# SOE Simulations Manuscript

*Sean Hardison, Charles Perretti, Andy Beet, Geret DePiper*

*June 29, 2018*

## Abstract

## Introduction

The development and analysis of indicators plays a key strategic role in implementing the Ecosystem Approach for a host of science, management, and intergovernmental organizations (e.g. NOAA 2006; ICES 2013; Secretariat of the Convention on Biological Diversity 2004; Perry, Livingston, and Fulton 2010; Garcia et al. 2003; Levin et al. 2009). At least partially in support of this, substantial effort has been invested in assessing indicator status and trends for the purpose of ecosystem reporting, in all of its guises (e.g. Garfield and Harvey 2016; NEFSC 2018a; NEFSC 2018b; Blanchard et al. 2010; O'Brien 2017; Butchart et al. 2010).

Ecosystem-level indicators often vary greatly with respect to the length of the series under investigation. The ultimate goal of providing integrated advice often leads analysts to truncate longer datasets; generating a consistent series length across indicators for comparison purposes (e.g. Blanchard et al. 2010; Shin and Shannon 2010; Shannon et al. 2010; Canales et al. 2015). Further reinforcing this approach is the fact that managers tend to focus on short-term issues (Secretariat of the Convention on Biological Diversity 2004; Wagner et al. 2013), which ultimately necessitates the assessment of trajectories at relatively short time scales.

These issues can lead to the use of short time series for the purpose of ecosystem reporting; i.e. less than 20 data points per indicator (Blanchard et al. 2010; Shin and Shannon 2010; Shannon et al. 2010; Canales et al. 2015; Mackas, Thomson, and Galbraith 2001; Nicholson and Jennings 2004). Statistical trend analysis of indicator data is the gold standard for managers, stakeholders, and analysts. However, in reality trend analysis in this context can be extremely difficult. Evidence indicates that the statistical power to identify trends using short time series may be limited in general (Nicholson and Jennings 2004; Wagner et al. 2013). The hydrological and climatological literature shows that autocorrelation in time series can falsely inflate trend detection rates when models are incorrectly specified assuming the independence of error terms (Kulkarni and Storch 1995; Storch 1999; Zhang et al. 2000; Wang and Swail 2001; Yue and Wang 2002; Bayazit 2015).The magnitude of assigned trends can also be inflated by the presence of autocorrelation, and both of these problems are amplified by short time series (Kulkarni and Storch 1995; Yue and Wang 2002).

Despite this, there has been no systematic investigation for the performance of models in detecting trends across the full breadth of indicators utilized in ecosystem reporting.

In this manuscript we abstract away from issues surrounding the identification and vetting of appropriate indicators, but note that this in itself can be a challenging undertaking for which Bundy, Gomez, and Cook (2017) present a survey of the literature. We focus, instead, on the ability to statistically identify trends for the broad array of indicators used in marine ecosystem reporting; ranging from large-scale climatological and oceanographic drivers through the benefits derived by human society. We use Monte Carlo simulations to assess the performance of the most commonly applied statistical models under a range of time series lengths, trend strengths, and autocorrelation regimes. The simulations are parameterized using the properties of indicators currently presented in the Mid-Atlantic and New England State of the Ecosystem Reports, which are annual ecosystem status reports tailored for the U.S. Mid-Atlantic and New England Fishery Management Councils respectively (NEFSC 2018a; NEFSC 2018b).

Results indicate that correctly identifying trends is problematic using less than 30 data points, with both Type I and Type II error common. Even under the strongest signal-noise ratio (i.e. strong trends and no autocorrelation) tests perform poorly with only ten data points. The simulations highlight problems associated with standardizing approaches across indicators, and suggest that further thought is warranted on status and trend analysis in the context of ecosystem reporting.

