ICES Journal of Marine Science

ICES Journal of Marine Science (2019), 76(7), 2060–2069. doi:10.1093/icesjms/fsz097

Original Article

A simulation study of trend detection methods for integrated ecosystem assessment

Sean Hardison [1,2]*, Charles T. Perretti[1,2], Geret S. DePiper[2], and Andrew Beet[1,2]

[1]Integrated Statistics, Woods Hole, MA 02543-1026, USA
[2]Northeast Fisheries Science Center, Woods Hole, MA 02543-1026, USA

*Corresponding author: tel: +1 508 495 2273; e-mail: sean.hardison@noaa.gov.

Hardison, S., Perretti, C. T., DePiper, G. S., and Beet, A. A simulation study of trend detection methods for integrated ecosystem assessment. – ICES Journal of Marine Science, 76: 2060–2069.

Received 22 October 2018; revised 18 March 2019; accepted 10 April 2019; advance access publication 7 June 2019.

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 are typically assessed using a single statistical method applied uniformly across all time series.

Keywords: ecosystem-based fisheries management, ecosystem indicators, trend analysis

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 (Garcia *et al.*, 2003; Secretariat of the Convention on Biological Diversity, 2004; NOAA, 2006; Levin *et al.*, 2009; Perry *et al.*, 2010; ICES, 2019). 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 (Blanchard *et al.*, 2010; Butchart *et al.*, 2010; Garfield and Harvey, 2016; NEFSC, 2017a, b, 2018a, b; Wiebe *et al.*, 2017).

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 data sets; generating a consistent series length across indicators for comparison purposes (Blanchard *et al.*, 2010; Shannon *et al.*, 2010; Shin and Shannon, 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. <20 data points per indicator (Mackas *et al.*, 2001; Nicholson and Jennings, 2004; Blanchard *et al.*, 2010; Shannon *et al.*, 2010; Shin and Shannon, 2010; Canales *et al.*, 2015; Karnauskas *et al.*, 2017). 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 show 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, 1992; Woodward *et al.*, 1997; Hamed and Rao, 1998; Storch, 1999; Zhang *et al.*, 2000; Nicholls, 2001; Wang and Swail, 2001; Yue and Wang, 2002; Roy *et al.*, 2004; 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, 1992; 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.

Assessments of ecosystem status and trend are important pieces of the Ecosystem Approach, particularly with respect to integrated ecosystem assessments (IEAs) (Zador *et al.*, 2017), and so it should be emphasized that the ability to detect trends has implications for future management outcomes. For example, in the Northeast United States indicators were used in a risk assessment based on their capacity to capture the potential threats to valuable ecosystem components (F/B status, food production, habitat quality, etc.), and trend analysis was a component of the developed risk rankings (Gaichas *et al.*, 2018). Mischaracterized indicator trends can therefore lead to biases in the risks chosen to assess the performance of management strategies against.

In this study, we abstract away from issues surrounding the identification and vetting of appropriate indicators but note that this can be a challenging undertaking for which Bundy *et al.* (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 (e.g. community well-being and stability). Given the known biases introduced by common structural aspects of time series in trend assessment, our goal here is to assess the feasibility of providing rigorous insights to managers in data limited situations. 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 US Mid-Atlantic and New England Fishery Management Councils, respectively (NEFSC, 2017a, b).

Results indicate that correctly identifying trends is problematic using <30 data points, with both Type I and Type II error common. Even under the strongest signal-to-noise ratio (i.e. strong trends and no autocorrelation) tests perform poorly when series lengths are <30. 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 normalized time series that were candidates for inclusion in the 2017 State of the Ecosystem (SOE) reports (NEFSC, 2017a, b) (Figure 1). Trends in these candidate time series were characterized by linear regression, with the mean and upper 95th percentile for the absolute value of slopes chosen for representation in simulations. We chose not to account for autocorrelated error structure when estimating slopes in this analysis, as our goal was simply to generate reasonable values to simulate from.

The first- and second-order autoregressive parameters (ρ_1, ρ_2) of SOE time series were estimated by fitting a second-order autoregressive [AR(2)] model via maximum likelihood estimation (MLE). The mean and 95th percentile of all estimated $\rho_1$values were chosen as our "medium" and "strong" autocorrelation parameters for series simulated with AR(1) error. Analysis of SOE time series to derive the average ρ_2 condition yielded a value close to 0, and so we chose to simulate from $\rho_2 = 0.2$ to better represent the influence of AR(2) error. To reasonably parameterize simulation series residual variance, we fitted all residual series with an AR(1) model estimating variance using MLE, and then calculated the mean, first quartile, and third quartile of the resulting distribution of variances. Data and R code for this work are available at https://github.com/seanhardison1/soe_simulations/.

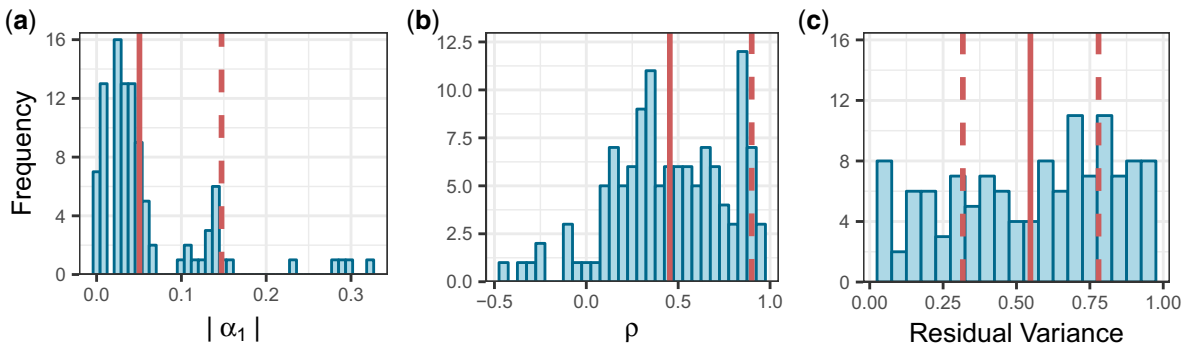

Figure 1. Frequency of estimated slopes (absolute values) (a), autocorrelation strengths (b), and time series residual variances (c) in 124 time series considered for inclusion in the 2017 SOE report. The solid lines (a–c) represent distribution means. The dashed lines in a and b show the 95th percentile for estimated trend slope and AR(1) error strengths, and the dashed lines in c show the lower and upper quartiles for series residual variance.

