International Journal of Forecasting 30 (2014) 944–962

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluating forecasts of political conflict dynamics

Patrick T. Brandt [a,*], John R. Freeman [b], Philip A. Schrodt [c]

[a] School of Economic, Political, and Policy Science, University of Texas, Dallas, United States
[b] Department of Political Science, University of Minnesota, United States
[c] Department of Political Science, Pennsylvania State University, United States

ARTICLE INFO

Keywords:
Conflict dynamics
Bayesian
Time series
Density evaluation
Verification rank histogram
Scoring rules

ABSTRACT

There is considerable interest today in the forecasting of conflict dynamics. Commonly, the root mean square error and other point metrics are used to evaluate the forecasts from such models. However, conflict processes are non-linear, so these point metrics often do not produce adequate evaluations of the calibration and sharpness of the forecast models. Forecast density evaluation improves the model evaluation. We review tools for density evaluation, including continuous rank probability scores, verification rank histograms, and sharpness plots. The usefulness of these tools for evaluating conflict forecasting models is explained. We illustrate this, first, in a comparison of several time series models' forecasts of simulated data from a Markov-switching process, and second, in a comparison of several models' abilities to forecast conflict dynamics in the Cross Straits. These applications show the pitfalls of relying on point metrics alone for evaluating the quality of conflict forecasting models. As in other fields, it is more useful to employ a suite of tools. A non-linear vector autoregressive model emerges as the model which is best able to forecast conflict dynamics between China and Taiwan.

1. Introduction

There is considerable interest today in forecasting international relations. Scholars are hard at work on attempts to produce early warnings of state failures, civil wars, and military conflicts between countries (Goldstone et al., 2010; Hegre, Karlsen, Nygård, Strand, & Urdal, 2012; O'Brien, 2010). Potentially, these forecasts could help policy makers to prepare for humanitarian crises. Some of these efforts are catalogued in Table 1 (Brandt, Freeman, & Schrodt, 2011).

There has been a lot of debate about how these forecasts should be evaluated. For example, Ward, Greenhill, and Bakke (2010) warn against forecasting using models that are based solely on in-sample tests of statistical significance. Instead, they recommend "out-of-sample heuristics" for evaluating models. Usually, when researchers conduct out-of-sample model evaluations, they employ point metrics like the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). However, these metrics are poor tools for evaluating the forecast accuracy; specifically, Armstrong and Collopy (1992) show that these metrics are unreliable and sensitive to outliers, and lack construct validity. More generally, RMSE and MAE contain little information about the multiple sources of forecast uncertainty (Tay & Wallis, 2000, p. 235).

Consider the genres in the lower two rows of Table 1. These studies aim to forecast conflict dynamics; that is,

* Corresponding author.
E-mail addresses: pbrandt@utdallas.edu (P.T. Brandt), freeman@polisci.umn.edu (J.R. Freeman), schrodt@psu.edu (P.A. Schrodt).

http://dx.doi.org/10.1016/j.ijforecast.2014.03.014
0169-2070/

Table 1
Some approaches to forecasting conflict in international relations.
Source: Brandt et al. (2011).

Type	Ex post	Ex post Counterfactual	Ex ante
One shot decision or event	Game theoretic/TOM (Brams & Togman, 2000)		Rational choice (Organski & Lust-Okar, 1997); game theoretic/EUA (de Mesquita, 2002, 2011); expert systems (Tetlock, 2005); role playing (Bennett & McQuade, 1996)
Probability of binary outcome	State failure project (Esty et al., 1998; King & Zeng, 2001); neural net models (Beck et al., 2000); integrated crisis early warning systems (ICEWS) (O'Brien, 2010)	Political instability task force	Prediction markets (Wolfers & Zitzewitz, 2004); predictions of civil war onset (Rost et al., 2009)
Conflict system: directed dyadic behavior	Cluster-density analysis (Schrodt & Gerner, 2000); BVAR (Brandt & Freeman, 2006); B-SVAR (Brandt et al., 2008)	BVAR (Brandt & Freeman, 2006)	International Crisis Group (ICG); Swiss Peace Foundation (FAST); VAR (Pevehouse & Goldstein, 1999)
Behavioral phase shifts	Crisis early warning systems (CEWS) (Alker et al., 2001; O'Brien, 2010); hidden Markov models (Schrodt, 2000)		

temporally disaggregated action–reaction sequences between two or more belligerents.[1] Stylized facts motivate much of this research. Scholars are convinced that conflicts move through phases *repeatedly*, and that belligerents exhibit different kinds of behaviors in each phase. Moreover, the sequence of phases through which a conflict moves as it escalates may differ in both kind and duration from the sequence in which it deescalates. Understanding these phase shifts is critical if we want to forecast conflict dynamics. Unfortunately, as Table 1 indicates, there are no statistical tests of these particular forecasts. Instead, these tests often amount to *ad hoc* narratives (Alker, Gurr, & Rupesinghe, 2001). Forecasts of systems of conflict analyze simultaneous exchanges between belligerents over time, for instance how Taiwanese behavior toward the Chinese is both a cause and a consequence of Chinese behavior toward the Taiwanese. Conflict system forecasts are often based on events data, and therefore make short-term predictions (Huth & Allee, 2002; Zeitzoff, 2011). Conflict system forecasters sometimes use reduced form multiequation time series models like vector autoregressions. In some cases, their forecasts are based on causality tests. One example is Pevehouse and Goldstein (1999), who produced forecasts of the consequences of NATO's bombing campaign on Serbian behavior. More recent research has produced multivariate forecasts of Israeli behavior toward the Palestinians and Palestinian behavior toward the Israelis (Brandt, Colaresi, & Freeman, 2008; Brandt & Freeman, 2006), but this work relies on the RMSE and MAE for evaluating the forecasts.

Taken together, the works in the bottom two rows of Table 1 suggest the need for a forecasting model that accounts for both *repeated* phase shifts and the possibility of different conflict dynamics within each phase. The possibility of repeated phase shifts makes this model inherently nonlinear. If such a model were constructed, how should it be evaluated? Are there other tools that could be used in addition to point metrics? According to research in meteorology, finance, and economics, the answer is *density evaluation*. Density evaluation has been shown to be useful for evaluating nonlinear forecasting models (Clements & Smith, 2000). To understand how it works, consider two analysts (models) A and B that produce the same point (mean) forecast and hence the same prediction error, but with different densities or coverages (i.e., different variances or longer tails in one of the densities). These forecasts are depicted in Fig. 1. If analyst A assigns a much higher probability to values of Y between Y_{t+1} and $\hat{Y}_{t+1}$ than B does, then using point metrics like RMSE will lead one to conclude that the two analysts perform equally well when, in actual fact, from a probabilistic standpoint, analyst A is a better forecaster than analyst B (since she assigns a higher probability to forecasts near the true value). To evaluate nonlinear models of conflict systems, we need tools that can take these differences in density coverage into account.

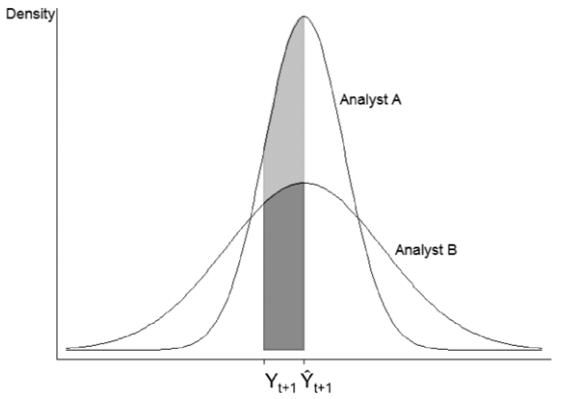

Fig. 1. Different model densities for a density comparison.

