# Inference and Attribution in Watershed Hydrology: Commentary on *Climate and agricultural land use change impacts on streamflow in the upper midwestern United States* (Gupta et al. 2015)

## FRST 590: Statistical Methods in Hydrology

**Submitted**: 29 March 2019 **Prepared by**: Dan Kovacek (35402767)

## 1.0 Motivation

A commentary on the current state of research into the effect of changing land use and land cover (LULC) on streamflow and floods at the catchment scale is presented in *Rogger et al. [2017]*. In the process of delineating gaps in the existing research, the authors describe the need for new approaches to obtain more general statements on impacts, citing the regularity with which studies obtain contradictory results for the same *kind of change*, or intervention. *Rogger et al. [2017]* highlights two such studies:

> *"Some recent publications such as the paper of Gupta et al. [2015] on the relative impacts of climate and land use changes on streamflow or that by Alila et al. [2009] about the effects of forest practices on floods have triggered scientific debates with the results being criticized by many scientists."*

To gain more quantitative insights into the impacts of LULC on hydrological trends, perhaps new quantitative approaches are needed, as *Rogger et al. [2017]* argues. A clearer understanding of the distinguishing characteristics and appropriate use of existing approaches may be equally valuable. *Cox [2006]* (p. 197) argues that the translation of a subject-matter problem into a formal statistical question is often the most critical part of the analysis. The aim of this paper is to determine whether the conclusions arrived at in *Gupta et al. [2015]* are justified by the approach. First, a general outline of statistical inference is presented to provide context for the subject-matter development and translation problem. A summary and discussion of the *Gupta et al. [2015]* study then follows to determine its capacity for inference, and finally the conclusions of the study are compared to the model's capacity for inference.

# 2.0 Background

## 2.1 Paradigms of Statistical Inference

Some of the difficulty in reviewing the statistical literature is due to the prevalence of value statements invoking blame, guilt, and fear, none of which contribute to the understanding of science. (Lloyd and Oreskes 2018) Certainly no discipline or body of literature is perfect, however some of the lack of understanding of statistics often decried in the literature may instead be an indication of the similarities between the established paradigms of statistical inference. The prominent statistician D.R. Cox broadly defined inferential statistics by the following paradigms, presented here in the briefest of summaries:

- **Frequentist**: inference of system behaviour is measured from data alone, assuming the unknown parameter of interest is *fixed*. (Cox 2006) (p. 24) The traditional approach of R.A. Fisher, Neymann, and Pearson is to formalize a set of rules to govern behaviour such that in the long run, we won't be wrong too often. (Lakens 2017)
- **Bayesian**: inference of system behaviour is measured from data, but prior knowledge is incorporated by assuming the unknown parameter of interest is *probabilistic*. (Cox 2006) (p. 24) Quality of evidence is expressed in terms of 'degrees of belief'. (Lakens 2017)

*Lindley [2000]* (p. 293) states that the concern of statistical analysis is evaluating uncertainty, and the fundamental problem of statistical inference is in using past data to predict future data. Uncertainty in quantifying some parameter of interest can be separated into two distinct and fundamental types: *natural* uncertainty is attributatble to the variability of the underlying stochastic process, while *epistemic* uncertainty lies in the incomplete understanding of the greater system under study. (Bruno Merz and Thieken 2005) Quantifying information about some unknown parameter or a system of interest is related to the separation of aleatoric (natural) and epistemic uncertainty. (Weijs, Van de Giesen, and Parlange 2013)

While the treatment of the statistical discipline in *Lindley [2000]* entirely avoids the language of causality and attribution, causal inference is a more recently established paradigm (despite independent origins in the 1920s from both Barbara Burks and Sewall Wright) putting causality central in the approach to statistical inference. (Pearl 2009) (p. 1) *Pearl [2018]* argues that causality is not just an extreme condition of association, as much of the field has historically treated it. The capacity to evaluate nonexistent *"what-if"* scenarios, or counterfactuals, is the more advanced level of inference that the field of artificial intelligence strives for, and mere association (i.e. linear regression, machine learning) is the most primitive level. (Pearl and Mackenzie 2018) Statistical inference can thus reasonably include both *associative* and *causal* sub-categories. *Likelihood* and *Information* are additional established paradigms of statistical inference that are beyond the scope of this discussion.

