A simulation study of trend detection methods for Integrated Ecosystem Assessment

Sean Hardison, Charles Perretti, Andy Beet, Geret DePiper

June 29, 2018

Abstract

The identification of trends in ecosystem indicators has become a core component of ecosystem approaches to resource management, although oftentimes assumptions of statistical models are not properly accounted for in the reporting process. To explore the limitations of trend analysis of short times series, we applied three common methods of trend detection, including a generalized least squares model selection approach, the Mann-Kendall test, and Mann-Kendall test with trend-free pre-whitening to simulated time series of varying trend and autocorrelation strengths. Our results suggest that the ability to detect trends in time series is hampered by the influence of autocorrelated residuals in short series lengths. While it is known that tests designed to account for autocorrelation will approach nominal rejection rates as series lengths increase, the results of this study indicate biased rejection rates in the presence of even weak autocorrelation for series lengths often encountered in indicators developed for ecosystem-level reporting (N = 10, 20, 30). This work has broad implications for ecosystem-level reporting, where indicator time series are often limited in length, maintain a variety of error structures, and typically are assessed using a single statistical method applied uniformly across all time series. If a hypothesis testing approach for indicator trend analysis is to be implemented, we suggest first characterizing candidate series based on suitability (e.g. based on variance, autocorrelation, and series length) rather than a uniform application of tests for trend. A parametric approach to trend assessment could then be used to provide estimates of uncertainty and trend strengths from probability distributions.

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 2017a; NEFSC 2017b; 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 (Bence 1995; Nicholson and Jennings 2004; Wagner et al. 2013). The hydrological, climatological, and statistical 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; Woodward, Bottone, and Gray 1997; Hamed and Rao 1998; Storch 1999; Nicholls 2001; Roy, Falk, and Fuller 2004, 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 across all series length. 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

Data

Parameters used in simulations were chosen based on preliminary analyses characterizing the distribution of trend and autocorrelation strengths across 124 time series that were candidates for inclusion in the 2017 State of the Ecosystem (SOE) reports (NEFSC 2017a; NEFSC 2017b) (Fig. 1). Trends in these candidate time series were characterized by linear regression, with the mean and upper 95% confidence interval for slopes chosen for representation in simulations. The ρ components of SOE time series were estimated by maximum likelihood estimation (MLE), and the distribution mean was chosen as our "medium autocorrelation" parameter for simulated series. To reasonably parameterize simulation innovation variance, we fit all residual series with an AR(1) model estimating innovation variance using MLE, and then found the mean of the resulting distribution of variances.

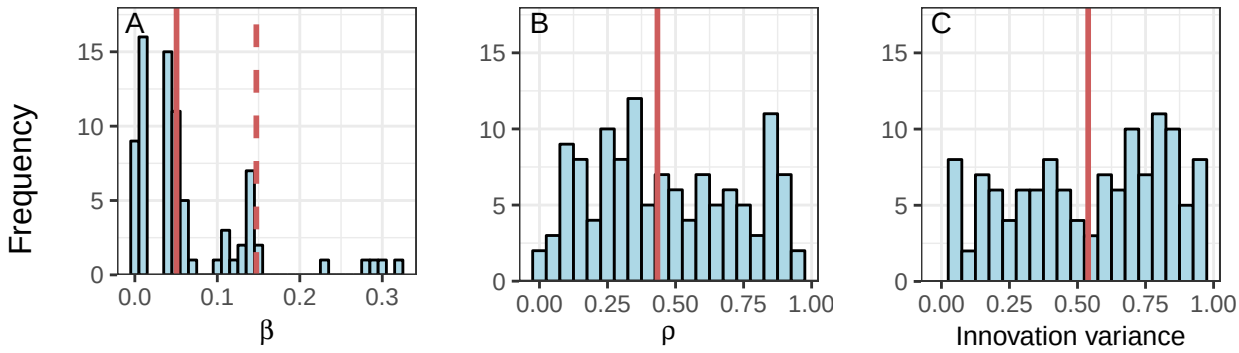

Figure 1: Frequency of estimated slopes (A), autocorrelation strengths (B), and innovation variances (C) in time series considered for inclusion in the 2017 State of the Ecosystem report. The solid red lines (A-C) represent distribution means, and the dashed red line (A) shows the upper 95% confidence interval for estimated trend slopes.