We show how to improve forecasts of conflict dynamics, and in particular, forecasts from the third and fourth

[1] The other studies in Table 1 aim to forecast single, one-off decisions of leaders (row 1) or the discrete onset of wars (row 2).

rows of Table 1.[2] We begin by briefly reviewing some of the theoretical motivation for the idea of conflict phase shifts, and the reason why a Markov-switching vector autogressive forecasting model captures it best. Next, we describe density evaluation. We explain the main criteria used in this kind of evaluation — calibration and sharpness — and describe some of the tools used in density evaluation: scoring rules, the probability integral transform (PIT), the verification rank histogram (VRH), and sharpness diagrams. We then use some familiar point metrics, together with these additional tools, to forecast conflict dynamics in both a stylized simulation and an actual international conflict. The former compares the performances of several Markov-switching vector autogressive, linear vector autoregressive, and bivariate normal forecasting models. It highlights the pitfalls of relying solely on conventional point forecasting metrics like RMSE for evaluating forecasting models of these kinds. The actual conflict we consider analyzes one of the world's flashpoints: the Cross-Straits conflict between Taiwan and China. Again, we stress the pitfalls of relying on point metrics like RMSE and MAE alone for evaluating forecasts of these and other conflict dynamics.

2. Motivation

2.1. Conflict phase shifts and nonlinear dynamics

The idea of conflict phase shifts is not new. It dates back at least to the work of Butterworth and Scranton (1976), and includes work like that of Bloomfield and Moulton (1989, 1997) (CASCON), and Sherman (1994) and Sherman and Neack (1993) (SHERFACS). Two more recent examples are Alker et al.'s (2001) six phase (CEWS) framework and Huth and Allee's (2002) four stage conflict model. The former yields "representations of multi-phased multi-perspective [conflict] histories" — catalogues of escalation and deescalation sequences. The defining features of the latter include the focus on the "repeated choices" made by belligerents and the fact that "disputes persist in time" (Huth & Allee, 2002, pp. 37, 42, 55). Because of this, Huth and Allee's framework is more sensitive to dynamics than Alker et al.'s. Interestingly, neither project analyzed the Cross-Straits case. Our investigation forecasts "repeated choices" of China and Taiwan.

As Diehl (2006) argues, there is little well-developed theory for conflict phase shifts.[3] The rational choice literature has suggested some possible theories. Wolford (2007) develops a game theoretic model of crisis bargaining in the shadow of leadership turnover. In his game, the equilibrium is semi-separating, implying switching between strategies where some protagonists accept their antagonists' initial proposals and earn reputations as doves, while other protagonists reject the initial proposal and earn reputations as hawks. In equilibrium, leaders lose office, so each new leader is tested. This leadership turnover produces

> "... 'cycles of conflict' in which new leaders are more prone to conflict than longer serving leaders.... Cycles of reputation building and conflict resulting from this turnover trap explain the occurrence of international conflict as a function of (1) a leader's time in office, (2) expectation over the behavior of potential successors, and (3) anticipated probabilities of political survival" (Wolford, 2007, pp. 779, 783).

Wolford's analysis implies two phases which repeat over time. This number of phases follows immediately from the structure of his game. Below, our simulation and the application to the Cross Straits includes an evaluation of a two phase model.

Another strand of rational choice theory stresses the effects of reciprocity. Building on the seminal work of Axelrod (1984) and research on human behavior in repeated non-zero-sum games, scholars find that cooperative behavior by one agent often begets cooperation by another. This cooperative behavior is rational if the shadow of the future (expected length of play) is sufficiently long. Huth and Allee (2002, 6ff.) associate tit-for-tat behavior with the norms-based account of conflict. Reciprocal strategies have been uncovered in (Cold War) superpower relations (Goldstein & Freeman, 1990), in Indian–Pakistani relations (Ward, 1982), in the Middle East (Goldstein, Pevehouse, Gerner, & Telhami, 2001), and in the Balkans (Goldstein & Pevehouse, 1997). Researchers have also found reciprocal behavior in subnational cases (Moore, 1998; Shellman, Reeves, & Stewart, 2007). Most of these investigations employ a one phase, frequentist vector autoregressive (VAR) model. We therefore also include this kind of model in both our simulation and our forecasting application for the Cross-Straits.

For conflict phase shifts, to the best of our knowledge, no studies to date have allowed for the possibility that reciprocity exists in one or a subset of conflict phases, or that leaders communicate their intentions effectively in one phase but not in another. Rational choice theories usually assume clear communication of the belligerents' preferences. For example, Wolford (2007) supposes that in phase two of his game, the antagonist learns the decision of the protagonist, and her type; thus, over time, international conflicts become games of complete information in which belligerents adopt a single, stationary strategy (Wolford, 2007, fn. 7). In fact, there is a considerable amount of evidence that belligerents switch strategies, and also that they often have trouble signaling their preferences. This is illustrated by studies of the Levant which reveal that, at times, Israelis and Palestinians use reciprocal strategies, and then switch to inverse (nonreciprocal) strategies.[4] As

[2] The works in Table 1 were reviewed by Brandt et al. (2011). Here, we suggest the addition of density evaluation — the use of a suite of tools that can gauge the forecast uncertainty — to the methods used in most of these genres. For a recent application of density evaluation to civil war forecasting, see Bagozzi (2014).

[3] Diehl (2006) reviews how work on selection effects, rational choice, path dependency, learning and issue linkage may be able to explain conflict phase shifts.

[4] For instance, Israeli leaders have complained explicitly about a lack of reciprocity on the part of Hamas and Islamic Jihad (Morris, 1999, p. 645); and Brandt et al. (2008, p. 360) find inverse rather than reciprocal behavior on the part of the Palestinians in the period April 1996–March 2005. Zeitzoff (2011) finds considerable evidence of "asymmetric response dynamics" during the Gaza conflict: Israel responded strongly in kind to conflict from Hamas, but Hamas did not exhibit any statistically significant responses to shocks in conflict from Israel.

regards communication, Ross (2000) and others stress the effectiveness of the signaling that went on between the governments in the 1995–1996 Cross-Strait crisis. Chinese officials communicated effectively (via U.S. officials several times in Beijing in both February and March 1996) that they did not intend to attack Taiwan (Tung, 2003, p. 157). However, these and other commentators also stress the mixed signals sent by President Lee in the 1990s about his commitment to Taiwanese independence.

Statistically, the possibility that belligerents use different decision rules in different conflict phases suggests, as was argued by Diehl (2006, p. 207), that conflict dynamics are nonlinear. Conflict processes switch between phases according to some matrix of transition probabilities that governs the sequencing structure of the dynamics. The data generating process for these conflict dynamics is therefore a mixture of distributions, a mixture representing belligerents' use, in each conflict phase, of possibly distinct decision rules (strategies). In turn, the switching process produces different patterns of escalation and deescalation over time, and different average durations in each phase. Brandt, Freeman, Lin, and Schrodt (2013) show that the Markov Switching vector autoregressive (MS-VAR) model is the most useful for capturing and forecasting these kinds of conflict dynamics. We provide more details about this model in the Appendix. We also include it in our simulated forecasting exercise and the application to the Cross Straits. Finally, the number of conflict phases is usually asserted by scholars, rather than established through statistical analysis. Not surprisingly, then, there is a tremendous amount of confusion about how many phases characterize conflicts like that in the Cross Straits. Analysts claim that the number of phases ranges from two to eight (e.g., Azar, 1972; Senese & Vasquez, 2008; Vasquez, Johnson, Jaffe, & Stamato, 1995). Here, we actually test for the number of phases in the Cross Straits, and find that a simple two phase model describes this case best.[5]

Evaluations of the forecasts from our MS-VAR and from the other models should take estimation uncertainty into account. The next section shows how this can be done.

2.2. Evaluating probabilistic forecasts: calibration and sharpness

The case for probabilistic forecasting is made by Dawid (1984). The idea is to use either an assessor (a judgmental forecast) or a model to produce a probability density function for future quantities or events. Thus, a density or probabilistic forecast is "a complete description of the uncertainty associated with a prediction, and stands in contrast to a point forecast, which by itself, contains no description of the associated uncertainty" (Tay & Wallis, 2000, p. 235).