## Methods

### Simulations

Simulated time series were generated through the addition of $AR(1)$ autoregressive processes to first-order linear models:

$$y = X\beta + \varepsilon_t, \text{ where } \varepsilon_t = \phi\varepsilon_{t-1} + \nu_t$$

where $X$ is the $n \times p$ model matrix, $\beta$ is a vector of model coefficients, and $\varepsilon_t$ is the $AR1$ error process; the strength of which is given by $\phi$. $\nu_t$ is assumed to be derived from Gaussian white noise. The levels of $\beta$ in our study were 0.004, .051, and .147, which we combined with four levels of $\phi$: 0, .43, .8, .95. These levels were chosen based on a preliminary analysis characterizing the distribution of trend and autocorrelation strengths across 2017 State of the Ecosystem time series data. Time series used in this preliminary step were normalized by $(y - \bar{y})/y_{sd}$ before parameter identification. 1000 simulations were implemented for all combinations of trend and autocorrelation strength. To test the null hypothesis of no trend in simulated time series, we used Generalized Least Squares, Mann Kendall test, and Mann Kendall test with trend-free pre-whitening.

### Generalized least squares

The Generalized Least Squares (GLS) model fitting process used here was an iterative model-selection approach. Two first order linear and two quadratic GLS models were fit to each simulated time series and best models were chosen using small sample AIC (AICc). One of each of the two linear and quadratic GLS models were specified with first-order autocorrelated error structure identical to the error process used to generate simulations.

Under $AR(1)$ error structure, the error-covariance matrix $\Sigma$ of the GLS estimator of $\beta$ is estimated by $\Sigma = \sigma^2 P$, where $P$ is a diagonal matrix composed of error variances and autocorrelations from the data at different lag times ($\rho_s$). Error autocorrelations $\rho_s$ were estimated by restricted maximum-likelihood (REML) using the *nlme* R package. The GLS estimator $b_{GLS}$ is given by

$$b_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y,$$

where $y$ is the response vector, and $X$ is an $n \times p$ model matrix. The covariance matrix of $b_{GLS}$ is

$$\mathrm{Var}(b_{GLS}) = (X'\Sigma^{-1}X)^{-1}.$$

The second pair of first order linear and quadratic GLS models were specified with normal error structure, $N(\mu, \sigma^2)$. After selecting the best performing model using AICc, we performed likelihood ratio tests between the selected and null models specifying a maximum-likelihood approach to test the null of no trend given $\alpha = 0.05$.

### Mann Kendall test

Further tests for trend in simulated time series were performed using the Mann-Kendall test (MK) (Mann 1945; Kendall 1975) and the more robust Mann-Kendall test with trend-free pre-whitening (MK-TFPW) (Yue et al. 2002). The MK test is a non-parametric test for trend that assumes sample data are independent and identically distributed; an assumption frequently violated in time series data. Serial correlation within sample data will lead to inflated rejection rates of the null hypothesis of no trend if no correction steps are applied to the MK test (von Storch 1995), such as residual pre-whitening, although pre-whitening is known to reduce the magnitude of existing trend (Yue et al. 2002). The Mann-Kendall with trend-free pre-whitening is a step-wise procedure developed by Yue et al. 2002 to address issues introduced by pre-whitening, and is further detailed below. Under both MK and MK-TFPW frameworks, Kendall's tau statistic is given by:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(y_j - y_i),$$

where $y$ is the response vector, $n$ is the length of the series, and

$$\operatorname{sgn}(x) = \left\{ \begin{array}{ll} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{array} \right\}.$$

When there are no ties in the data, the variance of $S$ is given by

$$V(S) = \frac{n(n-1)(2n+5)}{18},$$

and the distribution of $S$ when $n \geq 8$ is approximately normal and symmetric about a mean of 0. We then perform a two-sided test for trend using the standardized $Z$ statistic:

$$\mathrm{Z} = \left\{ \begin{array}{cc} \frac{S-1}{\sqrt{V(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{V(S)}} & S < 0 \end{array} \right\},$$

which is drawn from a normal distribution with mean of zero and variance of one. Following Yue et al. 2002, a $P$ value for the test is determined using the standard normal cumulative distribution function

$$P = \int_{\infty}^{Z} e^{-t^2/2} \mathrm{d}t.$$

For significance level $\alpha$, if $Z_{a/2} > |P|$, then we reject $H_0$ of no trend.