Simulations

Simulated time series were generated through the addition of AR(1) and AR(2) autoregressive processes to first-order linear models:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1 X_t + \epsilon_t \\ \epsilon_t &= \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \omega_t, \\ \omega_t &\sim N(0, \sigma^2) \end{aligned} \quad (1)$$

where Y_t is the simulated series at time t, α_1 is the slope component, and ϵ_t is the AR(k) process of order k. The error component ω_t was assumed to be derived from Gaussian white noise. Setting ρ_2 to 0 yielded the AR(1) error process. 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_1: 0, 0.43, and 0.9. Autoregressive parameters used to simulate from the AR(2) model were $\rho_1 = 0.43$ and $\rho_2 = 0.2$, which we crossed with all levels of trend. 10,000 simulated time series were assessed for each component of our study.

To test for trend in simulated time series, we used a generalized least squares (GLS) model selection process, Mann–Kendall (MK) test, and Mann–Kendall test with trend-free pre-whitening (MK-TFPW), which we describe in greater detail below. Both similar nested GLS approaches and MK tests have been used rather extensively in ecosystem-level reporting (Blanchard *et al.*, 2010; Butchart *et al.*, 2010; Shannon *et al.*, 2010; Canales *et al.*, 2015; NEFSC, 2017a, b, 2018a, b; Wiebe *et al.*, 2017). The MK-TFPW was included as an obvious extension for autocorrelated time series, given the test attempts to overcome the failed independent error assumption of the original MK specification.

We focused 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 (Yue, Pilon, and Cavadias, 2002; Yue, Pilon, Phinney, *et al.*, 2002; Yue and Wang, 2002). Furthermore, null hypothesis testing is often applied in ecosystem indicator reporting for assessing trend (NEFSC, 2017a, b, 2018a, b). Our first analysis tested for trend in simulations crossed with all levels of AR(1) error, trend strength, and series length. We then extended this analysis for the scenario of no trend and strong autocorrelation to larger sample sizes ($N = 50$–650) to highlight the shortcomings of small sample sizes when autocorrelated residuals are present. To address the role of time series variance in trend detection, we also simulated the fully crossed autocorrelation and trend strength scenario under low and high levels of series variance. Next, we simulated time series with an AR(2) error component at all levels of trend. Our final analysis compared the efficacy of the nonparametric Sen's slope to the GLS estimator for assessing trend effect size (i.e. slope) where trend was found to be significant ($p < 0.05$).

Generalized least squares

GLS models (with or without modified error structures) have in the past been a common approach to testing for trend in ecosystem indicator assessments (Blanchard *et al.*, 2010; Shannon *et al.*, 2010; Karnauskas *et al.* 2017; NEFSC, 2018a, b), and so a GLS model selection procedure was chosen for simulation testing. If simulations were generated with Gaussian or AR(1) error processes, we fit two first-order linear models to each simulated series: one with uncorrelated residuals (i.e. linear regression) and one with correlated residuals [Equation (2)]. The best model fit was then chosen using AIC corrected for small sample size (AICc). When simulating from a model with AR(2) error, we included a third linear model in the selection process with second-order correlated residuals. The above model follows the same notation as our simulated series. Setting ρ_1 and $\rho_2 = 0$ gave models with uncorrelated residuals. The model selection procedure was implemented in R using the packages *nlme* and *AICcmodavg* (Mazerolle, 2017; Pinheiro *et al.*, 2018; R Core Team, 2018).

MK test

Further tests for trend in simulated time series were performed using the MK test (Mann, 1945; Kendall, 1955) and the more robust MK-TFPW (Yue, Pilon, and Cavadias, 2002; Yue, Pilon, Phinney, *et al.*, 2002). The MK test, which has been used previously in ecosystem indicator reporting (NEFSC, 2017a, b; Gaichas *et al.*, 2018), is a nonparametric approach 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, 1992). 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 (Yue and Wang, 2002). The MK-TFPW is a step-wise procedure developed by Yue, Pilon, and Cavadias (2002) and Yue, Pilon, Phinney, *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:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(y_j - y_i), \quad (2)$$

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

$$sgn(x) = \begin{Bmatrix} 1 \text{ if } & x > 0 \\ 0 \text{ if } & x = 0 \\ -1 \text{ if } & x < 0 \end{Bmatrix}. \quad (3)$$

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

$$V(S) = \frac{n(n-1)(2n+5)}{18}, \quad (4)$$

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

$$Z = \begin{Bmatrix} \frac{S-1}{V(S)} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{V(S)} & S < 0 \end{Bmatrix}, \quad (5)$$

is normally distributed with mean of zero and variance of one. 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 function (Wang and Swail, 2001).