There are at least two major reasons for the use of probabilistic forecasts. First, decision theory shows that, in general, it is impossible to rank two incorrect forecast densities for two forecast users; if the forecast density corresponding to the true data generating process can be found, all forecast users will prefer that density regardless of their loss functions. Diebold, Gunther, and Tay (1998) demonstrate the advantages of finding a true forecast density, by comparing the action choice for what is believed to be the correct density, $p(y)$, for a variable y_t with realizations $\{y_t\}_{t=1}^{m}$, with the action choice for the actual true data generating process, $f(y)$. They show that, if two decision makers, i and j, have different loss functions, and neither of their forecast densities, $p_i(y)$ and $p_j(y)$, is correct, it is not possible to rank these densities. One decision maker might prefer to use her $p_j(y)$, while the other decision maker might prefer to use another $p_k(y)$ as his density. [6]

Second, because they can evaluate non-Gaussian predictive densities and mixtures of Gaussian densities, probabilistic forecasts help analysts to discriminate between linear and non-linear models. Clements and Smith (2000) explain why point metrics usually fail to show non-linear models as being superior for forecasting relative to linear models, and show how density forecast evaluation methods can be used to compare the performances of linear and nonlinear models (AR and SETAR), and linear and nonlinear multivariate models (VAR and non-linear VAR). As we will show, this is important in trying to forecast conflict dynamics that exhibit phase shifts and other kinds of non-linearity.

The main evaluation criteria for probabilistic forecasts are calibration and sharpness. *Calibration* is the statistical consistency between the distributional forecasts and observations. *Sharpness* has to do with the "concentration of [the] predictive distribution and is a property only of the forecasts" (Gneiting & Raftery, 2007, p. 359). The goal of probabilistic forecasting is to achieve a high degree of sharpness, subject to calibration.[7]

Among the tools used in these evaluations are scoring rules. These rules assign numbers representing the degree of association between the predictive distributions and observed events. The scores are used to rank the levels of success of forecasters and their models. In this way, scoring rules are an integral part of what meteorologists call forecast verification. A scoring rule is proper if it gives an assessor an incentive to reveal her true probability function rather than to hedge (e.g., supply equal probabilities

[5] Based on the international relations literature, Brandt et al. (2013) derive several propositions about when conflict phase shifts should occur, then use an MS-BVAR model to test them. For example, from the work on institutional determinants of international conflict, they derive the proposition that phase shifts should occur when there is a change to or from the provision for "irregular" (e.g., coups) to "regular" (e.g., electoral) removal of leaders (Chiozza & Goemans, 2011), and when an authoritarian government changes between military and partisan types (Lai & Slater, 2006; Weeks, 2012).

[6] Diebold and Lopez (1996) provide an overview of the application of loss functions to forecasting, including the idea of loss function differentials being related to forecasting models. Tay and Wallis (2000) explain that different loss functions may lead to different optimal point forecasts if the true density is asymmetric.

[7] See also Raftery, Gneiting, Balabdaoui, and Polakowski (2005). Hamill (2001, p. 551–552) equates the concepts of calibration and reliability. Gneiting, Balabdaoui, and Raftery (2007) develop three concepts of calibration: probabilistic, exceedence and marginal. We focus on the first of these here.

Table 2
Four well-known (proper) scoring rules for discrete random variable forecasts.

Rule	Form	Range
Quadratic $Q(r, d)$	$1 - \sum_i (r_i - d_i)^2$	$[-1, 1]$
Brier $PS(r, d)$	$1 - Q(r, d)$	$[0, 2]$
Spherical $S(r, d)$	$\frac{\sum_i r_i d_i}{(\sum_i r_i^2)^{\frac{1}{2}}}$	$[0, 1]$
Logarithmic $L(r, d)$	$ln(\sum_i d_i r_i)$	$(-\infty, 0]$

for each event). Model fitting can be interpreted as the application of optimum score estimators, of which a special case is maximum likelihood.[8]

There are several well-known scoring rules for discrete random variables.[9] Consider a variable that produces n discrete and mutually exclusive events, $E_1, \ldots, E_n$. Say that, at time t, an assessor supplies a probabilistic forecast about the values that the variable will assume at time $t + 1$. This probabilistic forecast is a vector $r = (r_1, \ldots, r_n)$, where r_i is her elicited probability that event i will occur. Let her *true* assessment be represented by the vector $p = (p_1, \ldots, p_n)$. Finally, let the row vector $d = (d_1, \ldots, d_n)$ denote the actual observation at $t + 1$, so that, if event i is realized, $d_i = 1$ and $d_j = 0$ for $j \neq i$. Then, an assessor is considered normatively perfect if her probability vector, r, is "coherent" – satisfies the laws of probability – and her vector corresponds completely to her true beliefs ($r = p$). A proper scoring rule is one that makes it rational for this last condition to be satisfied ($r = p$), or one for which reporting p maximizes the assessor's expected score or utility.

Examples of such proper binary scoring rules are described in Table 2. The logarithmic scoring rule was suggested by Good (1952), and is sometimes called an ignorance score. It is also a local rule, insofar as its value depends only on the probability assigned to the outcome that actually is observed, not on the probabilities assigned to outcomes that are not observed.[10]

As an illustration, consider the event that precipitated the Gulf War, namely Iraq's invasion of Kuwait. Suppose that there were only three possibilities: ground invasion, air attack, and no invasion. Furthermore, suppose that, through an elicitation tool, forecasters A and B supplied the vectors (0.35, 0.60, 0.05) and (0.30, 0.35, 0.35), respectively. Iraq launched a ground invasion of Kuwait, so event E_1 was realized. Table 3 shows that forecaster B would have received a higher score according to the quadratic and spherical scoring rules, whereas forecaster A would have received a higher score according to the logarithmic score. Because the interpretation of the Brier score is the reverse of the quadratic score, it would also rank forecaster B as more proficient than forecaster A. These rankings reflect the fact that the logarithmic scoring rule only considers the assessed probability of the event that actually occurred, whereas the other rules consider all of the assessed probabilities relative to this realization.[11]

Table 3
Illustrative scores for two hypothetical forecasters for Iraq's behavior in 1991. Forecaster A: (0.35, 0.60, 0.05); Forecaster B: (0.30, 0.35, 0.35).

Forecaster	$Q(r, d)$	Brier PS	$S(r, d)$	$L(r, d)$
A	0.215	0.785	0.503	−1.050
B	0.265	0.735	0.518	−1.204

When evaluating forecasts of ordered categories of a discrete random variable, the Ranked Probability Score (RPS) is often used. The RPS includes an evaluation of the "distance" between the realization and the relative probabilities assigned by a forecaster to the different (ordered) categories of events. Suppose that there are k such categories, and that the assessor's forecast is the row vector $(p_1, \ldots, p_k)$. Then, a proper scoring rule for when each observation j actually occurs is:

$$S_j = \frac{3}{2} - \frac{1}{2(k-1)} \sum_{i=1}^{k-1} \left[\left(\sum_{n=1}^{i} p_n \right)^2 + \left(\sum_{n=i+1}^{k} p_n \right)^2 \right] - \frac{1}{k-1} \sum_{i=1}^{k} |i - j| p_i, \quad (2)$$

where S_j ranges from 0 for the worse possible forecast to 1 for the best forecast (Epstein, 1969).

Say that a particular intra-state conflict moves around among four conflict phases,[12] and call the count of conflict-related events at time t, C_t. Assume that the phases are defined by an ordered set of categories of such counts. Specifically, let phase 1 be values in the range C_t = 0–10; phase 2: C_t = 11–20; phase 3: C_t = 20–30; and phase 4: $C_t > 40$. When asked what phase the conflict will be in at time $(t + 1)$, analyst A provides the forecast (0.1, 0.3, 0.5, 0.1), while analyst B provides the forecast (0.5, 0.3, 0.1, 0.1). According to this rule, the scores for the two analysts are reported in Table 4 for each observed category. If the highest phase of the conflict is realized, analyst A scores 0.67, while his counterpart scores 0.43. This is because analyst A assigns a higher overall probability to phases 3 and 4 than analyst B. Conversely, if phase 1 or 2 is realized, analyst B would perform better, since he assigns a higher probability to these phases than analyst A.

Suppose that the analyst or model produces a predictive density function for future values of a variable.

[8] Gneiting and Raftery (2007) provide a theory of proper scoring rules for general probability spaces. They explain the relationships between these rules and information measures and entropy functions. For estimation, they explain how proper scoring rules suggest useful loss functions from which optimum scoring estimators can be derived, and demonstrate the link between proper scoring rules and Bayesian decision analysis.

[9] See Winkler and Murphy (1968, pp. 753–755) and Gneiting and Raftery (2007, Section 3).