Simulations

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

linear models:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1 X_t + \mu_t \\ \mu_t &= \rho\mu_{t-1} + \omega_t \\ \omega_t &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

where Y_t is the simulated series at time t, α_1 is the slope component, and μ_t is the AR1 error process; the strength of which is given by ρ, with the error component ω_t assumed to be derived from Gaussian white noise. Through the preliminary analysis detailed above, the levels of α_1 were 0.026, 0.051, and 0.147, which we combined with three levels of ρ: 0, 0.43, and 0.8. For each trend strength, autocorrelation strength, and time series length, 1000 simulations were performed. To test the null hypothesis of no trend in simulated time series, we used a generalized least squares (GLS) model selection process, Mann Kendall test, and Mann Kendall test with trend-free pre-whitening.

We focus our analyses on rejection rates of the null hypothesis of no trend, as this methodology is a common framework for assessing the flexibility of trend models to deviations from assumptions [e.g. Yue and Wang (2002); Yue2002b]. Further, null hypothesis testing is often applied in ecosystem indicator reporting for assessing trend [e.g. NEFSC (2017a); NEFSC2018b]. We chose to extend this analysis of rejection rates for the scenario of no trend and strong autocorrelation to larger sample size (N = 50-650) to highlight the shortcoming of small sample sizes when strong autocorrelation is present. Our final analysis compared the efficacy of the non-parametric Sen's slope to the GLS estimator for assigning linear trend to data where trend was found to be significant ($p < 0.05$).

Generalized least squares

A GLS model selection procedure was implemented to test for trend in simulated series. Two first order linear and two quadratic GLS models were fit to each simulated time series and best models were chosen using AIC corrected for small sample size (AICc). Specifically, the models were 1) linear trend with uncorrelated residuals, 2) linear trend with correlated residuals, 3) quadratic trend with uncorrelated residuals, and 4) quadratic trend with correlated residuals. Component GLS models were derived from

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2 + \mu_t \\ \mu_t &= \rho\mu_{t-1} + \omega_t \\ \omega_t &\sim N(0, \sigma^2) \end{aligned} \tag{2}$$

The above model follows the same notation as our simulated series. Setting $\alpha_2 = 0$ yielded linear trend models, and $\rho = 0$ gave models with uncorrelated residuals.

Mann Kendall test

Further tests for trend in simulated time series were performed using the Mann-Kendall test (MK) (Mann 1945; Kendall 1955) and the more robust Mann-Kendall test with trend-free pre-whitening (MK-TFPW)(Yue et al. 2002). The MK test is a nonparametric test for trend that assumes sample data are independent and identically distributed. Serial correlation within sample data has been found to lead to inflated rejection rates of the null hypothesis of no trend if no correction steps are applied to the MK test (Kulkarni and Storch 1995). Residual pre-whitening is a common correction to address autocorrelation within MK tests, although pre-whitening is known to reduce the magnitude of existing trend [Yue2002a]. The MK 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:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(Y_j - Y_i), \tag{3}$$

where Y is the response vector, n is the length of the series, and

$$\text{sgn}(x) = \left\{ \begin{array}{ll} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{array} \right\}. \tag{4}$$

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

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

and the distribution of S when $n \geq 8$ is approximately normal and symmetric about a mean of 0 and variance, V(S). The standardized test statistic,

$$\mathrm{Z} = \left\{ \begin{array}{ll} \frac{S-1}{\sqrt{V(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{V(S)}} & S < 0 \end{array} \right\}, \tag{6}$$

is normally distributed with mean of zero and variance of one (Wang and Swail 2001). The null hypothesis of no trend is rejected at significance level α if the probability $1 - \Phi(|Z|) < \alpha$, where $\Phi(x)$ is the standard normal cumulative distribution funciton.