The variety of ways of expressing like methods is a natural outcome of the application of statistics across the breadth of academic disciplines with little reason or opportunity to share ideas. Proof of the apparent interchangeability of methods is easily seen in a random sample of titles by submitting to an academic journal database the key words “Frequentist” and “Bayesian”. Even the work of a single author may evolve over time to favour different paradigms, as well established statisticians have noted their support for one paradigm or other evolving over their career. ((Pearl and Mackenzie 2018), (Lindley 2000), p.336)

Similarly, part of the challenge in reviewing the hydrological literature lies in the nuanced description and integrated application of statistical methods. In the field of hydrological research, there are numerous and varied approaches to the measurement and prediction of runoff, as well as to the attribution of physical causes to trends in observed data. (Viglione et al. 2016) Causality is invoked by *Viglione et [2016]* by stating *“the attribution of physical causes”*. *Rogger et al. [2017]* also invokes the language of causality in their criticism of the discipline:

> *“Studies that examine the impact of land use changes on streamflow and floods often obtain contradictory results for the same kind of change.”*

Analysis of hydrometric data is undertaken in order to base decisions upon expectations of future behaviour of some unknown parameter of interest. To gain any level of practical understanding of runoff at the watershed level, a model of some form must be employed. Input variables to hydrological models are discrete observations in time and space, representing samples of component and mechanism behaviours of the hydrologic cycle. As such, hydrological analysis is inherently inferential, rather than merely descriptive.

One of the central tasks in the study of watershed hydrology is the determination of an appropriate model for the characterization of timing and quantity of runoff at a spatio-temporal scale of interest. It is the model development that determines the paradigm of statistical inference of the study. The established inferential paradigms are not mutually exclusive, rather there are a variety of valid approaches to characterizations of the system under study, and the validity of the approach is dependent upon on the question being asked of the data. ((Hoaglin et al. 1991), p. 24)

### 2.2 Modelling Processes: Deterministic, Stochastic, and In-between

> *“Rather than idealized angels of reason, scientific models are powerful clay robots without intent of their own, bumbling along according to the myopic instructions they embody.”* (McElreath 2018)

Statistical study has two fundamental steps according to *Lindley [2000]*,. The first is model construction, which is necessarily subjective and requires careful consideration in order to ensure the model is consistent with reality. The second is analysis, which is routine and ripe for automation. (Lindley 2000)(p. 303) The function of the model is to translate a subject-matter question into a formal statistical question. (Cox 2006) (p. 197) However, even an otherwise correctly developed model can introduce errors if it is applied beyond the range of calibration data. (Alila et al. 2009)

Process-based analysis investigates the mathematical relationships describing pathways for the movement of water (Bracken et al. 2013):

> *"While there is a current trend favouring process-based hydrological analysis over purely empirical approaches, there remains a lack of consensus in the definition and measurement of hydrological connectivity."*

The discussion of *"hydrological connectivity"* in *Bracken [2013]* suggests there is plenty of room for new developments in deterministic modelling, with no mention of stochastic processes or Bayesian inference. However, deterministic (event-based) approaches are not suited to common questions such as prediction of extreme event behaviour, where stochastic (frequency-based) approaches are better suited. (Alila et al. 2009)

Hydrological processes occur on many different scales, both deterministic and stochastic in nature. In describing the complexity of systems, *Sivakumar [2017]* places the two terms at opposing ends of the scale of complexity, and adds a third term to form a continuum between extremes:

- **deterministic**: order and dependence exist at certain spatiotemporal scales, such as daily discharge and daily temperature,
- **stochastic**: nonlinear interactions dominate the hydrologic cycle yielding random and irreproducible states of the real system, and
- **chaotic**: systems requiring three or more independent variables (degrees of freedom) to describe state exhibit chaotic behaviour, (Gleick 1987) which is deterministic in the short term, yet irreproducible and unpredictable in the long term due to sensitivity to initial conditions.

In the hydrologic cycle, interactions between components and mechanisms occur in many different ways, directly or indirectly, often in feedback forms, and with varying degrees of nonlinearity. Natural river discharge at a daily time scale has been shown to be a deterministic-chaotic process. (Kędra 2014) With a change in time scale, there is merit to deterministic approaches to represent the significant deterministic nature of seasonal or annual cycles of river flow, while stochastic approaches are suitable where complex interactions govern a system or process of interest and the ability to observe the data is limited. (Sivakumar 2017) Deterministic, or process-based models are combined with stochastic models in practice and can be complementary. ((Koutsoyiannis 2016), (Sivakumar 2017)) While in the short term there is determinism and order in a low complexity system, such as the rainfall-runoff response of a small, highly developed catchment, the sensitivity of even a simple rainfall-runoff model of few degrees of freedom can be highly sensitive to initial conditions and as a result unpredictable in the long-term. (Sivakumar 2017)

A hydrological model can either be designed to yield some level of certainty about an unknown parameter (or treatment effect) given specific requirements for input, or conversely can yield quantitative statements about the quality of estimation of the unknown parameter (or treatment effect) can be determined given fixed input. The latter is the more common case when observing natural systems due to the practical realities of collecting field measurements. As with the choice of statistical paradigm, the choice of model is highly dependend upon the question being asked of the data. How the quality of predictions are communicated is addressed next.