### Mann-Kendall trend-free pre-whitening (MK-TFPW)

The Mann Kendall trend-free pre-whitening procedure as developed by Yue et al. 2002 is composed of four steps:

1. *Removal of trend* - The slope of trend $b$ is estimated using the Theil-Sen estimator (Theil 1950a-c; Sen 1968) and removed from sample data if different from zero, where $b$ is given by

$$b = \text{Median}\left(\frac{X_j - X_i}{j - l}\right) \forall l < j.$$

Trend $b$ is removed from the series by

$$X_t^{'} = X_t - bt,$$

where $X_t$ is the original series at time step $t$.

2. *Trend-free pre-whitening* - A pre-whitening step is applied to the detrended series to remove the $AR(1)$ component. First, the lag-1 autocorrelation coefficient $\rho_1$ is computed using

$$\rho_k = \frac{\frac{1}{n-k}\sum_{t=1}^{n-k}[X_t - E(X_t)][X_{t+k} - E(X_t)]}{\frac{1}{n}\sum_{t=1}^{n}[X_t - E(X_t)]^2},$$

where $E(X_t)$ is the mean of the series and $\rho_k$ is the lag-$k$ autocorrelation coefficient. Serial correlation is then removed from the detrended series $X_t^{'}$ by

$$Y_t^{'} = X_t^{'} - \rho_1 X_t^{'}.$$

3. *Blending trend and residual series* - Trend $b$ is added to the independent residual series $Y_t^{'}$ by

$$Y_t = Y_t^{'} + bt.$$

4. *MK test* - Trend is assessed through the application of the Mann Kendall test as discussed above.

## Results

### Assessing power of trend detection tests under varying levels of autocorrelation, trend strength, and time series lengths.

Simulation results show that no test for trend exceeded in all scenarios of simulated trend strength, time series length, and autocorrelation strength. As has been documented elsewhere (Yue and Wang 2002; Yue et al. 2002), we show in Figure 1 that time series length has a large effect on the power of each test. Under no autocorrelation, tests for trend are not effective at detecting trends in series with N < 30. When N = 10 with no autocorrelation and strong trend ($\beta$ = 0.8), no test detected trend in >50% of series. The rate of failing to reject the null hypothesis in the presence of trend decreased to < 0.1 when N = 20 with no autocorrelation and strong trend. The effect of increased power with increasing series length and no autocorrelation diminished with reductions in trend strength across all tests. When $\beta$ was greater than 0.04 under no autocorrelation and $N \geq 20$, the GLS test showed the highest rejection rate compared to other tests. Although slightly inflated when N=10, all tests returned rejection rates near the nominal significance level of 0.05 under the no trend and no autocorrelation scenarios.