Mann-Kendall trend-free pre-whitening

The Mann Kendall trend-free pre-whitening procedure as developed by Yue and Wang (2002) is composed of four steps:

1. *Removal of trend* - The slope of trend b is estimated using the Theil-Sen estimator [@theill1992; @sen1968] and removed from sample data if different from zero, where b is given by

$$b = \text{Median}\left(\frac{y_j - y_i}{j - l}\right) \forall l < j. \tag{7}$$

   Trend b is removed from the series by

$$y_t^{'} = y_t - bt, \tag{8}$$

   where y_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 ρ_1 is computed using

$$\rho_k = \frac{\frac{1}{n-k}\sum_{t=1}^{n-k}[y_t - E(y_t)][y_{t+k} - E(y_t)]}{\frac{1}{n}\sum_{t=1}^{n}[y_t - E(y_t)]^2}, \tag{9}$$

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

$$Y_t^{'} = y_t^{'} - \rho_1 y_t^{'}. \tag{10}$$

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

$$Y_t = Y_t^{'} + bt. \tag{11}$$

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

Results

Throughout this study we adopt an alpha value of 0.05 to assess significance of statistical results. Overall, no method performed consistently well 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 find time series length has a large effect on the power of each test (Figure 2), and performance was generally best across autocorrelation and trend scenarios when N = 30. With no autocorrelation and trend present, trends were only detected with $> 90\%$ accuracy when trend was strong ($\alpha_1 = 0.147$). Even with a strong trend and no autocorrelation, no test detected a trend in greater than 50% of the series when N = 10. Again under no autocorrelation, the increased power associated with increasing series length diminished with reductions in trend strength across all tests. Under no autocorrelation, the GLS test showed the highest rejection rates compared to other tests, although this effect was minimal (mean difference of rejection rates between GLS and MK-TFPW was <3%). All tests returned rejection rates near the nominal significance level of 0.05 under the no trend and no autocorrelation scenarios, with the largest departures occurring when N = 10 ($\text{MK-TFPW}_{\text{sig}} = 0.096$, $\text{MK}_{\text{sig}} = 0.035$, $\text{GLS}_{\text{sig}} = 0.065$).