### 2.3 Communicating Effects: P-Value, Significance, Confidence, and Equivalence

Interpretations of model outcomes are communicated using specific metrics to make the information useful for a practical application, regardless of the model type. The way the effect of a treatment is measured and communicated has been the source of ongoing debate in the literature ((B. Merz et al. 2012), (Alila et al. 2009), (Lloyd and Oreskes 2018), (Pearl and Mackenzie 2018), (Gupta et al. 2015)) Since real systems tend to be highly complex, it is necessary to have different approaches to evaluate the effect of some parameter of interest or treatment. This discussion is limited to terms relevant to the methodology presented in *Gupta et al. [2015].*

The underlying principle of the frequentist analysis recognizes that drawing conclusions from data is error prone, assuming long-term use of the implications of data, or the unknown parameter of interest. (Cox 2006) (p. 24) Suppose a study aims to measure the effect of some treatment with the null hypothesis that there is no effect (H = no effect). The study wants to have a high level of certainty that the outcome, stated as the rejection or affirmation of H, will be the same *in the long run*, for future experiments (given the same number of observations). The confidence interval is *chosen* by setting rules for testing such that (typically) 95% of future outcomes are consistent with the assumption that H is true. The remaining studies resulting in the opposite outcome with respect to the assumption of H being true is the type 1 error rate, or the alpha level. The concept of the 95% confidence interval (CI) with long-term, finite sampling from a stationary population is illustrated in Figure 1, below.

The p-value is a measure of surprise in the data. The lower the p-value, the greater the surprise. The significance of a statistical test is determined by comparing the result of the test (the p-value) to the alpha level ($\alpha$), or type 1 error rate. Critics of the use of p-values point to the large number of studies reporting “no effect” for $P > \alpha$, when this is entirely not the case. (Lakens 2017) Assuming H is true, where the result of a significance test is $P > \alpha$, the only

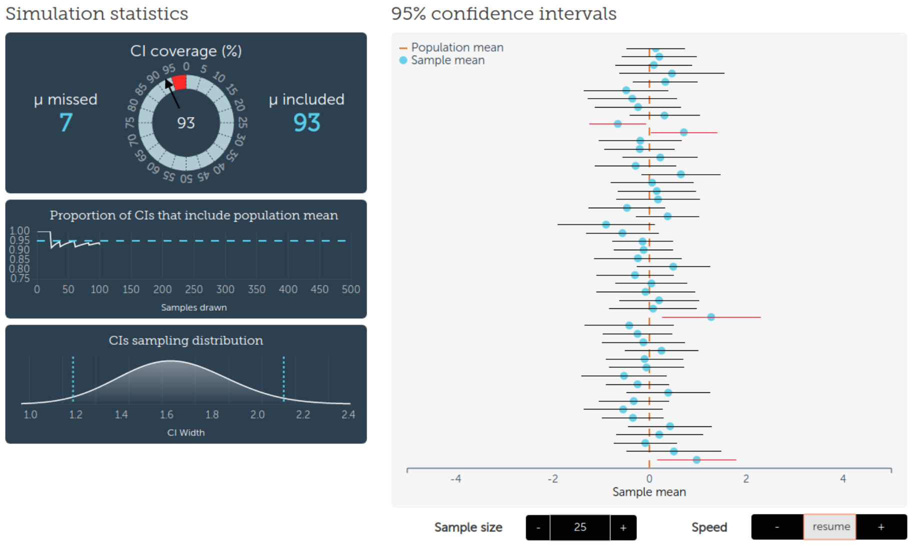

Figure 1: Interactive illustration of long-term sampling given n=25, CI=95% (source: Kristoffer Magnusson @ RPsychologist.com)

correct conclusion is that the data are not surprising. (Lakens 2017) However, it is still common in the literature to see statements that an effect is "statistically significant" if $P < \alpha$ and concluding "no effect" with $P > \alpha$. (Lakens 2017)

A statement of the quality, or confidence level, of an estimated parameter is only as informative as it relates to some size of effect that is interesting or useful (Lakens 2017). The size of an effect might be evaluated in terms of the difference in some parameter between two groups, one receiving a treatment and one not (the control). A statistical test in this case expresses the difference of the unknown parameter in terms of equivalence. The equivalence measure relates a *subjective* interval, or magnitude of an effect, that is considered to be of practical significance. For instance, if a parameter of interest is evaluated in two independent samples, and the difference between the two is determined to be within the measurement precision, it cannot reasonably be claimed that an effect has been measured. Note that this statement does not claim there is no effect. In many cases the measurement error may be small relative to some practical effect size of interest, and there may be a practical effect size related to some outcome, such as a materially different design or policy implementation. In this case, a statistical equivalence test should use bounds that are of practical significance to the application. Even when a test finds significance, it is a best a start for further analysis of implications. ((Hoaglin et al. 1991), p. 3)

*Lindley [2000]* (p. 299-300) argues that significance level and confidence, which are descriptions of parameters and not data, do not obey the probability calculus, and holds that the connection between two sets of data, expressed through a parameter $\theta$, can only be evaluated probabilistically. The distinction between significance, confidence, and probability is described as the following (where H is the hypothesis that the treatment has no effect) (Lindley 2000) (p. 299-300):

- **significance level**: the probability of some aspect of the data, given H is true,
- **probability**: your probablity of H, given the data
- **confidence**: probability that the interval includes $\theta$
- **probability (restated)**: probability that $\theta$ is included in the confidence interval

The differences in the above statements are subtle in print, but have important mathematical consequences, as *Lindley [2000]* details.