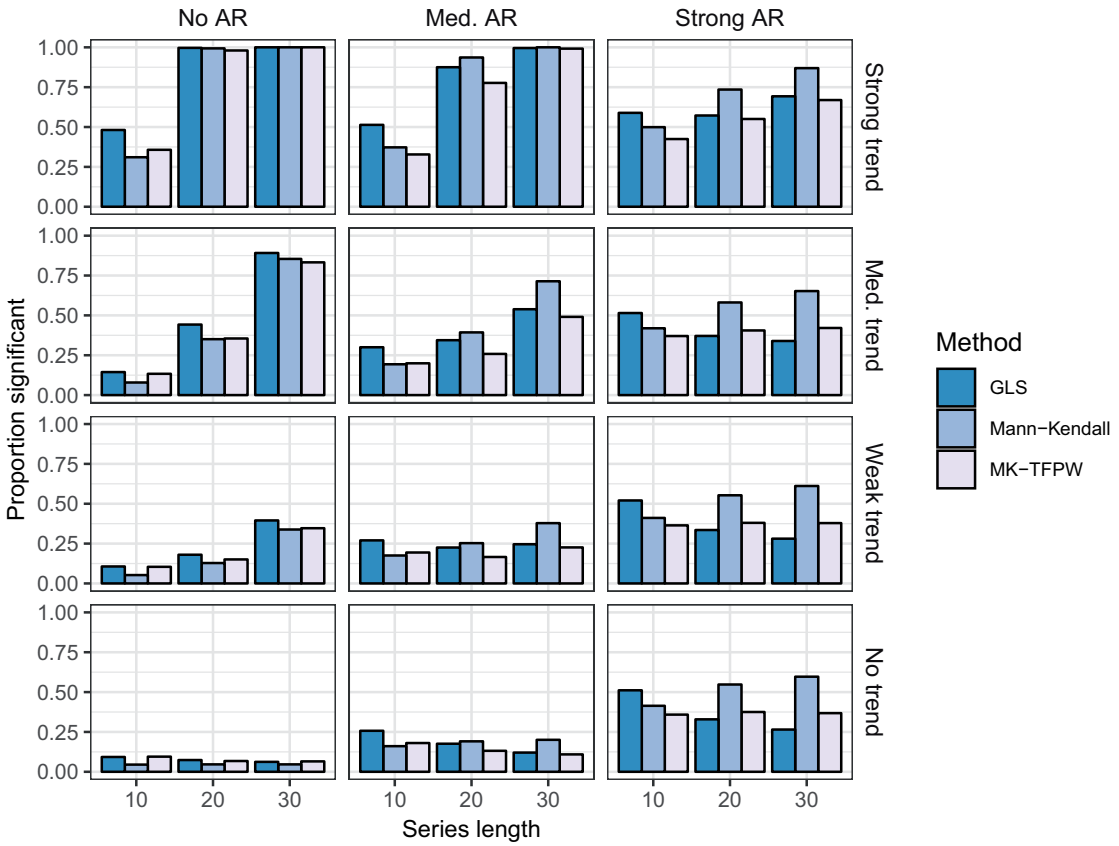

Figure 2. Test rejection rates of simulated time series ($p < 0.05$) among different combinations of AR(1) process strength ($\rho_1 = 0, 0.43, 0.9$) and trend strength ($\alpha_1 = 0, 0.026, 0.051, 0.147$). Bar colour indicates the test for trend that was applied.

Mann–Kendall trend-free pre-whitening

The MK-TFPW procedure as developed by Yue and Wang (2002) is composed of four steps:

(1) *Removal of trend*—The Theil–Sen estimator (Sen, 1968; Theil, 1992) is used to estimate the slope of trend b, which is removed from sample data if different from zero. b is given by

$$b = \text{Median}\left(\frac{y_j - y_i}{j - i}\right) \forall i < j, \tag{6}$$

where y_i and y_j are paired series values. Trend b is removed from the series by

$$y_t^{'} = y_t - bt, \tag{7}$$

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 found using

$$\rho_1 = \frac{1/n-1\sum_{t=1}^{n-1}\left[y_t - E(y_t)\right]\left[y_{t+1} - E(y_t)\right]}{1/n\sum_{t=1}^{n}\left[y_t - E(y_t)\right]^2}, \tag{8}$$

where $E(y_t)$ is the mean of the series and ρ_1 is the lag−1 autocorrelation coefficient. Serial correlation is then removed from the detrended series $y_t^{'}$ by

$$Y_t^{'} = y_t^{'} - \rho_1 y_t^{'}. \tag{9}$$

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

$$Y_t = Y_t^{'} + bt. \tag{10}$$

(4) *MK test*—Trend is assessed through the application of the MK test as discussed above.

The MK test and MK-TFPW were implemented using the Kendall and zyp packages (McLeod, 2011; Bronaugh and Werner, 2013).

Results

Throughout this study we adopt an alpha value of 0.05 to assess statistical significance. Overall, no method performed consistently well in all scenarios of simulated trend strength, time series length, and autocorrelation strength. We find time series length has a large effect on the sensitivity of each test (i.e. the true positive rate) (Figure 2), and performance was generally best across autocorrelation and trend scenarios when $N = 30$. With trend present and no autocorrelation, trends were only detected with >90% sensitivity when trend was strong ($\alpha_1 = 0.147$). Even with

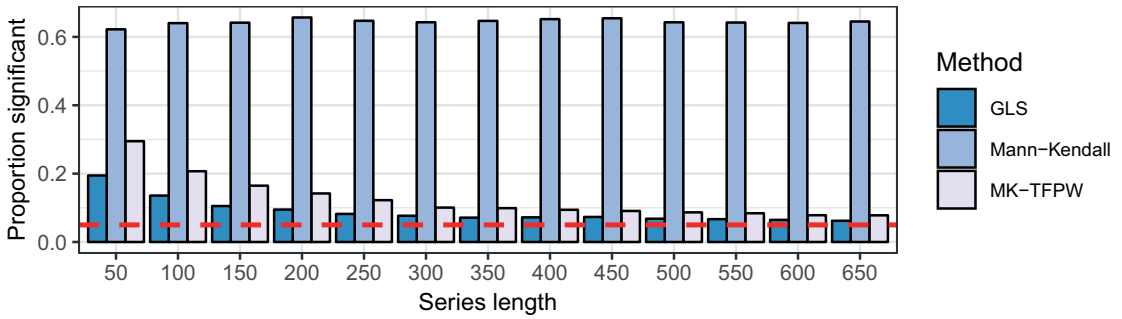

Figure 3. Test rejection rates ($p < 0.05$) when simulations were created under the parameters of no trend ($\alpha_1 = 0$), strong autocorrelation ($\rho_1 = 0.9$), and series lengths between $N = 50$ to $N = 650$. The dashed line shows the nominal rejection rate of 0.05.