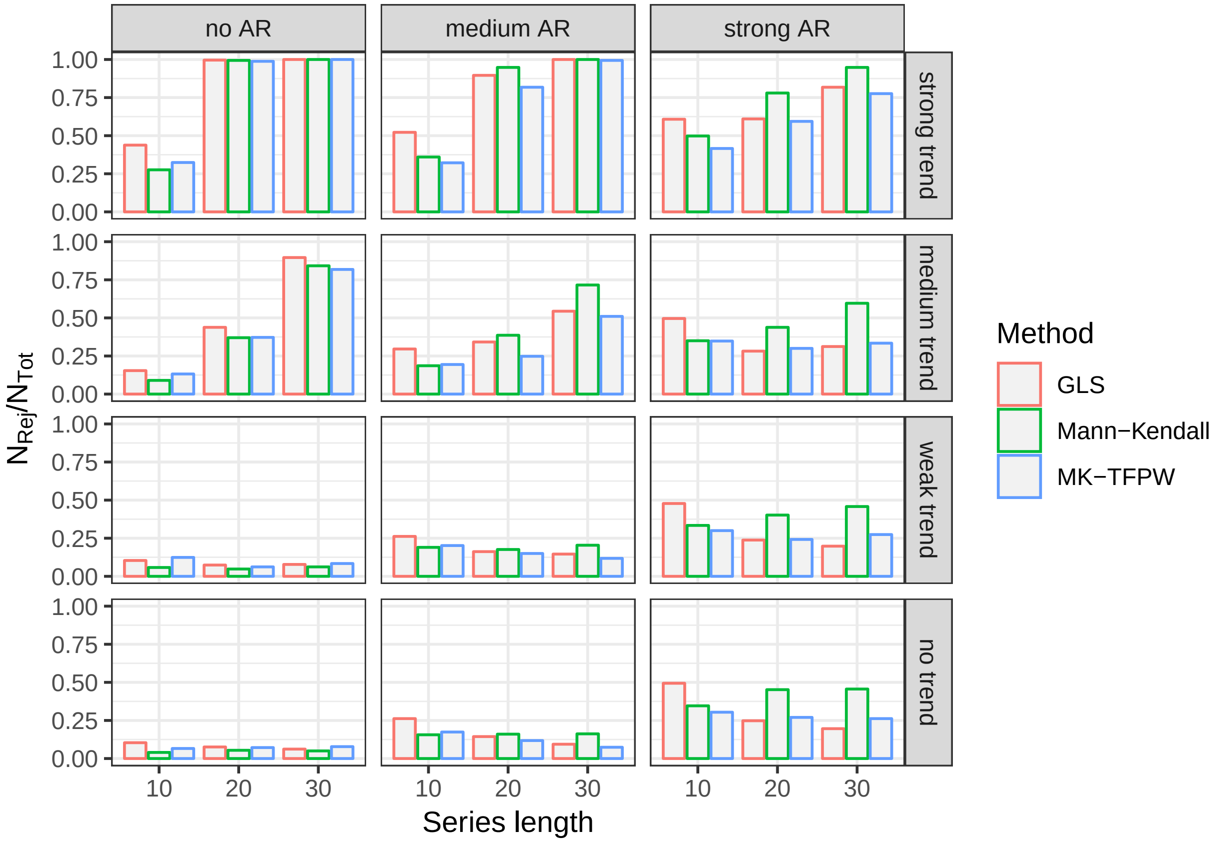


Figure 1: Barplots showing the ratio of number of rejections ($p$<0.05) to number of total simulations. Subplots are representative of different autocorrelation and trend scenarios, with time series length increasing along the x axis. Colored bars show results from different tests for trend.

Autocorrelation is known the reduce the power of the MK test by increasing the variance of the $S$ statistic (Yue and Wang 2002). Identifying this problem led to the development of the MK-PW and other stepwise approaches that sought to address issues introduced by the MK-test when assumptions of independence are violated (Wang et al. 2000, Yue et al. 2002). Our work agrees with these authors and others (von Storch 1995) showing that under no simulated trend, introducing autocorrelation leads to inflated rejection rates in the MK test. Figure 1 shows that under no trend and strong autocorrelation ($\rho = 0.433$ and $\rho = 0.8$), the rejection rate of the Mann Kendall test increases with series length. All other tests showed decreases in rejection rates under both medium and strong autocorrelation scenarios with increasing series length.

The GLS-B test was by far the strongest test under the no trend and strong autocorrelation scenario. When N = 30, the rejection rate of the GLS-B test was 0.046. At N = 20, the rejection rate for this test was slightly inflated to 0.078. Under all autocorrelation and trend strength parameters when N=30, the GLS-B test incorrectly rejected the null in only 5.5% of cases; correctly accepting the null in 94.8% of simulations (Figure 2). The next best test was the MK-PW approach. The rejection rate for this test when N = 30 was close to five times the rate of the GLS-B under the same scenario at 0.222. The GLS test showed the worst performance of tests accounting for $AR(1)$ error structure, although rejection rates followed the trend of decreasing with series length. When N = 30, the GLS test rejection rate was 0.262. Patterns in rejection rate under weak trend ($\beta = 0.004$) and strong autocorrelation were similar to those seen in the no trend and strong autocorrelation scenario (Figure **??**).

Under strong autocorrelation and strong trend, there were increasing rejection rates for all tests with increasing series length, except for the GLS, which had similar rejection rates in both N = 10 and N = 30 simulations. When $\beta \leq 0.8$, the pattern of increasing rejection rate with series length reverses, and for all tests except the Mann Kendall, rejection rates tend to decrease with series length. This result shows that tests developed to account for autocorrelation are not effective in detecting trend when autocorrelation is high unless the sample size is large and trend is strong. However, even with strong trend, power of these tests is limited by the presence of strong autocorrelation. As the GLS-B test was most effective in reducing the prevalence of false positives in the case of no trend and strong autocorrelation, it is the weakest in detecting a strong trend in the presence of high autocorrelation. When N = 30 with strong autocorrelation and strong trend, the GLS-B rejected the null hypothesis in less than half of the simulations ($N_{rej}/N_{Tot} = 0.440$).