### 2.4 Analysis of Variance (ANOVA)

Researchers are often interested in evaluating the effect of some treatment that cannot be directly measured. ANOVA is a statistical methodology that involves a response variable and structural components, or factors that classify the measured variable into subgroups. It is important to note that the appropriate application of ANOVA is dependent upon the assumptions of normal (gaussian) distribution of residuals and homogeneity of variances. Factors often represent treatments expressed in terms of levels, or versions. ((Hoaglin et al. 1991), p. 2, 50) The term *levels* is used to distinguish numeric categorization, and the term *versions* to represent non-numerical categorical data (i.e. male/female, or a binary variable such as treatment/no treatment). The simplest form of ANOVA is a 1-way layout, wherein each level (group) of a factor has a number of observations and we are interested in evaluating the difference of some factor between groups. ((Hoaglin et al. 1991), p 61)

ANOVA focuses on the differences in components by using overlays of groups to evaluate the group to group variability. The differences in variability in a group and between groups are what is evaluated in ANOVA to determine differences due to some effect. ((Hoaglin et al. 1991), p. 73). It is often the case that practical application calls for using the implications of analysis beyond the circumstances sampled, however "*...the interpretation cannot extend beyond the circumstances sampled, represented, or illustrated in the data before us*" ((Hoaglin et al. 1991), p. 336)

The preceding background discussion presented a general overview of statistical inference, modelling, and evaluation, and was written to provide specific context for the summary of *Gupta et al. [2015]* that follows in Section 3.

## 3.0 *Climate and agricultural land use change impacts on streamflow in the upper midwestern United States* (Gupta et al. 2015)

### 3.1 Summary

Analysis of measured runoff between 1909 and 2009 at 29 streamflow measurement stations in Iowa and Minnesota demonstrates an increasing trend of annual runoff, coincident with a positive trend in annual precipitation. *Gupta et al. [2015]* attempts to quantify the relative contributions of increased precipitation and changing land use and land cover (LULC) to the observed increase in runoff. A secondary goal of the study is to explain the observation of constant evapotranspiration (ET) over the same period of time, by attempting to disaggregate the effects of changing LULC (increasing ET) and loss of wetlands (decreasing ET).

Separating the measured record into two periods consistent with a Before-After-Control-Impact (BACI) analysis framework, *Gupta et al. [2015]* cites the extensive adoption of plastic drain tile in agricultural practices in the mid-1970s as the treatment (or intervention), consistent with the breakpoint adopted in previous studies. *Gupta et al. [2015]* tests for a change in the relationship between streamflow versus precipitation by using a series of linear regression models of varying complexity, presented in more detail in the subsequent section. Reporting results statistically significant at the 5% level (95% confidence interval, or $P < 0.05$) the study concludes that the relationship between precipitation and runoff was statistically similar between the first and second period, in other words no effect of LULC change was detected. However, some coefficients demonstrate a shift in the relationship, which is explained as a vestige of a simplified model. Visual evaluation of 5-year moving average plots of precipitation and runoff showed a shift of the data points along the axis of the best fit line, which is reported to suggest increased runoff is attributable to increased precipitation alone, although the ANOVA methodology is not applied to test equivalence of the best fit lines in this case. A single control watershed with limited agriculture and development found no statistical difference in the relationship between precipitation and runoff across the two periods. Given the results of the statistical tests, the authors conclude that increased streamflow over the study period is mainly due to increased precipitation, and that the LULC change had *no effect*.

In terms of the secondary question of the effect of ET on the relationship between precipitation and runoff, *Gupta et al. [2015]* concludes that the lack of effect of LULC change on streamflow is the result of comparable ET over the two periods. The focus of this commentary is primarily the first question of the effect of LULC change.

### 3.2 Discussion of the Study Assumptions and the Subject-Matter Problem

*Gupta et al. [2015]* asks a specific question of the data: how much of the observed increasing trend in runoff in the upper midwestern US is attributatble to improved soil drainage, and how much is attributable to the observed increasing trend in precipitation? Restated in the terms introduced in Section 2, what is the effect of the treatment (LULC change) on the parameter of interest (mean annual runoff)? Missing from the formulation of the subject-matter problem is the question of the effect size of interest, and a practical interpretation.