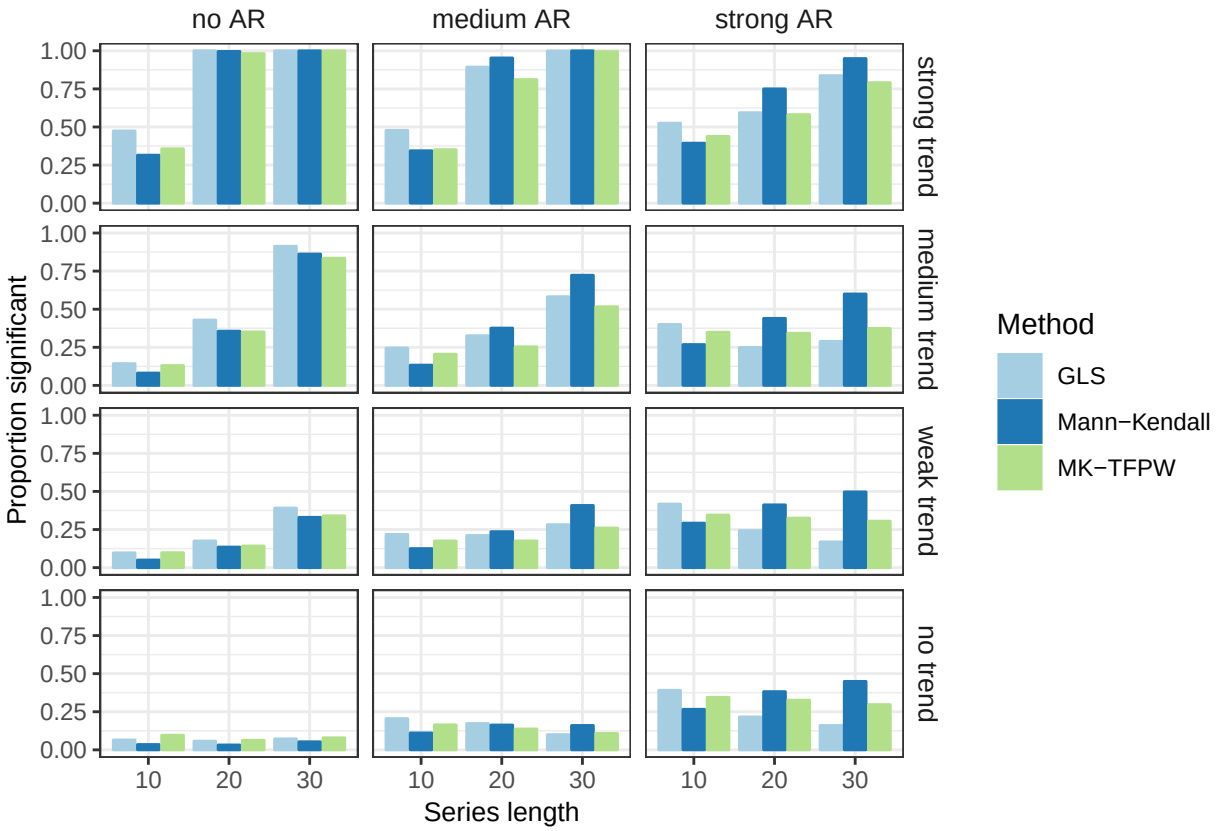

Figure 2: Barplots showing the propotion of significant trends (P<0.05) to number of total simulations. Subplots are representative of different autocorrelation ($\rho = 0, .43, .8$) and trend scenarios ($\alpha_1 = 0.026, .051$), with time series length increasing along the x axis. Colored bars show 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 (Storch 1999; Yue and Wang 2002), and our work also shows that under no simulated trend, introducing autocorrelation will lead to inflated rejection rates in the MK test. The bottom row of Figure **??** shows that under no trend and medium to 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.

The GLS procedure performed the best under the no trend and strong autocorrelation scenario: when N = 30, the rejection rate for the GLS was 0.16, 46% and 64% lower than the MK-TFPW and MK tests respectively. The performance of the GLS test was also more strongly affected by sample size than the MK-TFPW test. When there was strong autocorrelation and no trend, rejection rates of the MK-TFPW test decreased only 13.9% between N = 10 and N = 30. Under the same conditions and GLS approach, rejection rates decreased by 59%. However, the GLS approach also performed the worst under no trend and strong autocorrelation when N = 10. This shows that while there was improvement between both tests as series

lengths increased, neither test was effective in accounting for biases of autocorrelation when $N \leq 30$.

Extending this no trend and strong autocorrelation scenario out to longer series lengths shows that the GLS test reaches nominal rejection rates of 0.05 only when $N \geq 500$ (Figure 3). The MK-TFPW approach performed poorly in this analysis, and did not converge to nominal rejection rates for $N < 500$, although this work did not seek to identify a precise value of N where the MK-TFPW approach reached nominal levels. As expected, the MK test performed poorly in this scenario, and saw no reduction in rejection rates as N increased.

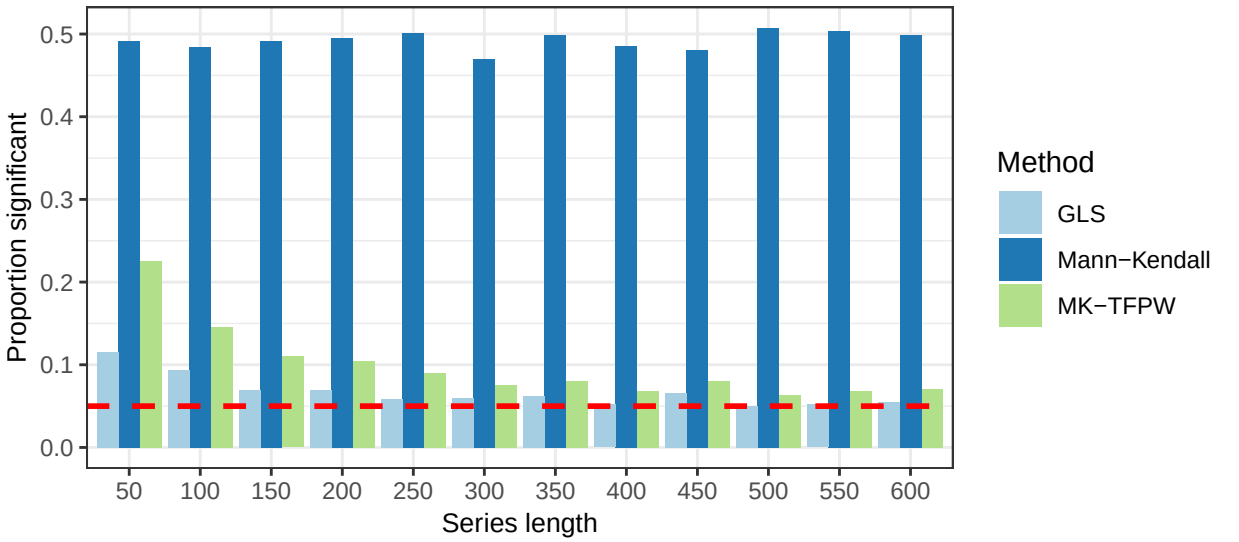

Figure 3: Barplot showing the ratio of number of rejections (P<0.05) to number of total simulations, when simulations were created under the parameters of no trend ($\alpha_1 = 0$), strong autocorrelation ($\rho = 0.8$), and series lengths between $N = 50$ to $N = 500$. The dashed red line shows the nominal rejection rate of 0.05.