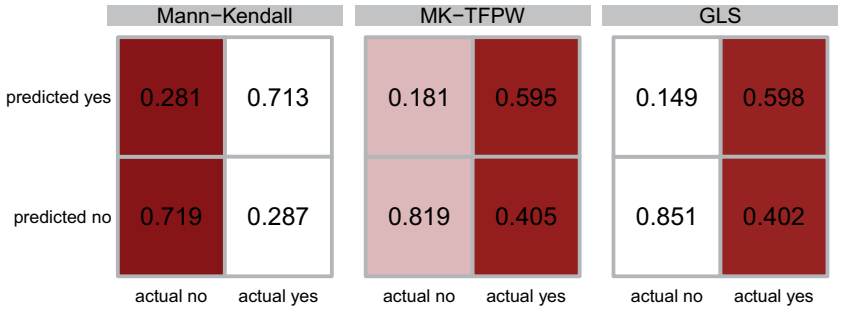

Figure 4. Confusion matrices showing aggregate results from testing for trend across all combinations of autocorrelation and trend strength when $N = 30$. Shading represents the performance of individual cells across tests, where darker shaded cells indicate a poorer outcome. For example, when $N = 30$, the GLS procedure falsely predicted a trend when there was none in 14.9% of cases (white), whereas this was true in 28.1% of MK simulations (shaded).

a strong trend and no autocorrelation, no test detected a trend in >50% of the series when $N = 10$. Again under no autocorrelation, the increased sensitivity associated with increasing series length diminished with reductions in trend strength across all tests. The GLS test showed the highest rejection rates compared to other tests under no autocorrelation (Figure 2, first column), although this effect was minimal (per cent increase in rejection rates between GLS and MK-TFPW was ~8%). For $N = 20$, 30 all tests returned rejection rates near the nominal significance level of 0.05 under the no trend and no autocorrelation scenarios. For $N = 10$, the MK-TFPW and GLS tests exceeded the expected nominal levels ($\text{MK-TFPW}_{sig} = 0.095$, $\text{GLS}_{sig} = 0.093$).

Under the no trend simulations, introducing autocorrelation was shown to lead to inflated rejection rates in the MK test, and the same bias in rejection rates can be seen for both GLS and MK-TFPW tests. The bottom row of Figure 2 shows that under no trend and medium to strong autocorrelation ($\rho = 0.433$ and $\rho = 0.9$), the rejection rate of the MK test increases with series length, but other tests showed decreases in rejection rates. Extending this no trend and strong autocorrelation scenario out to longer series lengths shows that the GLS test approaches nominal rejection rates of 0.05 only when $N > 650$ (Figure 3). The MK-TFPW approach performed poorly in this analysis, and also did not converge to nominal rejection rates for $N > 650$, although this work did not seek to identify a precise value of N where either test reached nominal levels. As expected, the MK test saw no reduction in rejection rates as N increased.

The GLS procedure performed the best under the no trend and strong autocorrelation scenario: when $N = 30$, the rejection rate for the GLS was 0.26; 26 and 56% 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 3% between $N = 10$ and $N = 30$. Under the same conditions and GLS approach, rejection rates decreased by 48%. However, the GLS approach also performed the worst under no trend and strong autocorrelation when $N = 10$.

Under strong autocorrelation ($\rho = 0.9$) 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 and series length on test results (Figure 2). Under these parameters, the GLS procedure was slightly more sensitive than the MK-TFPW test. The largest increase in sensitivity between the MK-TFPW and GLS tests in this scenario came when $N = 10$, where the GLS correctly identified trend 39% more often. Series length mattered least for the GLS in this scenario, as sensitivity decreased 3% between $N = 10$ and $N = 20$, but increased 30% for between $N = 10$ and $N = 20$ for the MK-TFPW.

When trend was weak or "medium" and autocorrelation was strong, neither the GLS nor MK-TFPW tests were able to detect trend in >55% of simulations regardless of series length. Interestingly, as series lengths increased when trend was weak (i.e. $\alpha_1 = 0.026$) and autocorrelation was strong ($\rho_1 = 0.9$), rejection rates tended to decrease for the GLS procedure, but remained stable for the MK-TFPW. 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.