To place the approach of *Gupta et al. [2015]* within the general overview of statistical inference paradigms described in Section 2.2, it is clear that language of causality is invoked throughout the paper (i.e. "*higher annual streamflows in recent periods are mainly* ***due to*** *higher precipitation*", "*there was* ***no effect*** *of land use changes on the streamflow versus precipitation relationship.*"), however there is also an explicit signal to the frequentist paradigm:

> "*As with many statistical analyses in which explanatory variable levels are not under control of the experimenter, relating streamflow to precipitation as was done in this study by itself does not suggest a cause and effect relationship.*"

How does the evaluation of statistical significance in *Gupta et al. [2015]* reflect the practical interpretation of the subject-matter problem? What is the purpose of evaluating the precipitation-runoff relationship at the *annual* scale? As *Belmont [2015]* points out, the annual timescale used in the study obscures more critical, higher frequency impacts of artificial drainage. Other research in LULC in the midwest US has demonstrated the link between LULC and hydrological response, in the practical context of sediment production and transport, nutrient cycling, and river ecology. (Foufoula-Georgiou et al. 2016) The link between increased streamflow and water quality is made at the outset, and addressed no further in the study.

Is there some probability of a quantifiable effect size of LULC change on average runoff on the *annual* scale that would warrant changes in policy or agricultural practice? If annual runoff is a proxy, or indicator, of changes to characteristics or processes relevant to agriculture, the approach of *Gupta et al. [2015]* does not address such connections. The development of the subject-matter problem in *Zhang and Libra [2006]*, also investigating increasing trends in runoff in Iowa, sets a practical context for the research question: changing baseflows in rivers across Iowa are changing the characteristics of water pollutant delivery. (Zhang and Schilling 2006) *Alila et al. [2009]* (and references therein) directly addresses the practical question of whether changes in annual means are a proxy for changes in variability, magnitude, and frequency of extremes.

Changes to seasonal runoff in terms relevant to agricultural productivity include timing and magnitude of extremes at different timescales, erosion, freshet (snowpack), and seasonal or monthly runoff distribution relevant to critical periods such as crop uptake. *Gupta et al. [2015]* goes no further to address such practical questions beyond qualitatively discussing the trends in runoff ratio increasing in May-June, and decreasing in September-October, despite precipitation trends in the opposite proportion. (Schottler et al. 2014) *Gupta et al. [2014]* addresses these observed trends in seasonal runoff ratio to defend to discuss changes in the annual soil storage distribution , but does not discuss the implications of increased soil moisture for floods, and disregards the practical context of the *Shottler [2014]* study in investigating the issue of erosion.

Pan evaporation is a measure of evaporative demand, and is driven by humidity gradients, temperature, wind speed, and solar insolation. (Roderick et al. 2007) Investigating a widely observed global trend in decreasing pan evaporation, *Roderick et al.* [2007] modeles the components of evaporative demand and attributed the decline in measured pan evaporation between 1975 and 2004 to a reduction in wind speed along with regional reduction in insolation. Trends in ET over the study period were evaluated for the *Gupta et al. [2015]* study based on pan evaporation data from a single location to represent ET across the midwestern US. Average wind speeds are highly spatially variable across Minnesota and Iowa, and crop changes have been connected with increased precipitation recycling in the midwest. (Harding and Snyder 2012) An increase in the proportion of precipitation originating from evaporation in the same region is surely an indication of the effect LULC change.

### 3.3 Discussion of the Statistical Methodology

First, temporal trends in annual precipitation are evaluatd to validate the results of other studies. Trend analysis of annual precipitation is done in two ways: one using the Mann-Kendall nonparametric test, and the other by calculating mean annual precipitation for three periods: 1920-1949, 1950-1979, and 1980-2009. Both methods indicate an increasing trend in precipitation over the study period.

The premise of the methodology for evaluating the effect of LULC change, or the statistical question as stated in *Gupta et al. [2015]*, is that a significant shift over time in the precipitation-runoff relationship suggests an effect of LULC change, whereas no shift indicates that increasing runoff is due to increasing precipitation alone. Mean annual precipitation versus mean annual runoff are divided into two groups: all years, and post-1975. The difference betweent the two groups associated with the input and response factors is evaluated by Analysis of Variance (ANOVA). The first group (all years) varies in start date for most stations, and the second period (post-1975) consistently ended in 2009. The breakpoint assumed to define the post-treatment period is 1975, corresponding to historical evidence of widespread adoption of plastic tile drainage in agriculture, and also consistent with related, independent studies. Using overlapping time periods for groupings instead of discrete and independent periods in ANOVA is consistent with avoiding the logical fallacy of composition.

The the number of samples comprising the parameter estimates (number of points the best-fit relationship of each station is based upon) in the first group varies ($29 \leq n_1 \leq 72$), and in the second group it is constant ($n_2 = 34$). Sensitivity of the results to variable sample sizes of group 1 is not addressed in the study, however follow up commentary point out that the methodology is not robust to slight changes in the time period, and that the methodology is not robust to the systematic removal of a single data point. (Belmont et al. 2016) *Hoaglin et al. [1991]* (p. 227) presents an example of why improper treatment of unbalanced cases can give unreliable results, though it is not known how the ANOVA methodology was handled for the unbalanced case in *Gupta et al. [2015]*.