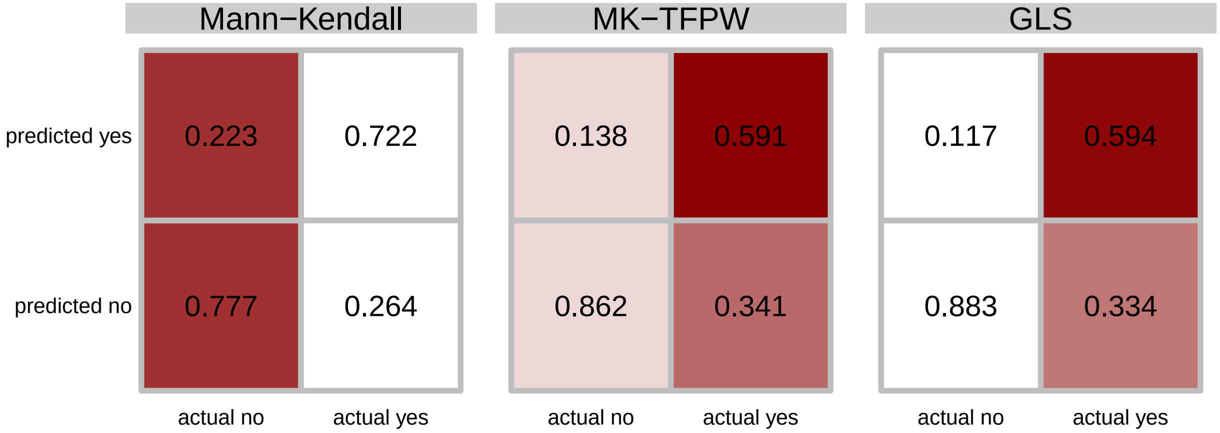
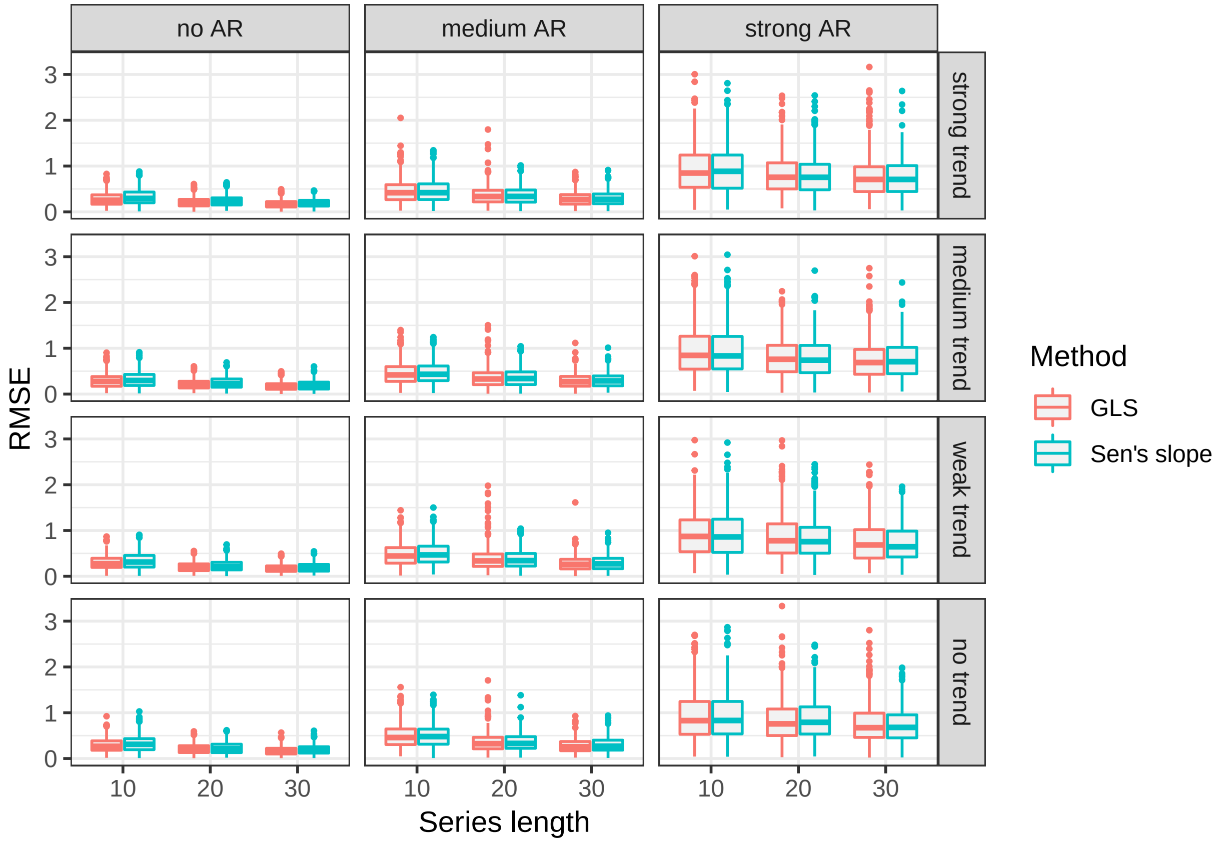