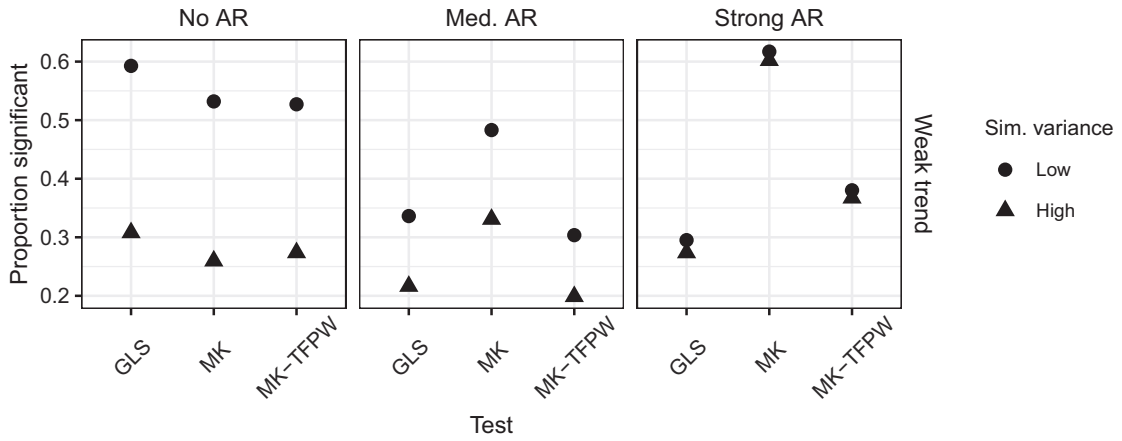

Figure 5. The proportion of significant results after simulating from two different levels of time series variance ($\sigma^2 = 0.3$ and $\sigma^2 = 0.8$) under weak trend across AR(1) error strengths when $N = 30$. The test rejection rates from the low variance simulations are shown by the circle, and high variance simulations by the triangle.

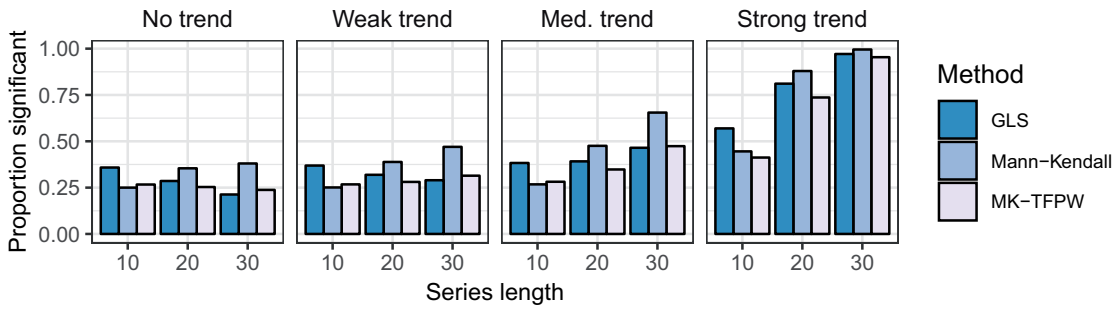

Figure 6. Barplots showing rejection rates of tests for trend under simulations generated with AR(2) error ($\rho_1 = 0.43$, $\rho_2 = 0.2$) and crossed with four levels of trend strength ($\alpha_1 = 0, 0.026, 0.051, 0.147$).

To demonstrate the effect of error variance on test sensitivity, we repeated the entire analysis with for two alternative values of residual variance ($\sigma^2 = 0.3$ and $\sigma^2 = 0.8$). We only provide a small subset of the results to illustrate the findings. As one would have expected (Figure 5), increasing error variance reduces the ability to successfully identify a trend. In addition, increasing autocorrelation will result in converging behaviour regardless of the value of sigma, the rate being dependent on the signal-to-noise ratio.

Testing for trend in simulations derived from an AR(2) process showed similar patterns of bias in rejection rates to the tests presented with strong AR(1) error (Figure 6). Rejection rates for the GLS approach, which included an AR(2) component in the model selection step, remained largely above nominal levels. Rejection rates also decreased slightly with series length when trend was weak or absent. When trend strength was medium, GLS rejection rates increased with series lengths. Under the MK-TFPW test, rejection rates across series lengths with weak or no trend remained largely the same, and rejection rates for the MK-TFPW did not start increasing with series lengths under trend strength was medium or strong. The MK test saw increasing rejection rates as both series length and trend strength increased, similar to its performance under the AR(1) scenario.

We next assessed the ability of each statistical approach to estimate the true trend slope (Figure 7). In the nonparametric case, we used Sen's slope [as derived in Equation (6)], which is a common statistic estimated alongside the MK and MK-TFPW significance tests. Sen's slope and the GLS estimator performed 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. Furthermore, trends falsely identified in the "no trend" scenarios tended to have the largest spread. As shown by the black median lines in Figure 7, both GLS and Sen's slope methods consistently overestimated trend slope when there was strong AR(1) error or series lengths were short. For example, when trend and AR(1) were strong, the median estimate of trend when $N = 10$ was 78.6% higher than the true value ($\alpha_{true} = 0.147; \alpha_{est} = 0.262$). When $N = 30$ under strong trend and AR(1), the median trend estimate was 21.4% higher than the true trend.

Discussion

Ecosystem reporting is vital to the development of IEAs, which lay out the framework for moving towards 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 changes to indicator status (Wagner *et al.*, 2013). Here, we addressed the shortcomings of identifying significant trends in indicator time series given the common problems of small sample size and autocorrelation.