Under strong autocorrelation ($\rho = 0.8$) and strong trend ($\alpha_1 = 0.147$), the relationship between time series length and rejection rate was positive, highlighting the importance of the trend signal strength on test results. Under these parameters, the MK-TFPW test performed slightly better than the GLS approach, although both tests were only able to detect trend in >50% of simulations when trend was strong and N = 20 or N = 30. When $\alpha_1 < 0.0147$ and autocorrelation was strong, neither the GLS nor MK-TFPW tests were able to detect trend in >50% of simulations regardless of series length. Interestingly, as series lengths increased when trend was weak (i.e. $\alpha_1 = 0.026$) and $\rho = 0.8$, rejection rates tended to decrease for GLS and MK-TFPW tests. The relative success of each test when N = 30 can be seen in Figure 4, which shows that the GLS approach was most effective in avoiding false positives, but performed similarly to the MK-TFPW test in terms of false negatives.

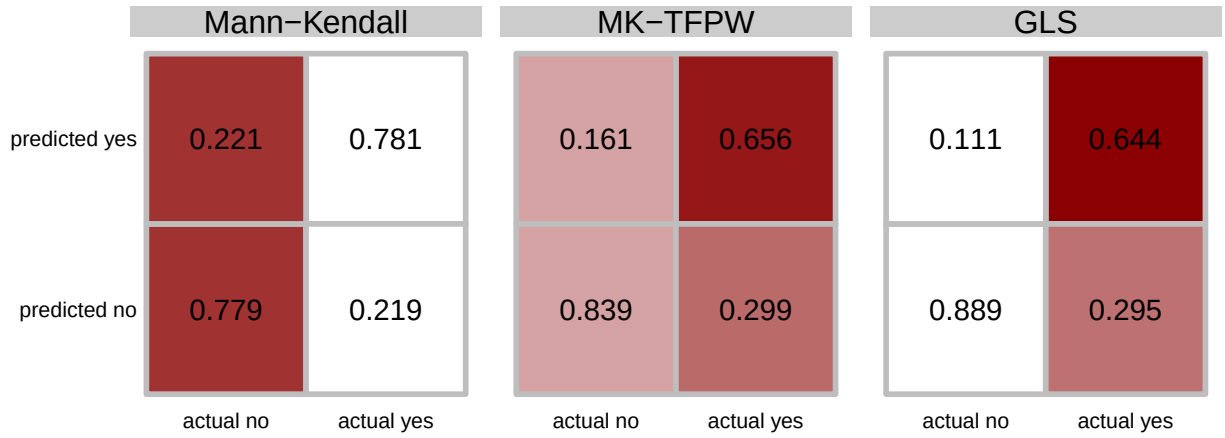

Figure 4: Confusion matrices showing aggregate results from testing for trend across all combinations autocorrelation and trend strength when N=30. Colors represent the performance of individual cells across tests, where cells shaded in red indicate a poorer outcome. For example, when N=30, the GLS procedure falsely predicted a trend when there was none in 11.1% of cases (white), whereas this was true in 22.1% of Mann-Kendall simulations (red).

We next assessed the ability of each statistical approach to estimate the true trend (Figures 5 and **??**). In the nonparametric case, we used Sen's slope (as derived in Equation 8), which is a common statistic estimated alongside the MK and MK-TFPW significance tests. Sen's slope and the GLS estimator perform similarly across all scenarios. For both methods, the spread of estimated trends increased with autocorrelation strength, although this effect was mediated by increasing series length (e.g. Fig **??**). Further, trends falsely assigned in the "no trend" scenarios tended to have the largest spread. As shown by the black median lines in Figure **??**, both GLS and Sen's slope methods consistently overestimated trend slope when it existed, although both performed well under the strong autocorrelation scenario when trend was strong and N = 30.