[10] The quadratic scoring rule, $Q(r, d)$ in Table 2, is proper (Winkler & Murphy, 1968, p. 754), since:

$$E(Q) = \sum_j p_j Q_j(r, d) = \sum_j p_j^2 - \sum_j (r_j - p_j)^2, \quad (1)$$

which is maximized when $r = p$. The spherical and logarithmic rules are also proper.

[11] When the quadratic score implies the best performance, $Q(r, d) = 1$, $PS = 1 - 1 = 0$. Conversely, when the $Q(r, d) = -1$, Brier $PS = 1 - (-1) = 2$.

[12] The following example is based on the weather illustration of Epstein (1969).

Table 4
Rank probability scores for conflict phase forecasts by two hypothetical analysts.

Observed category	Analyst *A* (0.1, 0.3, 0.5, 0.1)	Analyst *B* (0.5, 0.3, 0.1, 0.1)
1	0.61	0.90
2	0.87	0.90
3	0.94	0.70
4	0.67	0.43

Predictive densities can be obtained from models either by making distributional assumptions about the estimation uncertainty, or, as has become increasingly common, by means of computation (Brandt & Freeman, 2006, 2009). In this case, evaluative tools vary depending on whether they are based on binary scoring rules which are defined on the probability space (unit interval), or on payoff functions that are defined on the space of values of the variable of interest (real line).

A common scoring rule of this type is the Continuous Rank Probability Score (CRPS). The CRPS is defined as follows (Hersbach, 2000, pp. 560–561): let the forecast variable of interest be denoted by x, the observed value of the variable by x_a, and the analyst's (or model's) pdf by $\rho(x)$. The CRPS is

$$\mathrm{CRPS}(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx, \tag{3}$$

where P and P_a are the cumulative distributions:

$$P(x) = \int_{-\infty}^{x} \rho(y)dy \tag{4}$$

$$P_a(x) = H(x - x_a). \tag{5}$$

Here, H is the Heaviside function: $H(x-x_a) = 0$ if $(x-x_a) < 0$ and $H(x - x_a) = 1$ if $(x - x_a) \geq 0$. The CRPS is the difference between the total areas of the predicted and observed cumulative distributions, with lower values indicating better performances, as the forecast and observed densities are more closely matched, since this measure sums the probabilities of false positive and negative cases.

The CRPS is measured in units of the forecasted variable for each forecast point. In application, an average of the score is often calculated over this set of forecasts or grid points, k:

$$\overline{\mathrm{CRPS}} = \sum_k w_k \mathrm{CRPS}(P_k, x_a^k) \tag{6}$$

where w_k are weights set by the forecaster (typically, $w_k = \frac{1}{k}$).

The CRPS assesses both calibration and sharpness. The attractive properties of the CRPS are that it is sensitive to the entire range of x, it is defined in terms of the predictive cumulative density rather than the predictive marginal density, and it is readily interpretable as a integral over all possible Brier scores; see Gneiting and Raftery (2007, Section 4.2) for computational details.[13]

Verification rank histograms (VRH) are another tool which is used to assess calibration. The VRH for each forecast gives the *rank* of the observed value, which is tallied relative to the sorted (ranked) ensemble forecasts. Then, the population rank j is the fraction of times that the observed value, when compared to the ranked ensemble values, is between ensemble members $j - 1$ and j. Formally, this rank $r_j = \overline{P(x_{j-1} \leq V < x_j)}$, where V is the observed value, x_j is a sorted ensemble forecast of the indicated rank, j, and P is the probability. If the ensemble distribution is calibrated, these ranks produce a uniform histogram. The VRH can be expressed in terms of either relative frequencies or a continuous analogue, a density histogram. Gneiting et al. (2007, p. 252) call the VRH the "cornerstone" of forecast evaluation.

A related concept is the probability integral transform (PIT). The PIT is defined in terms of realizations of time series and their one-step-ahead forecasts. Let $\{y_t\}_{t=1}^m$ be a series of m realizations from the series of conditional densities $\{f(y_t|\Omega_{t-1})\}_{t=1}^m$, where Ω_{t-1} is the information set. If a series of one-step-ahead density forecasts, $\{p_{t-1}(y_t)\}_{t=1}^m$, coincides with $\{f(y_t|\Omega_{t-1})\}_{t=1}^m$, the series of PITs of $\{y_t\}_{t=1}^m$ with respect to $\{p_{t-1}(y_t)\}_{t=1}^m$ is i.i.d. $U(0, 1)$, or

$$\{z_t\}_{t=1}^m = \left\{ \int_{-\infty}^{y_t} p_{t-1}(u)du \right\}_{t=1}^m \sim U(0, 1). \tag{7}$$

For a collection of forecasts and series of observed values of a variable, PIT values can be calculated for a set of forecasts, after which these values can be tested for uniformity.[14]

The VRH of the PIT values gives some indication of the veracity of the predictive distribution. A U-shaped VRH indicates that the forecasting model is underdispersed, while a hump-shaped VRH indicates the forecasting model is overdispersed. Sometimes a cumulative density function (cdf) plot is used instead of the VRH. If the VRH is uniform, its cdf ought to be approximately a 45° line. A χ^2 test can be used to assess uniformity, and there are also various other tests available for this purpose, such as the Kolmogorov–Smirnov test (Diebold et al., 1998; Tay & Wallis, 2000). For time series data, it is important, before testing for uniformity, to establish (using a correlogram) that the PITs are i.i.d. (Clements & Hendry, 1998; Diebold et al., 1998, 1999; Gneiting et al., 2007). Hamill (2001) shows that the VRH can be flat, in spite of the fact that the forecasts suffer from conditional bias. Sampling from the tails of the distributions, from different regimes, and across

[13] The properties of the CRPS and the way in which it can be decomposed into reliability, uncertainty, and resolution components are discussed by Hersbach (2000). Gneiting and Raftery (2007) develop the decision theory for this and the other probability scores.

[14] For details of this definition, see Diebold et al. (1998); Diebold, Hahn, and Tay (1999). An explanation of the relationship between the PIT and Dawid's prequential principle is provided by Gneiting et al. (2007, p. 244).

space without accounting for the covariance at grid points can all produce mistaken inferences from the shape of the VRH. Hamill recommends forming the VRH from samples which are separated in both space and time.[15]

From Hamill's critique of the VRH and an examination of spread-error plots for some sample data, Raftery et al. (2005) and Gneiting et al. (2007) recommend that an assessment of sharpness be included in the evaluation of probabilistic forecasts. Sharpness has to do with the concentration of the predictive distribution. To assess this feature of the density forecast, Gneiting et al. (2007) use box plots for the central prediction intervals for competing models; on average, the shorter the central prediction interval for a model, the sharper its forecast.

One problem with the box plot is that it does not convey any information about the shape of the forecast density; it only depicts the width of a given credible interval. Another tool for gauging sharpness, the *violin plot*, addresses this issue. The violin plot is so shaped as to reflect the forecast density. More concentrated densities have violin shapes or less dispersion.

Probabilistic forecasting became prevalent in the late 1990s.[16] It is now at the heart of forecast verification in meteorology. Meteorologists use a suite of the tools described above to evaluate such forecasts. For instance, Gneiting, Larson, Westrick, Genton, and Aldrich (2006) use the CRPS, PIT and RMSE to evaluate the RST model. Gneiting et al. (2007) use a combination of calibration tests (PIT) and box plots, along with MAEs, log scores, and the CRPS, to rank three algorithms for forecasting windspeed.[17] Probabilistic forecasting is employed in finance for studying high frequency exchange rate series (Diebold et al., 1998) and stock returns (Weigend & Shi, 2000), and for evaluating portfolios (Tay & Wallis, 2000).

3. Illustrations: forecasting conflict dynamics

3.1. A stylized simulation

Forecasters in other disciplines often illustrate the value of density evaluation using a single, standard normal DGP. The forecasting models are assumed to be of the form $N(\mu, \sigma^2)$, and the VRH is studied for alternative (incorrect) values of μ and σ^2. However, theoretical and empirical investigations in political science suggest that, for conflict dynamics, the true DGP is a finite mixture distribution. The DGP for conflict dynamics is a nonlinear simultaneous equation system, a collection of equation systems between which conflicts move repeatedly over time. Thus, for students of conflict dynamics, a Markov switching vector autoregressive process, MS-VAR, is a more meaningful DGP. This model includes a separate equation for the actions and reactions of each belligerent, and allows for repeated phase shifts or changes in the parameters in each equation system across a set of regimes, $h = \{1, 2, \ldots\}$.