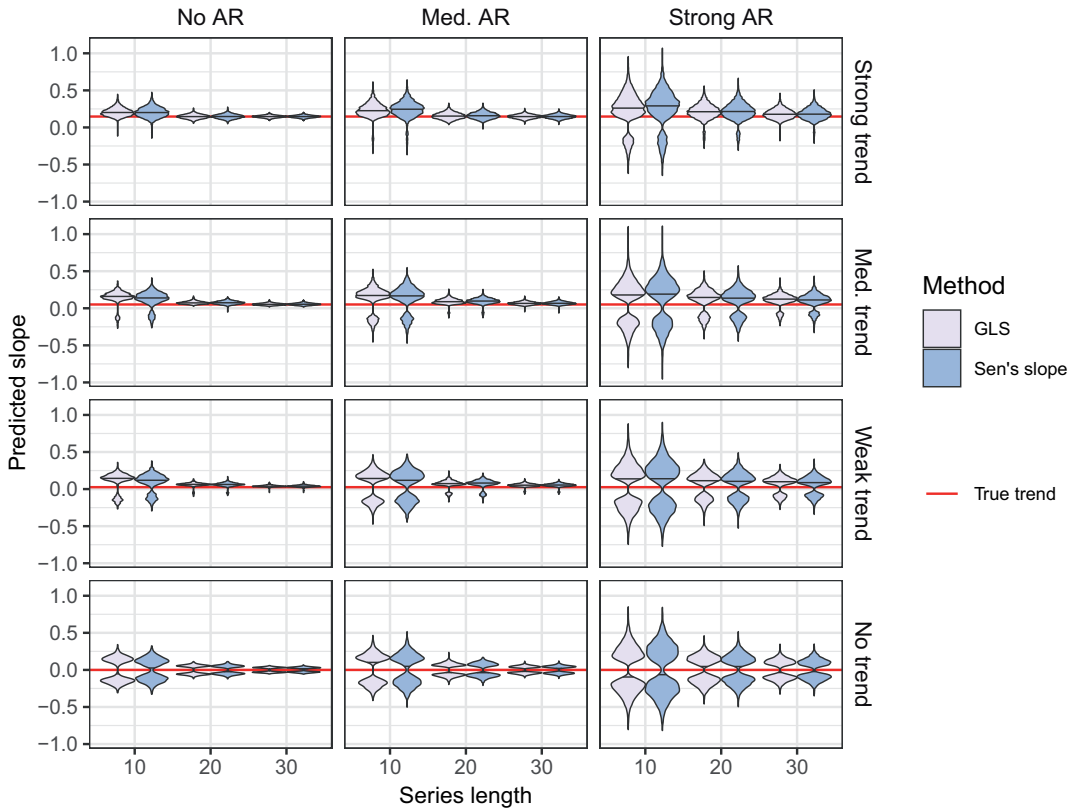

Figure 7. Violin plots showing probability densities of significant ($p < 0.05$) estimated trends from GLS and Sen's slope procedures under varying autocorrelation scenarios ($\rho_1 = 0, 0.43, 0.9$), simulation lengths ($N = 10, 20, 30$), and trend strengths ($\alpha_1 = 0, 0.026, 0.051, 0.147$). The size of each density estimate corresponds to the number of tests rejecting the null hypothesis of no trend under each scenario. Narrow lines represent the median slope estimate for each simulated scenario, and thicker background lines represent the true trend.

The key result from this study was that when no trend was present, none of the tests we examined returned rejection rates at the nominal 0.05 level under even weak amounts of autocorrelation, regardless of *a priori* incorporation of known simulation error structures (Figures 2 and 6). This held true for all lengths of time series in the study. Given this outcome, we advise caution when testing for trend in indicator time series using null hypothesis significance testing, and suggest a thorough examination of error structure and distribution family be accomplished prior to implementing tests for trend.

If we consider only simulations where $N = 30$, the GLS procedure we applied minimized false positives across AR(1) error strengths, although at smaller series lengths error rates between the MK-TFPW test and GLS were more similar. The GLS procedure also approached nominal rejection rates under the extended scenario of strong AR(1) error and no trend more rapidly than the MK-TFPW (Figure 3); however, neither reached 0.05 while $N \leq 650$. When no trend or autocorrelation were present, rejection rates for the GLS and MK-TFPW hovered above the nominal 0.05 level. This result was likely due to the influence of small sample size in both tests, as the GLS procedure relied upon a likelihood ratio tests that is known to be biased at small sample sizes (Bartlett, 1937). The breakdown of the MK-TFPW when $N < 20$ is not fully understood as references in the literature (Yue, Pilon, and Cavadias, 2002; Yue, Pilon, Phinney, *et al.*, 2002; Yue and Wang, 2002) limit its use to $N > 20$.

In assessing magnitude of trend slope (Figure 7), the MK-TFPW test and GLS performed similarly: both tended to overshoot estimates of trend strength when autocorrelation was present or series lengths were small. That rejection rates and parameter estimates are biased by autocorrelation in tests for trend is not a new concept (Storch, 1999; Yue, Pilon, and Cavadias, 2002; Yue, Pilon, Phinney, *et al.*, 2002; Yue and Wang, 2002; Beale *et al.*, 2010). However, by framing these results in the context of IEA, we hope to identify where current methodologies to assessing trend in time series may be improved for improving management outcomes. While we understand ecological data exists in many forms (for example continuous data, count data, data of proportions), in this study we focused our attention exclusively on continuous data to allow the comparison of GLS to the nonparametric methods described above. Using GLS on other types of data would not be appropriate. The natural model choice for these other data types would be a generalized linear model, although this was out of scope for this study.

In the context of hydrological literature, the upper limit of time series lengths seen in our ecosystem indicator data sets in the Northeast United States would be considered short (Bayazit, 2015). As discussed above, testing for trend in such short time series may result in an increased rate of false positives, but the failure to identify trend when it exists due to the presence of autocorrelation may also occur. We found that this was especially true when simulating from models with weak trend and

autocorrelated residuals. This effect was mediated by simulation variance, as simulations with lower variance had higher rejection rates than simulations with higher variance, although the effect of variance on rejection rate diminished as autocorrelation increased. This suggests that even when time series variance was relatively low (i.e. the 25th percentile of variance in empirical data), the presence of autocorrelation effectively masked the detection of trends by both GLS and MK-TFPW tests.