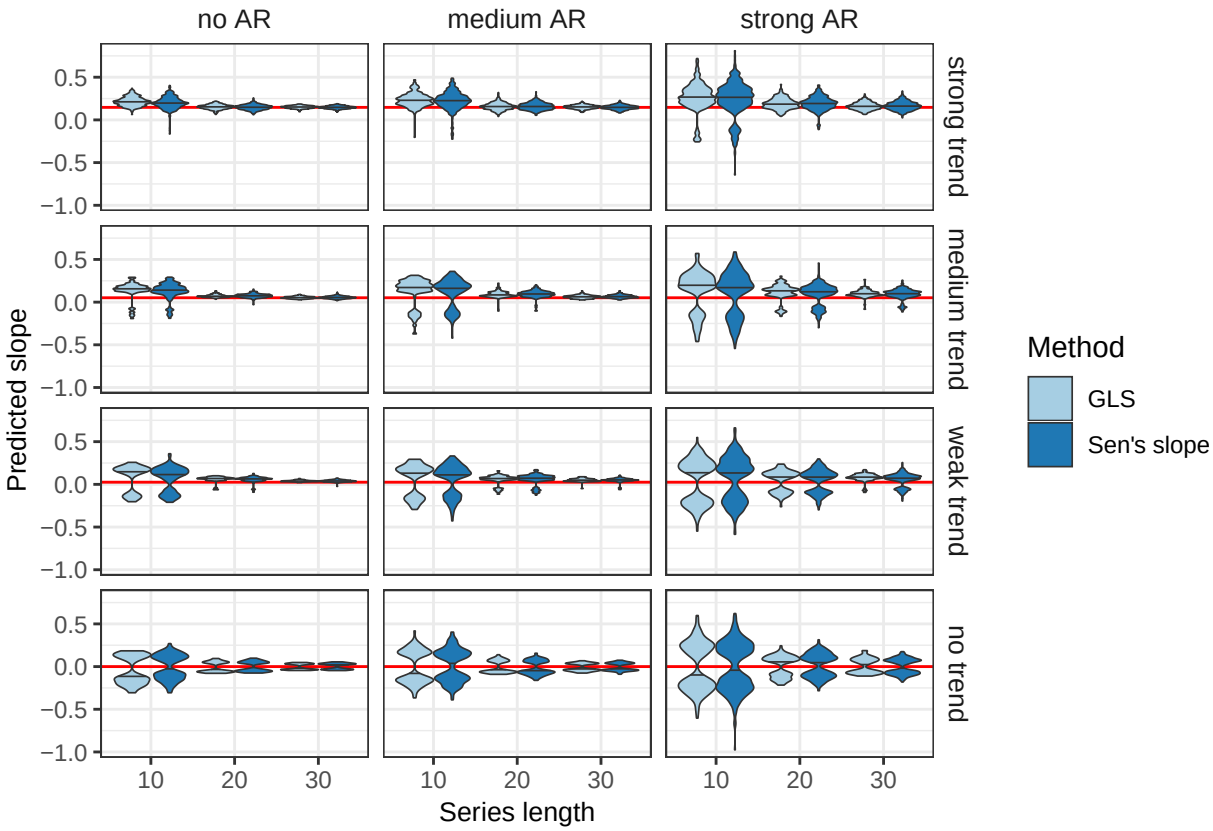

Figure 5: Violin plots showing probability densities of estimated trends from GLS and Sen's slope procedures under varying autocorrelation scenarios ($\rho = 0, 0.43, 0.8$) and simulation lengths (N = 10, 20, 30). Black lines represent the median slope estimate, and red lines the true slope. For this exercise, the GLS model selection procedure was constrained to fit only linear models of trend.