It is an evaluation of models which could be used to forecast data from the kind of non-linear time series process on which we focus. Our illustration demonstrates the pitfalls of using point metrics alone when the DGP is an MS-VAR.[18] In Appendices A and B we also show the usefulness of incorporating density evaluation when the DGP is a linear VAR.

Our simulated data are from a model with two equations and two regimes — a stylized version of peace v. war between two belligerents. Suppose that we have continuous measures of the levels of directed behavior between two belligerents, where positive values denote cooperation and negative values denote conflict. This simple two-equation dynamic MS-VAR model allows the net directed behavior between belligerents to evolve endogenously, and also allows the parameters governing the conflict dynamics to change in each regime.

The MS-VAR DGP processes is

$$\text{Regime 1:}\quad \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0.7 & 0.7 \\ 0.1 & 0.1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

$$\text{Regime 2:}\quad \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 0.2 & 0.2 \\ 0.1 & 0.1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}.$$

The error processes for each regime are the same, and are given by

$$\begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \right).$$

Finally, the Markov-transition matrix across regimes 1 and 2 is

$$Q = \begin{pmatrix} 0.8 & 0.2 \\ 0.05 & 0.95 \end{pmatrix},$$

where the (1,1) element gives the probability of remaining in the first regime; the (1,2) element gives the probability of changing from regime 1 to regime 2, etc. This is a first order MS process, where each simultaneous equation system is a VAR(1). Regime 1 is the main dynamic driver; it is, in effect, the conflict phase. Regime 2 is the non-conflict phase for the belligerents.

[15] The extension of the PIT to multivariate and multi-step analysis involves the decomposition of each period's forecast into its conditionals (Clements & Smith, 2000; Diebold et al., 1999).

[16] Fildes and Stekler's (2002) claim that probabilistic forecasting has not caught on in the natural and social sciences is simply inaccurate. See Gneiting and Raftery (2005) in *Science*, and the introduction to the special issue of the *Journal of Forecasting* by Timmerman (2000).

[17] Information about forecast verification at the University of Washington research center can be found at http://isis.apl.washington.edu/bma/index.jsp and http://probcast.washington.edu.

[18] Hamill (2001, pp. 553–555) considers a case where the DGP is a regime switching model, and the way in which this can confound inference in the VRH. The idea of regime switching is now common in meteorology; see for instance Gneiting et al. (2006), who incorporated switches in the wind direction (attributed to pressure differences over sea and land plus the topography of the Columbia river basin) into forecasting models.

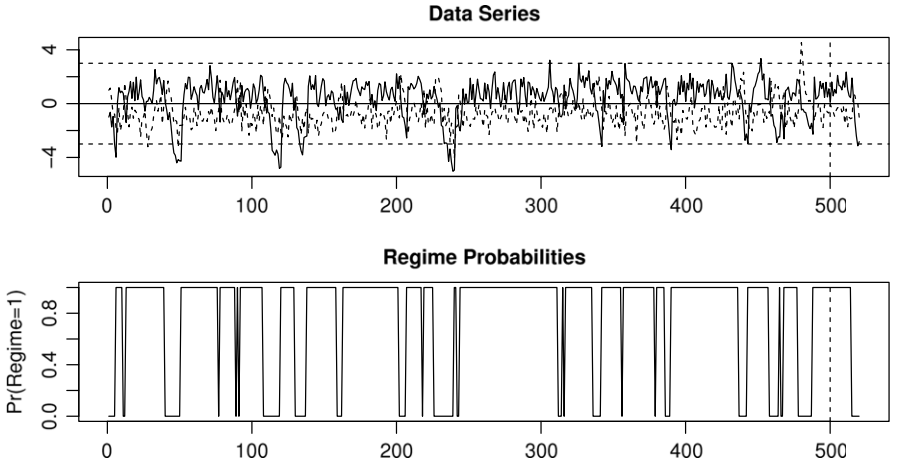

Fig. 2. Simulated data from a stylized conflict DGP. The solid and dashed lines in the top plot are the bivariate series of interest. The horizontal dashed lines show the two-standard-deviation extremes. The vertical dashed lines denote the split between the sample and the forecasting horizon.

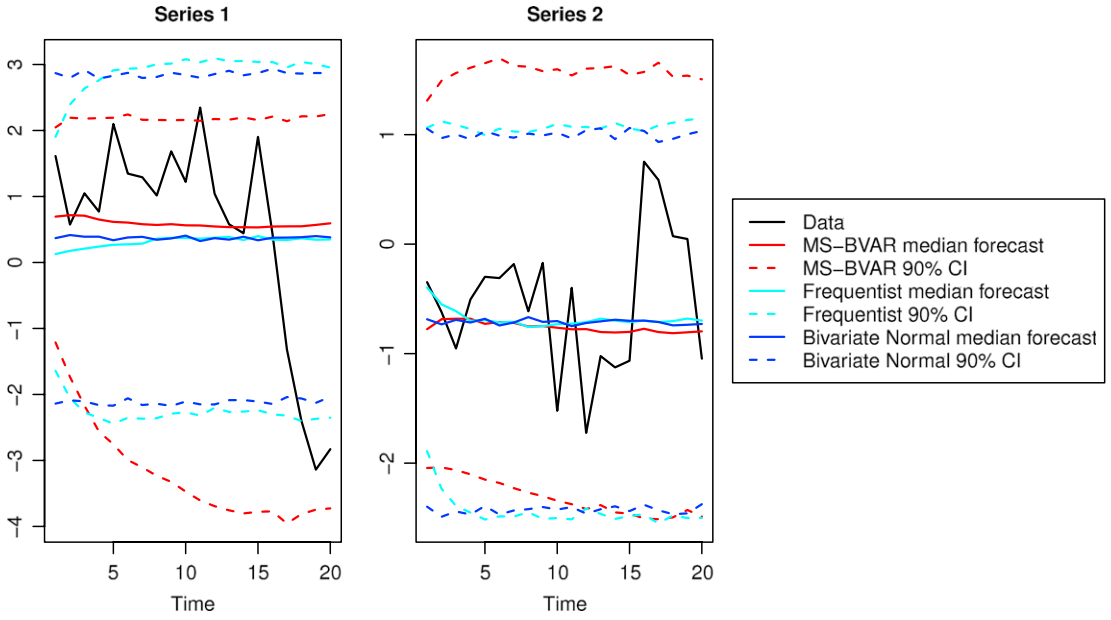

Fig. 3. Forecasts and 90% credible intervals from competing models. Solid lines are median predictions and dashed lines give the coverage intervals.

We simulated 520 bivariate observations from this process — observations of (y_1, y_2), corresponding to the behaviors of two belligerents across two regimes. We used the first 500 observations to estimate competing forecasting models, and the remaining 20 bivariate observations were used to evaluate the forecasts. Fig. 2 depicts the simulated data for the two series, along with the true probabilities that the process at each point in time is in regime 1.

We then ran a "horse race" between three forecasting models: (1) an MS-BVAR with one lag on both variables, two regimes, and a diffuse prior; (2) a frequentist VAR with one lag; and (3) a bivariate normal distribution based on sample moments of the observed (simulated) series. The MS-BVAR is intended to be the most accurate model of the DGP. The frequentist VAR captures the simultaneous dynamics in the system but omits the phase shifts. The bivariate normal model ignores both the phase shifts and the simultaneous dynamics.

Fig. 3 shows the three forecasts for the two series at time points 501–520. The error bands are 90% credible (confidence) intervals, generated by producing 5000 forecasts from each model. For each forecasting model, the credible intervals were calculated from the quantiles of the 5000 forecasts at each time point between 501 and 520. Note that the MS-BVAR model does a better job of forecasting the downturn in the y_1 series from time points 515 to 520.

Table 5 reports the results for two common point metrics, RMSE and MAE, and two probabilistic scoring rules,

Table 5
Point metrics and scoring rules for competing forecasting models.