We have shown that there is no solution in small sample sizes (Figure 3), but refrain from suggesting there is no value in testing for trends in time series. Instead, we advise that a "shotgun" approach to assessing trends in many indicator time series without consideration of error structures and series lengths will likely lead to both Type I and Type II error. Furthermore, the implications of trade-offs in Type I and Type II error must be confronted prior to applying tests for trend, as detection of "false" trend does not imply an absence of biological meaning to the observed phenomena. As discussed in Vogel *et al.* (2013) and Bayazit *et al.* (2015), the management and societal impacts of an inflation in rates of mischaracterized trend must also be considered. Specifically, practitioners must weigh the consequences of over-preparation if a false trend is acted upon against under-preparation if a true trend is missed. From the perspective of IEA, the mischaracterization of trends in ecosystem reporting has the potential to propagate into risk assessments, ecosystem models, and potentially management decisions, leading to mismanagement of resources and eroded stakeholder trust in the scientific process.

Null hypothesis significance testing for trend is fraught with pitfalls related to interpretation of *p*-values showing "statistical significance" (see Wasserstein and Lazar, 2016 for the ASA statement on *p*-values). 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). Smoothing techniques, such as those implemented by the OSPAR Coordinated Monitoring Program for environmental contaminants (OSPAR Commission, 2014), have been used to assess status and trend in a more limited setting, and the impact of autocorrelation on performance of these models should be investigated further.

A different approach to trend assessment departs from null hypothesis testing altogether in favour of a Bayesian framework. Wagner *et al.* (2013) suggests Dynamic Linear Models (DLMs) for indicators of small sample size. Bayesian 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 residual variance but a biologically significant trend would be considered non-significant by a frequentist approach, but was 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. how likely is it that an undesirable trend exists in a time series? While Bayesian methods cannot side-step the reality of small sample sizes, their use provides managers with a probabilistic framework for decision-making that can be more intuitive than 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. 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.

Acknowledgements

We would like to thank the members of the State of the Ecosystem Synthesis Working Groups and members of the ICES Working Group on the Northwest Atlantic Regional Sea (WGNARS) for providing valuable feedback during the preliminary phases of writing this manuscript.

Funding

Support for this work was provided by the NOAA Integrated Ecosystem Assessment Program.

References

Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences, 160: 268–282.

Bayazit, M. 2015. Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review. Environmental Processes, 2: 527–542.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. 2010. Regression analysis of spatial data. Ecology Letters, 13: 246–264.

Bence, J. R. 1995. Analysis of short time series: correcting for autocorrelation. Ecology, 76: 628–639.

Blanchard, J. L., Coll, M., Trenkel, V. M., Vergnon, R., Yemane, D., Jouffre, D., Link, J. S. *et al.* 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: 732–744.

Bronaugh, D., and Werner, A., 2013. zyp: Zhang + Yue–Pilon trends package. http://cran.r-project.org/package=zyp (last accessed 1 March 2019).

Bundy, A., Gomez, C., and Cook, A. 2017. Guidance framework for the selection and evaluation of ecological indicators. Canadian Technial Report of Fisheries and Aquatic Sciences No. 3232. Fisheries; Oceans Canada, Dartmouth, Nova Scotia.

Butchart, S. H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P., Almond, R. E., Baillie, J. E. *et al.* 2010. Global biodiversity: indicators of recent declines. Science, 328: 1164–1168.

Canales, T. M., Law, R., Wiff, R., and Blanchard, J. L. 2015. Changes in the size-structure of a multispecies pelagic fishery off Northern Chile. Fisheries Research, 161: 261–268.

Gaichas, S. K., DePiper, G. S., Seagraves, R. J., Muffley, B. W., Sabo, M., Colburn, L. L., and Loftus, A. L. 2018. Implementing

ecosystem approaches to fishery management: risk assessment in the US Mid-Atlantic. Frontiers in Marine Science, 5: 442.

Garcia, S., Zerbi, A., Aliaume, C., Do Chi, T., and Lasserre, G. 2003. The ecosystem approach to fisheries. Issues, terminology, principles, institutional foundations, implementation and outlook. FAO Fisheries Technical Paper, 443. 71 pp.

Garfield, T. D., and Harvey, C. 2016. California Current Integrated Ecosystem Assessment (CCIEA) State of the California Current Report, 2016. Pacific Fishery Management Council, NMFS Report 1. pp. 1–20.

Hamed, K. H., and Rao, A. R. 1998. A modified Mann–Kendall trend test for autocorrelated data. Journal of Hydrology, 204: 182–196.

ICES. 2019. ICES Strategic Plan. International Council for the Exploration of the Sea, Copenhagen, Denmark. https://issuu.com/icesdk/docs/ices_stategic_plan_2019_web (last accessed 1 May 2019).

Karnauskas, M., Kelble, C. R., Regan, S., Quenee, C., Allee, R., Jepson, M., Freitag, A. *et al.* 2017. Ecosystem status report update for the Gulf of Mexico. Southeast Fisheries Science Center. http://www.aoml.noaa.gov/ocd/ocdweb/ESR_GOMIEA/report/GoMEcosystemStatusReport2017_NMFS-SEFSC-706_FINAL.pdf (last accessed 1 February 2019).

Kendall, M. G. 1955. Rank Correlation Methods. Hafner Publishing Co, Oxford, England.

Kleisner, K. M., Fogarty, M. J., McGee, S., Hare, J. A., Moret, S., Perretti, C. T., and Saba, V. S. 2017. Marine species distribution shifts on the US Northeast Continental Shelf under continued ocean warming. Progress in Oceanography, 153: 24–36.

Kulkarni, A., and Storch, H. V. 1992. Monte Carlo experiments on the effect of serial correlation on the Mann–Kendall test of trend. Meteorologische Zeitschrift, 4: 82–85.

Levin, P. S., Fogarty, M. J., Murawski, S. A., and Fluharty, D. 2009. Integrated ecosystem assessments: developing the scientific basis for ecosystem-based management of the ocean. PLoS Biol, 7: 23–28.

Mackas, D., Thomson, R. E., and Galbraith, M. 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: 685–702.

