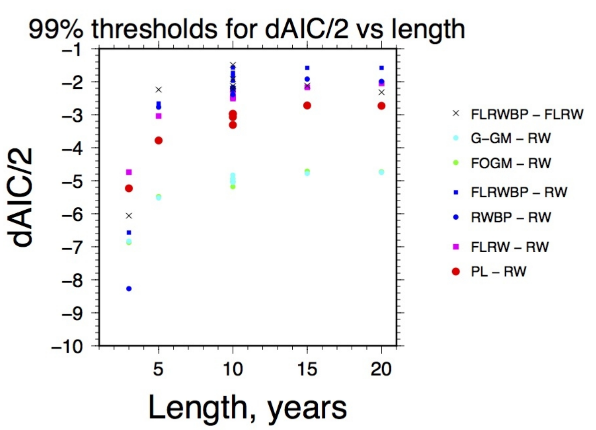

Figure 8. Results from Figure 3a (Input RW is 1 mm/yr$^{0.5}$) that rescale dMLE to dAIC; dAIC is plotted against the length of the simulated data.

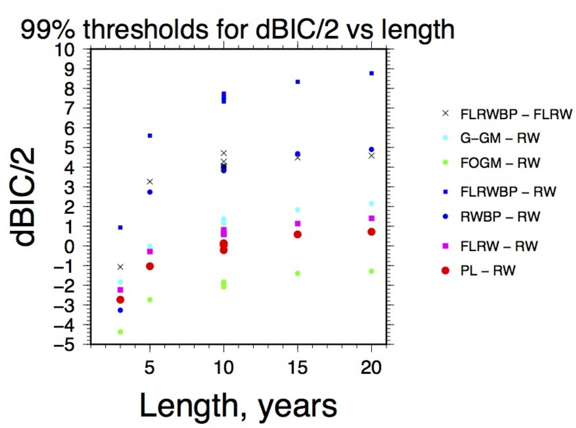

Figure 9. Results from Figure 3a (Input RW is 1 mm/yr$^{0.5}$) that rescale dMLE to dBIC; dBIC is plotted against the length of the simulated data.

The results shown in Figures 8 and 9 show that neither that dAIC nor dBIC removes the dependence on thetime-series length nor do they remove the dependence upon model types. In fact, they dBIC results amplify the length dependence. Consequently, I do not recommend using either AIC or BIC as means to discriminate between models.

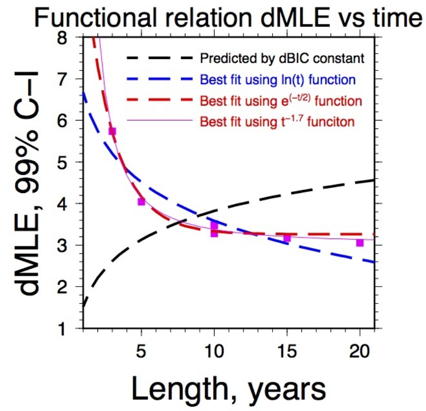

Figure 10. dMLE 99% thresholds for FLRW - RW test from Figure 3a are plotted with magenta squares. In addition, four functions that relate time-series length are plotted.

The poor performance of dBIC is demonstrated in Figure 10. Here, the dMLE threshold for the 99% confidence level is replotted from Figure 3a for testing the FLRW model against the null, RW model as magenta squares. If dBIC is suppose to be invariant with time-series length, then dMLE should be proportional to ln(365.25*t), with t being the length of the time-series. The value of this function increases with time-series length, shown with a dashed black curve, but that prediction is not consistent with the threshold dMLEs. On the other hand, three other functions are evaluated and their best fits to the threshold dMLEs are shown in Figure 10. These functions, ln(t), $e^{-t/2}$, and $t^{-1.7}$, are shown with a dashed blue curve, dashed red curve, and a solid magenta curve. The best fitting, at least for this set of dMLEs, is the $t^{-1.7}$ curve; although the exponential is a close second.

Flicker Noise Only – Maximum detectable RW noise needed to reject FL in favor of FLRW noise: This part is only a partial note with respect to detecting RW noise in the presence of FL noise. Often with GNSS time series, especially with vertical displacement solutions, est_noise suggests that the best noise model is that of FL (and white noise). However, work with borehole strainmeters and other ultra-sensitive instruments used to measure crustal deformation indicates that a major source of noise has RW characteristics that are probably related to the coupling that instrument to the ground; the belief is that the ground exhibits Brownian motion which is a random-walk. Although other statistical processes such as a strong flicker component, may mask the presence of random-walk. Although random-walk might not be detectable, its presence can impact the estimates of rate uncertainty. This is explored more in Langbein (2012).