	Series 1			Series 2		
	MSBVAR	VAR	Normal	MSBVAR	VAR	Normal
RMSE	2.02	2.06	2.05	1.27	1.21	1.19
MAE	1.72	1.78	1.74	1.04	1.00	0.98
CRPS	0.87	0.91	0.89	0.40	0.39	0.64
IGN	1.81	1.89	1.87	1.23	1.16	1.59

the CRPS and the Ignorance Score (IGN).[19] As expected, the MS-BVAR model comes out the "winner" in all but one case. For series y_2, the RMSE and MAE do not discriminate as well as the CRPS and the IGN. In fact, according to the RMSE criterion, the VAR and bivariate normal models do nearly as well as the MS-BVAR model. Similar results are obtained for the MAE criterion. Even for the CRPSs, the MS-BVAR model is only slightly preferred as a forecasting model for the second series. The main conclusion here, though, is that using the point metrics alone provides an incomplete picture of performance, since the CRPS and IGN metrics select the MSBVAR forecasts over the alternative models.

The PITs for the forecasts are reported in Fig. 4. For each variable and forecasting model, the PITs and VRHs were computed from the empirically estimated forecast densities for each series over the 20 periods. The p-value for a Kolmogorov–Smirnov statistic for the null of a uniform distribution is reported at the bottom of each PIT histogram. These PIT and VRH results allow for an evaluation of the calibration of the forecasts from each model. They provide more evidence in favor of the MS-BVAR model.

For both y_1 and y_2, the PITs for the MS-BVAR model are much more uniform in shape than those for the VAR or bivariate normal forecasts. For y_1, the forecasts from the VAR and Normal model have PITs that deviate from uniformity. For both series' forecasts, the MS-BVAR model produces PITs which show little evidence against uniformity, based on the reported p-values. The other models' PITs are not likely to be uniform for the y_1 forecasts — a point suggested by the results in Fig. 3.

The VRHs for the MS-BVAR model (not reported here, but available on request) are also more spread out, being akin to a uniform density. This uniformity is suggested by the reported Kolmogorov–Smirnov (KS) test statistics. For the MS-BVAR model, the KS tests have p-values greater than 0.12 for both variables, but for the VAR and Normal forecasts, the KS p-values are less than 0.03 for y_1 and less than 0.18 for y_2. The VRHs for y_1 have poor coverages over the range of the forecast density produced by the VAR and Normal models. This is the cause of the histograms having few forecasts (nearly 1500 of 5000) that do not rank below the observed values, again confirming what is seen in the coverage intervals in Fig. 3.

In this stylized simulation of conflict dynamics, the probabilistic scoring rules and density evaluation do a better job of picking the winner than the more familiar point metrics. Based on the point metrics in Table 5, the RMSE rankings of the models are inconsistent, preferring the MS-BVAR model for y_1, but not for y_2. A similar confusion is seen with the other metrics, though the CRPS gives a slight edge to selecting the MS-BVAR forecasting model. However, the PIT and VRH results provide consistent evidence in favor of the (correct) MS-BVAR model that accounts for the phase shifts in the forecast horizon.

3.2. Conflict forecasting example

We now use our suite of evaluation tools to learn which of several forecasting models do the best job of predicting an important substantive case, the Cross-Straits conflict between China and Taiwan. We analyze this conflict over the period 1 January 1998 to 31 December 2006. This period witnessed a relatively high level of conflict between the countries and several changes in political leadership, changes that could cause phase shifts.[20] The data come from the Event Data Project (EDP) at Penn State (http://eventdata.psu.edu). There were 366,482 machine coded events in this period. These are coded using the CAMEO event data coding format (Gerner, Schrodt, & Yilmaz, 2009), which classifies the events in a dyadic relationship (Chinese events toward Taiwan, etc.), as well as into material and verbal conflict/cooperation. The data are first aggregated into monthly time series for the numbers of events directed by China toward Taiwan and by Taiwan toward China according to the following categories:

- *Verbal cooperation*: The occurrence of dialogue-based meetings (e.g., negotiations, peace talks), or statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors.
- *Material cooperation*: Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc.
- *Verbal Conflict*: A spoken criticism, threat, or accusation, often related to potential acts of material conflict.
- *Material Conflict*: Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc.

We subtract the number of material and verbal conflict events from the respective number of cooperative events, so as to construct four time series which will summarize the conflict dynamics. Under this scaling, positive values indicate a net cooperation and negative values indicate a net conflict. The data series are Chinese net material

[19] We use a independent normal density for the computation of the CRPS and IGN. The IGN is the negative of the log scoring rule. The smaller the IGN, the better the probabilistic fit of the forecasting model.

[20] By "China" we mean the People's Republic of China and its affiliated actors. "Taiwan" refers to The Republic of China and its affiliated actors.

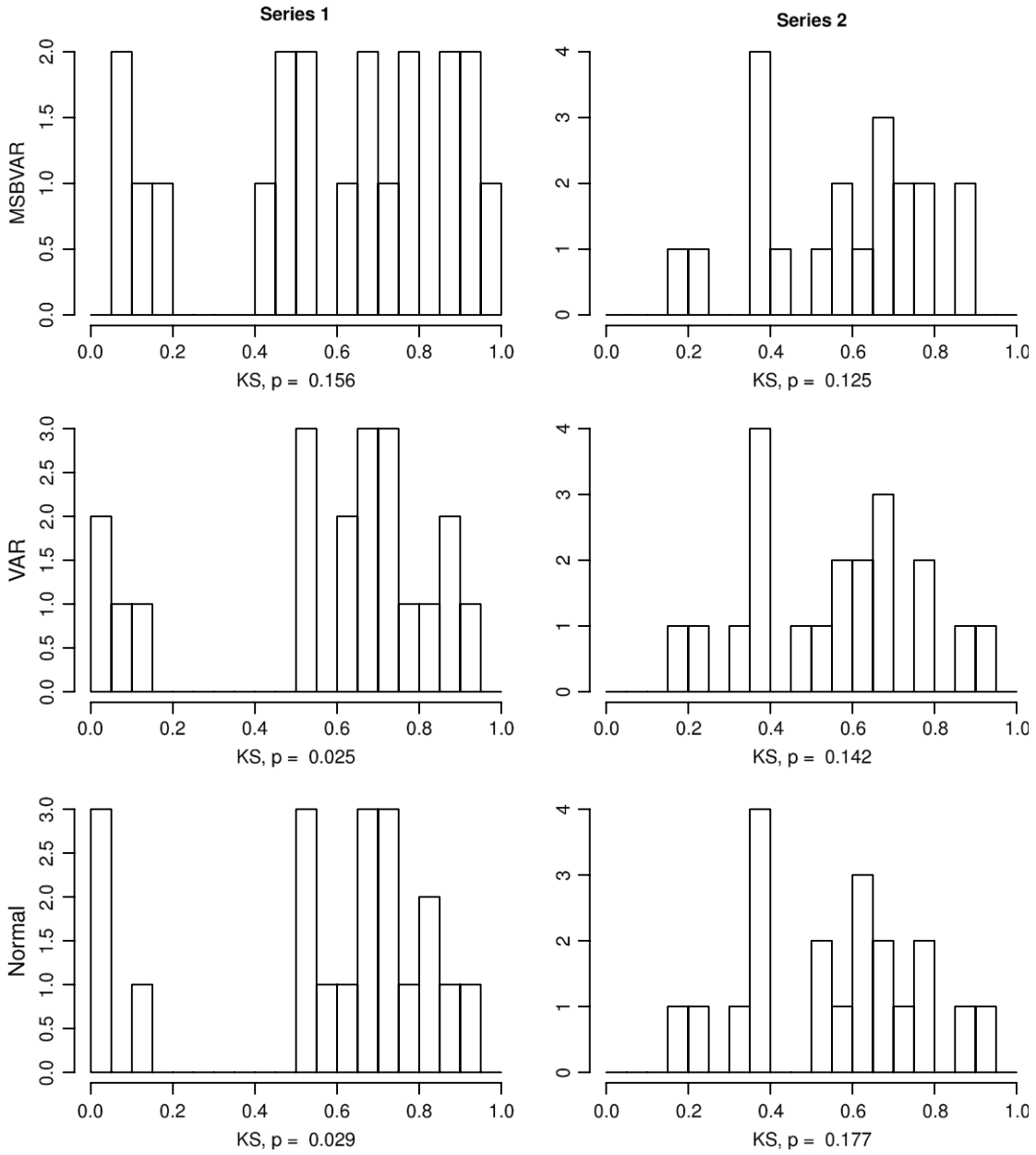

Fig. 4. PITs for competing forecasting models, stylized simulation.

actions toward Taiwan (C2T-M), Chinese net verbal actions toward Taiwan (C2T-V), Taiwan material actions toward China (T2C-M), and Taiwan verbal actions toward China (T2C-V).