Figure 2: Confusion matrices showing aggregate results from testing for trend across all combinations autocorrelation and trend strength when N=30. Colors represent the performance of individual cells across tests, where cells shaded in red indicate a poorer outcome. For example, when N=30, the GLS-B test falsely predicted a trend when there was none in 5.5% of cases (white), whereas this was true in more than 23% of Mann-Kendall simulations (red).



## Discussion

Ecosystem reporting is vital to the development of Integrated Ecosystem Assessments (IEA), which lay out the framework for moving toward Ecosystem-Based Fishery Management (EBFM)(Levin et al. 2009). The key analytical foundations to all IEA products revolve around the concept of indicator change; whether in the short or long-term, although oftentimes managers are most interested in short-term, abrupt changes to indicator status (Wagner et al. 2013). Here we addressed the shortcomings of assigning significant trends to indicator time series given the common problems of small sample size and autocorrelation. Our results show that commonly used statistical methods used to detect trend in the presence of autocorrelated residuals fail when assessing data with small sample size.

In the Northeast US, most indicators considered in the ecosystem assessment process are annual data typically ranging between 10-60 years in length. In the context of hydrological literature, the upper limit of time series lengths seen in our indicator data sets would be considered short (Bayazit 2015), and although it may be tempting to say otherwise, our work here highlights the dangers of assigning trends to short time series even while attempting to address problems of autocorrelation. The influence of autocorrelation in short series inevitably increases Type II error rates, or the failure to identify trend when it exists. This was especially true under scenarios of strong autocorrelation, which effectively masked the detection of trend unless trend was strong and N was large. An increase in Type I error, or the false rejection of the null hypothesis, was also seen, although results suggest that the GLS bootstrap approach was successful in reducing Type I error to acceptable levels under strong autocorrelation (0.05). All other tests showed high Type I error under strong autocorrelation.

As discussed by Bayazit (2015), the societal cost of Type II error (i.e. under-preparedness) may exceed that of Type I error (over-preparedness).*linking statement* Implementing the GLS-B approach as shown in this work would greatly reduce the likelihood of practicioners committing Type I error in assigning significance to trends that do not exist. However, simply because a signficant trend does not exist does not necessarily mean that nothing of biological importance is occurring. A surge in autocorrelated observations may be indicative of a "status" change occurring in the near-term; knowledge of which provides benefits to managers. Given the complexities and shortcomings of assigning significant trend, a way forward could be to move away from assigning significance to indicator time series altogether. Nicholls (2001) suggests that confidence intervals for trend effect size would suffice as alternatives to assigning signficance. In this case, if the confidence interval were to contain $\beta = 0$, then the null hypothesis of no trend would be accepted.