The series of models used in *Gupta et al. [2015]* to test the relative effect of precipitation and LULC change are described by the following equations:

$$ln(Q_{all}) = \beta_0 + \beta_1 \cdot P_{all} + \beta_2 \cdot I + \beta_3 \cdot P_{post} \cdot I \tag{1}$$

$$ln(Q_{all}) = \beta_4 + \beta_5 \cdot P_{all} + \beta_6 \cdot I \tag{2}$$

$$ln(Q_{all}) = \beta_7 + \beta_8 \cdot P_{all} \tag{3}$$

In the first model (1), $\beta_0$ and $\beta_2$ represent the intercepts of the best fit lines for the entire record, while $\beta_1$ and $\beta_3$ represent the slope of the best fit lines. In each model, $I$ has a value of 0 or 1 based upon the period, such that the post-change period is assigned a separate coefficient from the coefficient assigned to the full record. Using ANOVA with a BACI framework to test equivalance of the coefficients, the LULC change is said to have an effect if the coefficents of the second period are statistically significantly different than those describing the entire record. The results state that for all watershds the two periods are not statistically different based on model (1).

The second model (2) was then tested to focus on the intercept value of the precipitation-runoff relationship, assuming a fixed slope for the post-change period. What advantage there is to assuming a constant slope is unclear, as the variance is artificially suppressed by doing so. For 19 of 29 watersheds, the two periods were again reported not to be statistically different. The remaining 10 out of 29 cases resulting in statistically unique intercepts is roughly equivalent to the type II error rate, (Foufoula-Georgiou et al. 2016) highlighting the critical lack of power of the statistical test due to the small sample size. *Gupta et al. [2015]* explain the 10 significant results by stating, "*The shift was primarily because some of the variability in the slope between the two periods shifted to its corresponding intercepts when the slope was kept constant in model 2.*" It is not mathematically clear how this redistribution of variability occurs.

Model (3) is said to be applied for those watersheds where $\beta_6$ is not significant to describe the precipitation-runoff relationship for the whole record. However, the results state the opposite, that model (3) was applied to watersheds exhibiting a significant shift in $\beta_6$. In any case, if the simplification of assuming a constant slope artificially adds variability to the intercept coefficient as *Gupta et al. [2015]* states (it is not clear it does), it is unclear why a further simplified model (3) is investigated.

Part of the difficulty with relying too heavily on significance testing is that the true value of any effect or quantity is non-zero. ((Hoaglin et al. 1991), p. 3). In the case of evaluating the effect of a treatment in a hydrological cycle that has many interactions, if there are large interactions between factors, changing one factor at a time can be ineffective ((Hoaglin et al. 1991), p 19). A comparison of relative importance depends upon the variety of versions or levels considered for a factor, and when factors interact, the idea of relative importance can be very complicated ((Hoaglin et al. 1991), p 26) Confidence limits in ANOVA tests for significance are highly sensitive to the assumption of normality in the sample, and even small departures from normality can produce a large index of kurtosis, with the effect of dramatically increasing the confidence interval of sample variance. ((Hoaglin et al. 1991), p. 160-161)

Since many relationships worth investigating are nonlinear, the transformation of the response variable allows for nonlinear relationships to obey the assumptions of normal distribution and homogeneity of variance, and may lead to a simpler fit with fewer interactions that use only main effects. ((Hoaglin et al. 1991), p. 366-369) The log transformation of the response variable is standard practice in some disciplines, however a comparison of methods of investigating nonlinear processes in the biological field has pointed out how the log transform can alter interpretation of the outcome. (Stanton and Thiede 2005) In addition, the results of standard statistical tests performed on log-transformed data are often not relevant to the original data, (Feng et al. 2014) and may result in misleading measures of effect size. (Wilcox 1995)

Finally, the BACI methodology itself has shown susceptibility to overstating evidence for associations between groups of response variables, particularly when some level of serial autocorrelation exists in the data. (Murtaugh 2002) Longer-term oscillations of annual-scale cycles, such as the Pacific Decacal Oscillation, suggest the presence of serial autocorrelation, and again raises the issue of the use of annual timescale for the investigating the treatment in question.

### 3.4 Conclusions

Research into the nature of anthropogenic effects on the natural hydrological cycle must have consistency between the problem statement, the construction of the statistical question, and real hydrological systems. While individual components of the study presented in *Gupta et al. [2015]* are standard practice in research, there are significant issues with consistency of application and a lack of practical interpretation.