Data from January 1998 to December 2005 ($T = 96$) are used to produce 6- and 12-month *ex post* forecasts for four forecasting models. The first is an MS-BVAR model with two regimes and one lag (meaning that it is the same as in the earlier example).[21] The second is a Bayesian VAR with an informed prior (see Brandt & Freeman, 2006). The informed prior centers on a random walk model and is parameterized with a Sims–Zha prior (Sims & Zha, 1998) based on the in-sample performance (i.e., data prior to December 2005). The third model is a flat or diffuse prior BVAR. For both VAR models, a lag length of 6 is chosen. The final forecasting model is a set of independent, univariate autoregressive (AR) models for each of the four data series. In these univariate models, the lag length for the AR model for each series is selected by the AIC.

[21] Other choices of the lag length (1–3 lags) and number of regimes (2 and 3) were evaluated. Based on the log marginal data density values, this specification has the best fit, according to the model selection criteria for Markov-switching models discussed by Frühwirth-Schnatter (2006). Adding additional lags or regimes worsened the MS-BVAR in-sample fit.

Table 6
Performance of the three Cross-Straits forecasting models over forecasting horizon k. The forecast evaluations are based on an ensemble of 5000 forecasts for each model and each equation, using the average RMSE, average MAE, and CRPS values. The CRPS was computed under a normal density assumption. The entry in bold for each forecast horizon, variable, and metric is considered the "winner".

k	Model	RMSE			
		C2T-M	T2C-M	C2T-V	T2C-V
6	MS-BVAR	6.76	5.44	50.03	57.72
6	BVAR informed	10.17	13.02	67.52	56.61
6	BVAR flat	9.67	15.26	72.84	74.48
6	Univariate	**6.47**	**5.09**	**47.59**	**56.41**
		MAE			
6	MS-BVAR	5.39	4.30	**37.72**	**42.95**
6	BVAR informed	7.75	9.73	52.75	43.36
6	BVAR flat	7.48	11.67	56.23	55.52
6	Univariate	**5.21**	**4.08**	38.33	45.08
		CRPS			
6	MS-BVAR	**2.06**	1.79	**17.31**	**17.04**
6	BVAR informed	2.48	2.84	19.70	17.13
6	BVAR flat	2.51	3.49	21.37	20.67
6	Univariate	2.19	**1.61**	19.09	20.75
		RMSE			
12	MS-BVAR	6.62	5.46	48.11	58.48
12	BVAR informed	11.30	14.14	68.87	61.18
12	BVAR flat	10.69	16.26	75.49	79.29
12	Univariate	**6.35**	**5.39**	**44.79**	**56.32**
		MAE			
12	MS-BVAR	5.25	4.33	36.49	**44.29**
12	BVAR informed	8.58	10.67	54.03	47.28
12	BVAR flat	8.26	12.48	58.80	60.80
12	Univariate	**5.09**	**4.31**	**35.90**	45.37
		CRPS			
12	MS-BVAR	**1.85**	1.82	**13.87**	**15.04**
12	BVAR informed	2.58	3.15	18.07	16.33
12	BVAR flat	2.56	3.66	20.28	21.31
12	Univariate	2.05	**1.76**	15.77	20.86

For each of the forecasting models, an ensemble of 5000 unconditional draws was generated in order to summarize the forecast density (after a burn-in period of 1000 forecasts) over the 6- and 12-period horizons. The average RMSE, MAE, and CRPS values for each equation (variable), forecast model, and forecast horizon, relative to the true *ex post* realizations of the data, are reported in Table 6. The entries in bold are those that are considered the "winner" for each criterion.

Several of the pitfalls commonly encountered when evaluating and comparing forecasts are evident in this table. First, the oft-heard time series forecasting mantra that univariate forecasts are superior to multivariate (system) forecasts is seen in the 6- and 12-period forecast comparisons. The RMSEs for both forecast horizons are smaller for the univariate AR models than for the other forecast models. However, the MS-BVAR RMSEs are only slightly larger than the univariate AR ones. The MAE metric results are split between preferring the MS-BVAR and the univariate AR models, depending on the data series. This is because these pooled univariate model forecasts are overconfident and do not do a good job predicting events in the tails of the forecast densities. Despite these results, the MS-BVAR forecasts have RMSEs and MAEs that are roughly on a par with those of the univariate models, and better than the same metrics for the BVAR models.

The CRPSs reported for each variable in Table 6 indicate the superior fit of the forecast densities for the MS-BVAR model for all but the T2C-M series (and even here, the CRPSs are very close for the MS-BVAR and univariate AR models). Recall that this measure evaluates both the calibration and sharpness, and summarizes the deviations of the cumulative forecast density relative to the observed data. Again, the evidence for the MS-BVAR models over the BVAR models is strong.

These failures in forecasting coverage over the different horizons — which are the source of the results in Table 6 — are more evident in a graphical presentation of the forecasts with their 68% error bands. Fig. 5 presents the forecasts for each model over the 12-period forecast horizon. Here, some of the forecasting models do better than others for some of the variables. The forecasts from the two BVAR models have 68% forecast credible intervals (CIs) that cover the true data well across all of the data series, but their CIs are the largest. This is why these estimators have better CRPSs than those seen in the univariate model for three of the four series. The reverse is seen in the C2T-V

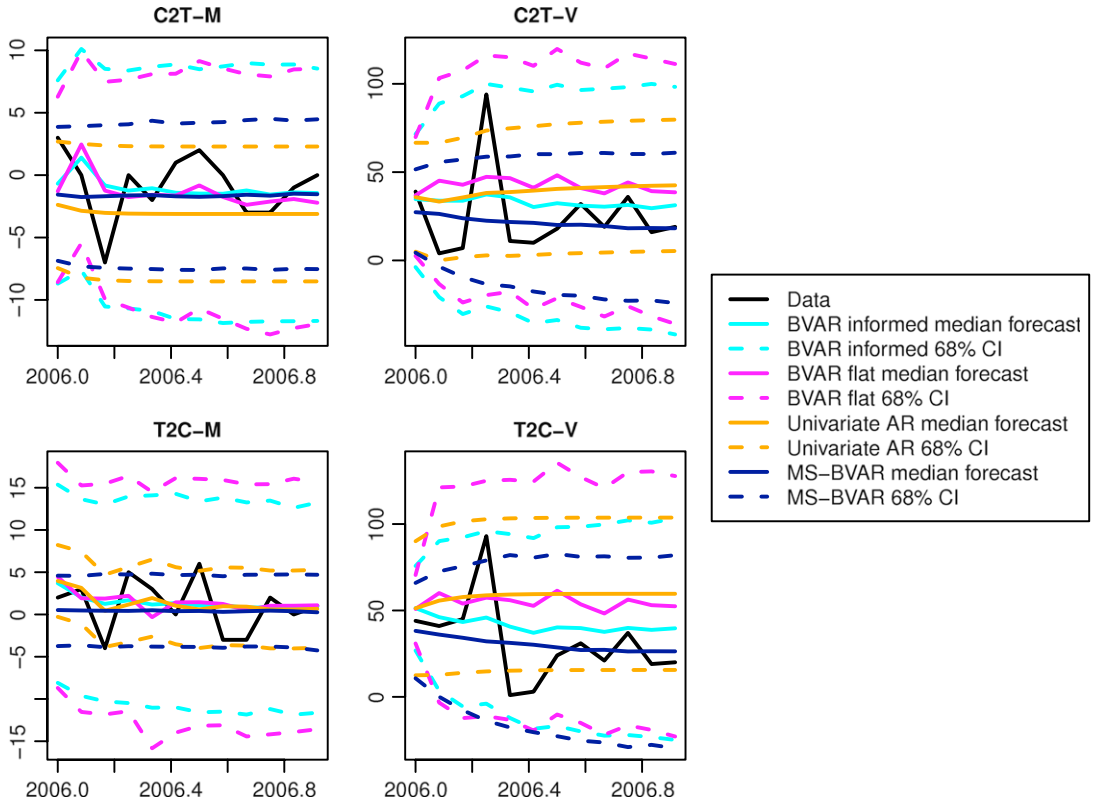

Fig. 5. Cross-Straits forecasts, 2005(12)–2006(12), based on an ensemble of 5000 forecasts. The error bands are 68% around the median forecast.

and T2C-V series, where there is evidence that the univariate forecasts (orange) have a slightly better coverage of the realized data, because the BVAR-based forecasts have coverages that are too wide. The best forecasts are those for the MS-BVAR model (dark blue). These have very narrow 68% forecast intervals that are quite close to the observed data. The forecast densities show us how and when these forecasting methods succeed and fail.