Our work also focused on the capacity of tests to detect weak, but empirically not uncommon, trend strengths in simulated time series. We found that under the limitations imposed by series lengths $\leq 30$, no test was effective in detecting weak trends. This result supports the work of others [e.g. Wagner2013] suggesting that small changes in trend (<1% change in our case) are difficult to detect in time series regardless of autocorrelation strength for small samples sizes. The scenario of weak trend may also benefit from moving away from the binary outcomes of a significant or non-significant result, and away from the frequentist alternative outlined by Nicholls (2001).

Wagner et al. (2013) suggest the use of Bayesian inference frameworks, which provide a more flexible approach to trend analysis than the binary outcomes of hypothesis testing. Specifically, their work cites Dynamic Linear Models (DLM) as a possible way forward for indicators with small samples size. A DLM approach allows for model coefficients (e.g. slope) to change with time while providing probabilities of rate changes. This approach introduces greater complexity into the common "up or down" model subscribed to by current ecosystem status reports. However, given the results of this study, we suggest that the current pathway for indicator trend analyses is likely insufficient for small samples.

The simulations revealed that no statistical approach excelled under all conditions of series length, trend, and autocorrelation strength. However, results indicate that relying on the GLS bootstrap method may prevent the incidence of Type I error above the 0.05 level regardless of autocorrelation strength, an outcome not captured by any other test for trend. Deriving trends from oftentimes disparate ecosystem indicators is challenging in part due to the goal of applying a single statistical approach to time series with a wide range of series lengths and error structures. The complexity of the chosen method must be balanced with its applicability to a wide range of indicators, and also with the interpretability of its results.

## References

Bayazit, Mehmetcik. 2015. "Nonstationarity of Hydrological Records and Recent Trends in Trend Analysis: A State-of-the-art Review." *Environmental Processes* 2 (3): 527–42. doi:10.1007/s40710-015-0081-7.

Blanchard, J. L., M. Coll, V. M. Trenkel, R. Vergnon, D. Yemane, D. Jouffre, J. S. Link, and Y. J. Shin. 2010. "Trend analysis of indicators: a comparison of recent changes in the status of marine ecosystems around the world." *ICES Journal of Marine Science* 67 (4): 732–44. doi:10.1093/icesjms/fsp282.

Bundy, A, C Gomez, and AM Cook. 2017. "Guidance framework for the selection and evaluation of ecological indicators." Dartmouth, Nova Scotia: Fisheries; Oceans Canada. https://www.researchgate.net/profile/Alida{_}Bundy/pu
framework-for-the-selection-and-evaluation-of-ecological-indicators.

Butchart, Stuart H.M., Matt Walpole, Ben Collen, Arco Van Strien, Jörn P.W. Scharlemann, Rosamunde E.A. Almond, Jonathan E.M. Baillie, et al. 2010. "Global biodiversity: Indicators of recent declines." *Science* 328 (5982): 1164–8. doi:10.1126/science.1187512.

Canales, T. Mariella, Richard Law, Rodrigo Wiff, and Julia L. Blanchard. 2015. "Changes in the size-structure of a multispecies pelagic fishery off Northern Chile." *Fisheries Research* 161: 261–68. doi:10.1016/j.fishres.2014.08.006.

Garcia, S.M, A. Zerbi, C. Aliaume, T. Do Chi, and G. Lasserre. 2003. "The ecosystem approach to fisheries. Issues, terminology, principles, institutional foundations, implementation and outlook." *FAO Fisheries Technical Paper 443*, 71. doi:10.1079/9781845934149.0000.

Garfield, Toby D, and Chris Harvey. 2016. "California Current Integrated Ecosystem Assessment (CCIEA) State of the California Current Report, 2016." *Pacific Fishery Management Council*, no. March: 1–20.

ICES. 2013. "ICES Strategic Plan 2014-2018." International Council for the Exploration of the Sea. doi:ISBN: 978-87-7482-146-5.

Kulkarni, A, and H Von Storch. 1995. "Monte Carlo experiments on the effect of serial correlation on the Mann-Kendall test of trend." *Meteorologische Zeitschrift* 4 (JANUARY): 82–85. http://cat.inist.fr/?aModele=afficheN{\&}cpsidt=3505933.