Concluding remarks: Using 5000 simulations, I am able to characterize some of the properties of dMLE used as a threshold to distinguish between the null model and a more complex noise model. I found that the threshold value of dMLE will depend upon the two models being compared. In addition, the threshold dMLE shows a dependence on the length of the time-series. On the other hand, dMLE seems insensitive to the amount amplitude of noise in the time series.

Like earlier analysis by Langbein (2004), the threshold dMLE for Gauss-Markov noise tends to be much higher, by about 2x, than thresholds for the other noise models including Band-passed filtered. However, that larger threshold can be understood in terms of trade-off between rate and the parameters that describe the GM noise.

Although dAIC and dBIC factor in the number of unknowns and the length of the time-series, neither of these statistical parameters provide any significant help relative to dMLE with discriminating a more complex model from the null model.

Finally, although dMLE is a valuable metric, I suggest that one needs to examine the plots of “drift” or wander real data in comparison to the drift computed from the estimated noise models.

APPENDIX 1:

Testing FLRWBP against the null, FLRW model. Following earlier recommendation of Langbein (2012?), it is advantageous to use the FLRW noise model to represent the background noise in GPS time series. This cuts-out computations needed for RW, FL, and PL models as it is assumed that GPS colored noise is a combination of random-walk noise of the GPS monument and flicker noise introduced by the GPS system. In addition, relative to using the nearly equivalent PL model of noise, the presence of the RW component in the FLRW model provides a slightly, more conservative estimate on the "error-bars" in the parameters describing the time-dependence, namely the rate. Consequently, using FLRW as the base noise model, one only needs to test whether additional BP noise is present in the time-series. Figure A1 shows the results of determining the threshold dMLE to discriminate between these two choices.

Like the tests used to construct Figures 2 and 3 in the main text, the results presented use 5000 simulations and a numerical sort of the difference between the MLEs for the more complex FLRWBP model and the null, FLRW model. However, rather than using synthetic time series using noise having only RW (plus WN), the time series for both methods 1a and 1b were constructed having each 0.5 mm/yr$^{0.5}$ RW, 0.5 mm/yr$^{0.25}$ FL and 0.5 mm of WN. The only difference between the two "methods" is that 1a was done by me and 1b was done by Jerry Svarc using the same program, est_noise, but coded in a different unix shell-script than mine. Consequently, the results shown in Figure A1, (cross and x) essentially overlay each other.

In contrast, method 2 is a re-sorting of the results of the experiments that went into Figure 3a. In those experiments, the synthetic time series have 1 mm/yr$^{0.5}$ random walk with no FL contribution. In this method, I simply assigned the MLE for the FLRW model as the "null" model, and recompiled and sorted a list of dMLE with the more complex FLRWBP model. The results for method 2 indicates that the threshold dMLE is about 10 to 20% larger than the threshold dMLE for method 1.

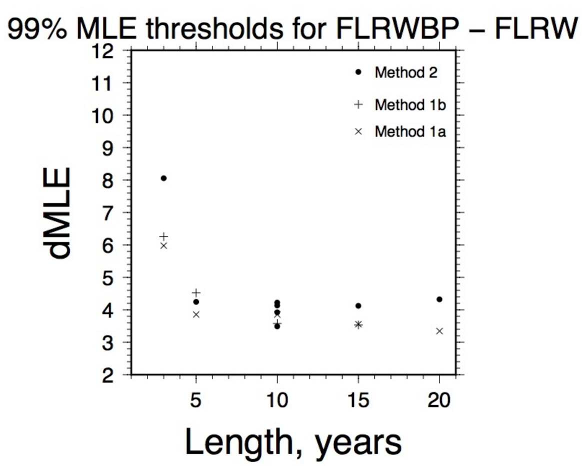

Figure A1: Threshold dMLE for rejection the null FLRW model in preference to FLRWBP noise. See text for description of the methods. The values for Method 1 are listed in Appendix 2.

APPENDIX 2:

Tables with threshold dMLE for 99% confidence level to reject the null hypothesis:

Essentially, the two tables that follow are the values that are plotted in Figures A1 and 3a. I have taken the liberty of rounding up by roughly 0.1 units, and even then, that might be not enough to be "robustly" considered a 99% C-I

FLRWBP vs null FLRW see appendix 1		
Time series length, years	**dMLE, 99% method 1a/b***	**dMLE, 99% method 2***
3	6.2	8.1
5	4.5	4.3
10	3.9	4.2
15	3.6	4.2
20	3.4	4.3

dMLE with 99% confidence level to reject null RW model						
Length of time series, years	**PL**	**FLRW**	**RWBP**	**FLRWBP**	**FOGM**	**G-GM**
3	6.3	5.8	10.4	9.7	8.0	8.9
5	4.9	4.1	4.9	5.8	6.6	7.6
10	4.1	3.4	4.2	4.8	6.0	7.0
15	3.8	3.3	4.0	4.7	5.8	6.9
20	3.8	3.1	4.1	4.7	5.8	6.8