Mann, H. B. 1945. Nonparametric test against trend. Econometrica, 13: 245–259.

Mazerolle, M. J. 2017. AICcmodavg: Model Selection and Multimodel Inference Based on (q)AIC(c). https://CRAN.R-project.org/package=AICcmodavg (last accessed 1 March 2019).

McLeod, A. 2011. Kendall: Kendall Rank Correlation and Mann–Kendall Trend Test. https://CRAN.R-project.org/package=Kendall (last accessed 1 March 2019).

NEFSC. 2017a. State of the Ecosystem—Mid-Atlantic Bight. Northeast Fisheries Science Center, Woods Hole, MA.

NEFSC. 2017b. State of the Ecosystem—Gulf of Maine and Georges Bank. Northeast Fisheries Science Center, Woods Hole, MA.

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

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

Nicholls, N. 2001. The insignificance of significance testing. Bulletin of the American Meteorological Society, 81: 981–986.

Nicholson, M. D., and Jennings, S. 2004. Testing candidate indicators to support ecosystem-based management: the power of monitoring surveys to detect temporal trends in fish community metrics. ICES Journal of Marine Science, 61: 35–42.

NOAA. 2006. Evolving an Ecosystem Approach to Science and Management Through NOAA and Its Partners. National Oceanic; Atmospheric Administration. https://sab.noaa.gov/sites/SAB/Reports/EETT/eERRT%20-%20Final%20Report%20to%20NOAA%20Oct%2006.pdf (last accessed 1 February 2019).

OSPAR Commission. 2014. Levels and trends in marine contaminants and their biological effects—CEMP Assessment Report 2013. OSPAR Working Group on Monitoring and on Trends and Effect of Substances in the Marine Environment. https://www.ospar.org/documents?d=7366 (last accessed 1 May 2019).

Perry, R., Livingston, P., and Fulton, E. 2010. Ecosystem indicators. *In* Ecosystem-based Management Science and its Application to the North Pacific. PICES Scientific Report, 37. 184 pp.

Pinheiro, J., Bates, D., and R-core. 2018. NLME: Linear and Nonlinear Mixed Effects Models. https://CRAN.R-project.org/package=nlme (last accessed 1 March 2019).

R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Roy, A., Falk, B., and Fuller, W. A. 2004. Testing for trend in the presence of autoregressive error. Journal of the American Statistical Association, 99: 1082–1091.

Secretariat of the Convention on Biological Diversity. 2004. The Ecosystem Approach. https://www.cbd.int/doc/publications/ea-text-en.pdf (last accessed 1 February 2019).

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63: 1379–1389.

Shannon, L. J., Coll, M., Yemane, D., Jouffre, D., Neira, S., Bertrand, A., Diaz, E. *et al.* 2010. Comparing data-based indicators across upwelling and comparable systems for communicating ecosystem states and trends. ICES Journal of Marine Science, 67: 807–832.

Shin, Y. J., and Shannon, L. J. 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: 686–691.

Theil, H. 1992. A rank-invariant method of linear and polynomial regression analysis. *In* Henri Theil's Contributions to Economics and Econometrics, pp. 345–381. Ed. by B. Raj and J. Koerts. Springer, Dordrecht.

Vogel, R., Rosner, A., and Kirshen, P. 2013. Brief communication: likelihood of societal preparedness for global change: trend detection. Natural Hazards and Earth System Sciences, 13: 1773–1778. Copernicus GmbH.

von Storch, H. 1999. Misuses of statistical analysis in climate research. *In* Analysis of Climate Variability, pp. 11–26. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.377.1673 (last accessed 1 February 2019).

Wade, P. R. 2000. Bayesian methods in conservation biology. Conservation Biology, 14: 1308–1316.

Wagner, T., Irwin, B. J., Bence, J. R., and Hayes, D. B. 2013. Detecting temporal trends in freshwater fisheries surveys: statistical power and the important linkages between management questions and monitoring objectives. Fisheries, 38: 309–319.

Wang, X. L., and Swail, V. R. 2001. Changes of extreme wave heights in northern hemisphere oceans and related atmospheric circulation regimes. Journal of Climate, 14: 2204–2221.

Wasserstein, R. L., and Lazar, N. A. 2016. The ASA's statement on *p*-values: context, process, and purpose. The American Statistician, 70: 129–133.

Wiebe, P., Atkinson, A., O'Brien, T. D., Thompson, P. A., Hosie, G. W., Lorenzoni, L., Meredith, M. P. *et al.* 2017. What are Marine Ecological Time Series telling us about the Ocean? A status Report. Chapter 2: Methods and Visualization. IOC-UNESCO, IOC Technical Series, 19–35. 297 pp.

Woodward, W. A., Bottone, S., and Gray, H. 1997. Improved tests for trend in time series data. Journal of Agricultural, Biological, and Environmental Statistics, 2: 403–416.

Yue, S., Pilon, P., and Cavadias, G. 2002. Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. Journal of Hydrology, 259: 254–271.

Yue, S., Pilon, P., Phinney, B., and Cavadias, G. 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrological Processes, 16: 1807–1829.

Yue, S., and Wang, C. Y. 2002. Applicability of prewhitening to eliminate the influence of serial correlation on the Mann–Kendall test. Water Resources Research, 38: 4-1–4-7.

Zador, S. G., Holsman, K. K., Aydin, K. Y., and Gaichas, S. K. 2017. Ecosystem considerations in Alaska: the value of qualitative assessments. ICES Journal of Marine Science, 74: 421–430.

Zhang, X., Vincent, L. A., Hogg, W., and Niitsoo, A. 2000. Temperature and precipitation trends in Canada during the 20th century. Atmosphere-Ocean, 38: 395–429.

Handling editor: Marta Coll