Levin, Phillip S., Michael J. Fogarty, Steven A. Murawski, and David Fluharty. 2009. "Integrated ecosystem assessments: Developing the scientific basis for ecosystem-based management of the ocean."

doi:10.1371/journal.pbio.1000014.

Mackas, D.L., Richard E. Thomson, and Moira Galbraith. 2001. “Changes in the zooplankton community of the British Columbia continental margin, 1985-1999, and their covariation with oceanographic conditions.” *Canadian Journal of Fisheries and Aquatic Sciences* 58 (4): 685–702. doi:10.1139/cjfas-58-4-685.

NEFSC. 2018a. “State of the Ecosystem - Gulf of Maine and Georges Bank.” Woods Hole, MA: Northeast Fisheries Science Center.

———. 2018b. “State of the Ecosystem - Mid-Atlantic Bight.” Woods Hole, MA: Northeast Fisheries Science Center.

Nicholls, N. 2001. “The insignificance of significance testing.” *Bulletin of the American Meteorological Society* 81 (5): 981–86.

Nicholson, Mike D., and Simon Jennings. 2004. “Testing candidate indicators to support ecosystem-based management: The power of monitoring surveys to detect temporal trends in fish community metrics.” doi:10.1016/j.icesjms.2003.09.004.

NOAA. 2006. “Evolving an ecosystem approach to science and management through NOAA and its partners.” National Ocean; Atmospheric Administration. https://sab.noaa.gov/sites/SAB/Reports/EETT/eERRT - Final Report to NOAA Oct 06.pdf.

Perry, RI, P Livingston, and E Fulton. 2010. “Ecosystem Indicators. In: Ecosystem-based Management Science and its Application to the North Pacific.” In *PICES Scientific Report No. 37*, 184.

Secretariat of the Convention on Biological Diversity. 2004. *The Ecosystem Approach.* Vol. 2. 1. doi:10.1007/BF00043328.

Shannon, Lynne J., Marta Coll, Dawit Yemane, Didier Jouffre, Sergio Neira, Arnaud Bertrand, Erich Diaz, and Yunne Jai Shin. 2010. “Comparing data-based indicators across upwelling and comparable systems for communicating ecosystem states and trends.” *ICES Journal of Marine Science* 67 (4): 807–32. doi:10.1093/icesjms/fsp270.

Shin, Yunne Jai, and Lynne J. Shannon. 2010. “Using indicators for evaluating, comparing, and communicating the ecological status of exploited marine ecosystems. 1. the indiSeas project.” *ICES Journal of Marine Science* 67 (4): 686–91. doi:10.1093/icesjms/fsp273.

Storch, Hans von. 1999. “Misuses of Statistical Analysis in Climate Research.” In *Analysis of Climate Variability*, 11–26. doi:10.1007/978-3-662-03744-7_2.

Wagner, Tyler, Brian J. Irwin, James R. Bence, and Daniel B. Hayes. 2013. “Detecting Temporal Trends in Freshwater Fisheries Surveys: Statistical Power and the Important Linkages between Management Questions and Monitoring Objectives.” *Fisheries* 38 (7): 309–19. doi:10.1080/03632415.2013.799466.

Wang, X. L., and V. R. Swail. 2001. “Changes of extreme Wave Heights in northern Hemisphere Oceans and related atmospheric circulation regimes." *Journal of Climate* 14 (10): 2204–21. doi:10.1175/1520-0442(2001)014<2204:COEWHI>2.0.CO;2.

Yue, Sheng, and Chun Yuan Wang. 2002. "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test." *Water Resources Research* 38 (6): 4–1–4–7. doi:10.1029/2001WR000861.

Yue, Sheng, Paul Pilon, Bob Phinney, and George Cavadias. 2002. "The influence of autocorrelation on the ability to detect trend in hydrological series." *Hydrological Processes* 16 (9): 1807–29. doi:10.1002/hyp.1095.

Zhang, Xuebin, Lucie A. Vincent, W.D. Hogg, and Ain Niitsoo. 2000. "Temperature and precipitation trends in Canada during the 20th century." *Atmosphere-Ocean* 38 (3): 395–429. doi:10.1080/07055900.2000.9649654.