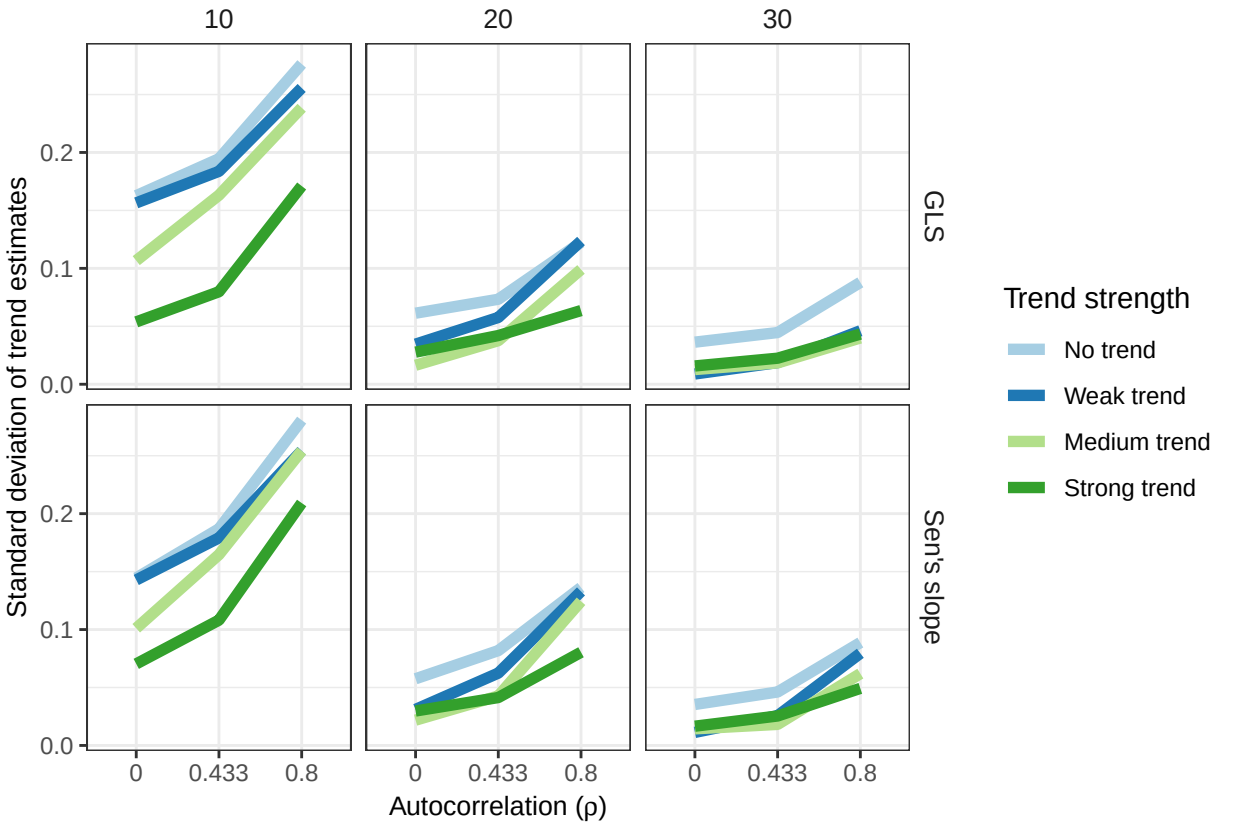

Figure 6: Standard deviations of trend estimates derived from Sen's slope and GLS methods across autocorrelation strengths ($\rho = 0, 0.433, 0.8$), trend strengths ($\alpha_1 = 0, 0.026, 0.052, 0.147$), and trend methods.

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; with managers 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 statistical methods commonly used to detect trend in the presence of autocorrelated residuals may be misleading when applied to short time series.

In the Northeast US, 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). This study 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 (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. Possibly even more problematic, an increase in Type I error, or the false rejection of the null hypothesis, occurred as the strength of autocorrelation increased. A departure from nominal rejection rates was seen from all tests with even moderate amounts of autocorrelation.

Our work focused on the capacity of tests to detect weak trend strengths in simulated time series. We found that under the limitations imposed by series lengths ≤ 30, no test was effective in detecting weak trends, even without autocorrelation (e.g. Fig. 2 no AR and weak trend scenario). Decreasing rejection rates with increasing series lengths for GLS and MK-TFPW tests under weak trend and strong autocorrelation further illustrate the "masking" effect of autocorrelation on trend detection. This result supports the work of others (e.g. Wagner et al. 2013) that have found small trend changes in short time series difficult to detect regardless of autocorrelation. The role of autocorrelation in making trend can in part can be explained by the variance of the random error used in our simulations ($\sigma^2 = 0.54$). While this may be considered large relative to the our definition of weak trend, it was selected based on the properties of real ecosystem indicators.