The question of the effect of LULC change on the relationship between precipitation and runoff on the annual timescale does not communicate any practical scientific, socio-economic, or environmental information. Further, the conclusion that LULC has no effect on increased runoff is inappropriate given the limited capacity of the approach for explaining effect sizes. Artificial suppression of variability, whether by transformation of the response variable or by model development, can produce misleading measures of effect sizes. Statements of "no-effect" backed by statistical significance tests, especially with low power, communicate an inflated message of confidence in statistical tests with low capacity for such conclusions.

The final section is added as an optional appendix to take a closer look at the consistency between the statistical question and the real hydrological system using the causal inference paradigm laid out in *Pearl [2009]*.

## 4.0 Additional Notes on Causality

Taking the subject-matter problem presented in *Gupta et al. [2015]*, and restating and simplifying using the counterfactual paradigm of *Pearl [2018]*, the consistency between the subject matter question, the statistical problem, and the reality of the hydrological system is inspected closer. According to *Pearl [2009]* (p. 183), an investigation of the statistical association between two variables (A and B) is confounded by a third variable that influences both A and B. Evaporation is discussed in *Gupta et al. [2015]* in a separate investigation of the trend of evaporation over time, and is only qualitatively addressed in terms of its interrelatedness with precipitation and runoff. However evaporation influences both precipitation and soil moisture (and thus runoff), so it is considered a confounding variable.

Suppose we have two hypothetically identical basins, and we subject them to identical precipitation over some period of time. If we apply a treatment of drain tile to one plot (LULC change), will annual runoff increase? Restating the problem in this way highlights the issue with problem formulation in *Gupta et al. [2015]*, as represented by equations (1) to (3) from Section 3.3, where the only factors considered are precipitation (as a continuous variable) and drainage (treatment as a binary variable). A simplified graphical representation of this system representation is shown in Figure 2.

If infiltration and evaporation are effectively constant over the time period, as *Gupta et al. [2015]* assumes, there will be no arrows to connect these processes to the input or response variable. However, if we recognize that ET is a process interacting with precipitation and soil storage, and also recognize that soil storage is a mediating factor, we arrive at the system representation shown in Figure 3.

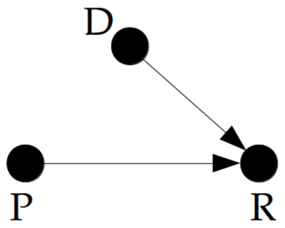

Figure 2: Graphical representation of the basic subject matter problem of precipitation (P), and runoff (R), and a treatment: drainage (D).

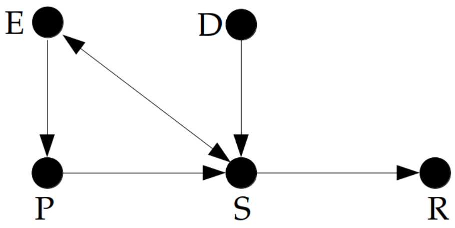

Figure 3: Graphical representation of the modified subject matter problem incorporating evaporation (E) as a factor interacting with soil moisture and precipitation, and representing soil moisture as a mediating factor.

Note the arrow indicating a directional causal effect of evaporation on precipitation allows for the study of the proportion of precipitation coming from precipitation recycling in the midwest as investigated by *Harding [2012]*. Precipitation is shown to affect evaporation through soil moisture, although given evaporation is negligible during precipitation events, we could also make a bi-directional connection between E and P. Soil moisture (or soil storage) is a mediating factor between precipitation and runoff, and the addition of drainage is imposed (or not, as a treatment) upon the soil storage factor, which mediates between precipitation, evaporation, and runoff.

The causal diagram shown in Figure 3 shows the evaporation (E) factor is a confounding variable, which creates issues for evaluating the statistical association between precipitation and runoff as treated in *Gupta et al. [2015]*.

## 4.0 References

Alila, Younes, Piotr K. Kuraś, Markus Schnorbus, and Robert Hudson. 2009. "Forests and Floods: A New Paradigm Sheds Light on Age-Old Controversies." *Water Resources Research* 45 (8): W08416.

Belmont, Patrick, John R. Stevens, Jonathan A. Czuba, Karthik Kumarasamy, and Sara A. Kelly. 2016. "Comment on 'Climate and Agricultural Land Use Change Impacts on Streamflow in the Upper Midwestern United States,' by Satish c. Gupta et Al." *Water Resources Research* 52 (9): 7523.

Bracken, L. J., J. Wainwright, G. A. Ali, D. Tetzlaff, M. W. Smith, S. M. Reaney, and A. G. Roy. 2013. "Concepts of Hydrological Connectivity: Research Approaches, Pathways and Future Agendas." *Earth-Science Reviews* 119: 17–34.

Cox, D. R. 2006. *Principles of Statistical Inference.* Cambridge, UK;New York; Cambridge University Press.

Feng, Changyong, Hongyue Wang, Naiji Lu, Tian Chen, Hua He, Ying Lu, and Xin M. Tu. 2014. "Log-Transformation and Its Implications for Data Analysis." *Shanghai Archives of Psychiatry* 26 (2): 105.