As shown in Figure 2, there is no solution in small sample sizes; however, we do not suggest there is no value in assigning trends to time series. Instead, we advise that a "shotgun" approach to assessing trends in many indicator time series without consideration of error structure and series lengths will likely lead to both Type I and Type II error. If the binary approach of hypothesis testing is to be implemented in ecosystem indicator reporting, a more hands-on approach should be implemented to determine indicators that are well-suited for trend analysis (e.g. a long series with low variance and weak autocorrelation). Further, the results of this study showed that the GLS approach to modeling trend ameliorated error rates compared to the Mann-Kendall test with trend-free pre-whitening. The parametric approach also carries the benefits of modeling trend uncertainty and coefficients, given the assumed probability distribution.

A more intuitive and flexible approach to trend assessment would be to simply present more information with each assessed time series. Nicholls (2001) suggested that the arbitrary (i.e. "$p < 0.05$") null hypothesis testing framework be replaced by the presentation of confidence intervals for trend effect size. This approach has the potential to provide more contextual information to managers, but as we show above, is limited by the reality that trends (and therefore confidence intervals for effect size) are often misrepresented when series length is small and autocorrelation exists. Supplementing ecosystem reporting documents with methodological summaries could be useful to highlight these limitations and provide realistic expectations for managers (Wagner et al. 2013).

A different approach to trend assessment departs from null hypothesis testing altogether in favor of a Bayesian framework. Wagner et al. (2013) suggests Dynamic Linear Models (DLMs) for indicators of small sample size. DLMs allow 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, and could therefore provide greater insight to managers. In an example of Bayesian regression, Wade (2000) showed how a series with larger variance but a biologically significant trend would be considered non-significant by a frequentist approach, but properly assessed by Bayesian methods. This framework could be adopted by analysts to answer specific questions that resource managers are interested in addressing; e.g. what is the probability that indicator X declined by Y% between this year and last? While Bayesian methods cannot side-step the reality of small sample sizes, their use provides managers with a probabilistic framework for decision-making that is lacking in the frequentist approach (Wade 2000; Wagner et al. 2013).

Deriving trends from 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 the interpretability of its results. Our work shows that blindly implementing this approach will likely result in assigning spurious trends or missing important patterns. However, programmatic consideration of candidate series for trend analysis would likely ameliorate some instance of error. Implementation of a parametric test for trend (e.g. the GLS procedure in our study) then has the benefit of providing estimates of uncertainty and trend based on a probability distribution. A subtler approach for trend analyses in ecosystem reporting would provide better outcomes for economic, ecological, and social systems in the context of EBFM decision-making.

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.

Bence, James R. 1995. "Analysis of short time series: correcting for autocorrelation." *Ecology* 76 (2). Wiley Online Library: 628–39.

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.

Hamed, Khaled H, and A Ramachandra Rao. 1998. "A modified Mann-Kendall trend test for autocorrelated data." *Journal of Hydrology* 204 (1-4). Elsevier: 182–96.

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

Kendall, Maurice G. 1955. "Rank correlation methods." Hafner Publishing Co.

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.

Mann, Henry B. 1945. "Nonparametric tests against trend." *Econometrica: Journal of the Econometric Society*. JSTOR, 245–59.

NEFSC. 2017a. "State of the Ecosystem - Mid-Atlantic Bight." Woods Hole, MA: Northeast Fisheries Science Center. https://static1.squarespace.com/static/511cdc7fe4b00307a2628ac6/t/58de8227bf629a46b8ab35ad/1490977355678/Tab02_2017-04_State-of-the-Ecosystem-and-EAFM.pdf.

———. 2017b. "State of the Ecosystem - Mid-Atlantic Bight." Woods Hole, MA: Northeast Fisheries

Science Center. http://s3.amazonaws.com/nefmc.org/2_2016-State-of-the-Ecosystem-Report.pdf.

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

———. 2018b. "State of the Ecosystem - Gulf of Maine and Georges Bank." 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.

Roy, Anindya, Barry Falk, and Wayne A Fuller. 2004. "Testing for trend in the presence of autoregressive error." *Journal of the American Statistical Association* 99 (468). Taylor & Francis: 1082–91.

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.

Wade, Paul R. 2000. "Bayesian methods in conservation biology." *Conservation Biology* 14 (5). Wiley Online Library: 1308–16.

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.

Woodward, Wayne A, Steven Bottone, and HL Gray. 1997. "Improved tests for trend in time series data." *Journal of Agricultural, Biological, and Environmental Statistics*. JSTOR, 403–16.

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.