Foufoula-Georgiou, Efi, Patrick Belmont, Peter Wilcock, Karen Gran, Jacques C. Finlay, Praveen Kumar, Jonathan A. Czuba, Jon Schwenk, and Zeinab Takbiri. 2016. "Comment on 'Climate and Agricultural Land Use Change Impacts on Streamflow in the Upper Midwestern United States' by Satish c. Gupta et Al." *Water Resources Research* 52 (9): 7536.

Gleick, James. 1987. *Chaos: Making a New Science.* New York, N.Y., U.S.A: Viking.

Gupta, Satish C., Andrew C. Kessler, Melinda K. Brown, and Francis Zvomuya. 2015. "Climate and Agricultural Land Use Change Impacts on Streamflow in the

Upper Midwestern United States." *Water Resources Research* 51 (7): 5301–17.

Harding, Keith J., and Peter K. Snyder. 2012. "Modeling the Atmospheric Response to Irrigation in the Great Plains: Part Ii: The Precipitation of Irrigated Water and Changes in Precipitation Recycling." *Journal of Hydrometeorology* 13 (6): 1687–1703.

Hoaglin, David C., Frederick Mosteller, John W. Tukey, and Wiley Online Library. 1991. *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley.

Kędra, Mariola. 2014. "Deterministic Chaotic Dynamics of Raba River Flow (Polish Carpathian Mountains)." *Journal of Hydrology* 509: 474–503.

Koutsoyiannis, Demetris. 2016. "Generic and Parsimonious Stochastic Modelling for Hydrology and Beyond." *Hydrological Sciences Journal* 61 (2): 225–44.

Lakens, D. D. 2017. "Equivalence Tests : A Practical Primer for T Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science Social Psychological and Personality Science* 8 (4): 355–62.

Lindley, Dennis V. 2000. "The Philosophy of Statistics." *Journal of the Royal Statistical Society. Series D (the Statistician)* 49 (3): 293–337.

Lloyd, Elisabeth A., and Naomi Oreskes. 2018. "Climate Change Attribution: When Is It Appropriate to Accept New Methods?" *Earth's Future* 6 (3): 311–25.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 1st ed. Vol. 122. Boca Raton: CRC Press/Taylor & Francis Group.

Merz, B., S. Vorogushyn, S. Uhlemann, J. Delgado, and Y. Hundecha. 2012. "HESS Opinions 'More Efforts and Scientific Rigour Are Needed to Attribute Trends in Flood Time Series'." *Hydrology and Earth System Sciences* 16 (5): 1379–87.

Merz, Bruno, and Annegret H. Thieken. 2005. "Separating Natural and Epistemic Uncertainty in Flood Frequency Analysis." *Journal of Hydrology* 309 (1): 114–32.

Murtaugh, Paul A. 2002. "On Rejection Rates of Paired Intervention Analysis." *Ecology* 83 (6): 1752.

Pearl, Judea. 2009. *Causality*. New York: Cambridge University Press.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. First. New York, NY: Basic Books, Hachette Book Group.

Roderick, Michael L., Leon D. Rotstayn, Graham D. Farquhar, and Michael T. Hobbins. 2007. "On the Attribution of Changing Pan Evaporation." *Geophysical Research Letters* 34 (17): L17403.

Schottler, S.P., J. Ulrich, P. Belmont, R. Moore, J.W. Lauer, D.R. Engstrom, and J.E. Almendinger. 2014. “Twentieth Century Agricultural Drainage Creates More Erosive Rivers.” *Hydrological Processes* 28 (4): 1951–61.

Sivakumar, Bellie. 2017. *Chaos in Hydrology. Bridging Determinism and Stochasticity.* Dordrecht: Springer Netherlands.

Stanton, Maureen L., and Denise A. Thiede. 2005. “Statistical Convenience Vs Biological Insight: Consequences of Data Transformation for the Analysis of Fitness Variation in Heterogeneous Environments.” *The New Phytologist* 166 (1): 319–37.

Viglione, Alberto, Bruno Merz, Nguyen Viet Dung, Juraj Parajka, Thomas Nester, and Günter Blöschl. 2016. “Attribution of Regional Flood Changes Based on Scaling Fingerprints.” *Water Resources Research* 52 (7): 5322–40.

Weijs, S. V., N. C. Van de Giesen, and M. B. Parlange. 2013. “Data Compression to Define Information Content of Hydrological Time Series.” *Hydrology and Earth System Sciences, 17 (8), 2013* 17 (8): 3171–87.

Wilcox, Rand R. 1995. “ANOVA: A Paradigm for Low Power and Misleading Measures of Effect Size?” *Review of Educational Research* 65 (1): 51–77.

Zhang, Y. -., and K. E. Schilling. 2006. “Increasing Streamflow and Baseflow in Mississippi River Since the 1940 S: Effect of Land Use Change.” *Journal of Hydrology* 324 (1): 412–22.