The assessment of the calibration and sharpness of the Cross-Straits forecasts is done using the PIT, the VRH, and sharpness diagrams. Recall that the PIT measures how well the density of the forecasts matches the actual density of the data, using Eq. (7). Two factors to note here are, first, the forecast density evaluation in Eq. (7) is a theoretical construct; and, second, the forecast density is a three-dimensional object, a multivariate prediction comprising a series of conditional vectors over time. To compute the probability integral transforms for the forecasts, we generate the PIT values empirically from the forecast density over all 12 periods for each variable. Since our interest is in the holistic evaluation of the forecasts, we choose the toughest case: the empirical forecast density from the 5000 forecast draws over the entire twelve-period forecast horizon (rather than one step at a time). This is a very stringent test, since, as Fig. 5 showed, the forecast density coverages are quite poor for some of the variable-model combinations.

Fig. 6 shows the PIT histograms for each model and variable over the 12-period forecast horizon. The rows correspond to the four estimators employed, and the columns to the forecasted variables. These PIT histograms show the rather poor performances of the forecasting models over the forecast horizon. For each model, at least one of the series is forecast poorly; that is, the forecast density is mis-centered or has overly generous tails. This is shown clearly in the PIT histograms. For the univariate AR forecasts, the mis-centering is evident from the spottiness and spikes of the PIT. For the BVAR forecasts, the inverted-U shapes of the PITs indicate poor tail coverages relative to the observed values. For the MS-BVAR forecasts, the coverages are poor for the C2T-M and T2C-M series in the upper tail of the density, but comparable to those of the BVAR model for the C2T-V and T2C-V series. For each plot, the Kolmogorov–Smirnov test statistics for a null of uniformity are reported. Only the MS-BVAR forecast consistently gives a set of p-values that do not reject uniformity across the four variables in the model.

The conclusion we reach based on both the earlier results and the PIT histograms is that the BVAR and univariate forecasts are poorly calibrated. The calibration of the MS-BVAR forecasts is slightly better, especially in terms of the 68% coverages reported in Fig. 5. The real differences between the models arise in the tails. To assess the calibration and discrimination of the forecasts jointly, Fig. 7 presents the verification rank histograms. Here, the relevant forecast ensemble is the 5000 *vectorized* forecasts for the four variables over the forecast horizon (meaning that the length is 24 (48) for the 6- (12-)period horizon) for each model. This gives us up to 24 ranked forecasts for the

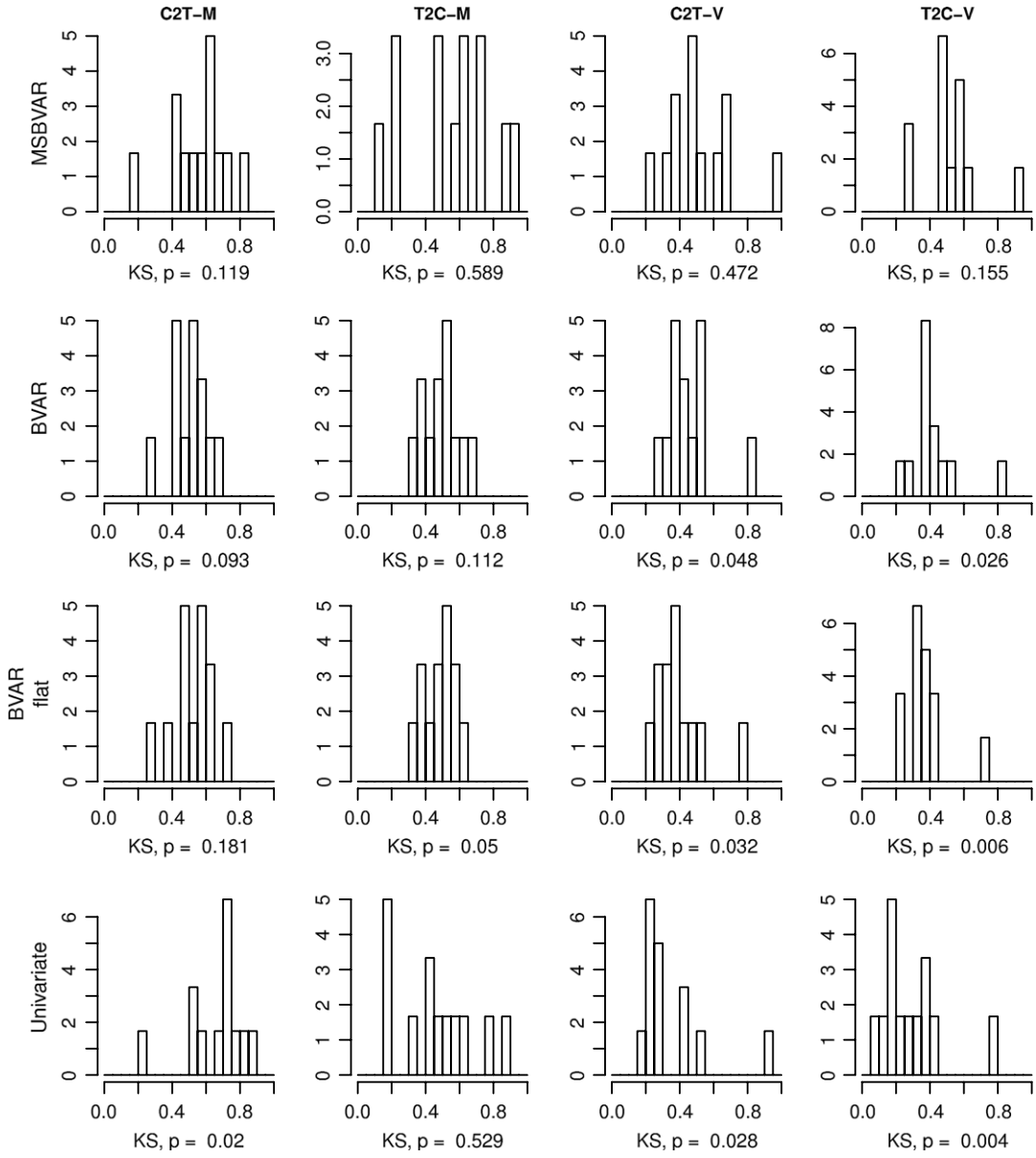

Fig. 6. Cross-Straits forecast PITs for 12-period forecasts. The *p*-values for the Kolmogorov–Smirnov tests for uniformity are give below each plot.

6-period model and up to 48 ranked forecasts for the 12-period model. The verification rank asks how many times this vector of forecasts is below the observed values across the ensemble of 5000 forecasts.

Fig. 7 gives the clearest explanation of the forecast performances over all of the measures discussed. Starting with the six-period forecasts, the MS-BVAR VRHs (first row and column) are uniform, based on the Kolmogorov–Smirnov test (p-value $=$ 0.56). The only other estimator with the same conclusion is the six-period univariate forecast (p-value $=$ 0.86). For the BVAR models, the six-period forecasts offer evidence rejecting the uniformity of the VRH, since the forecast density tail coverages are too large. This is because the BVAR forecast VRHs are too dispersed, since there are very few extreme ranks. Once we consider the 12-period forecasts, there is evidence that the forecasting models are beginning to break down. The VRHs look less uniform in the second column of Fig. 7, as is confirmed by the Komolgorov–Smirnov tests for all but the univariate model over this longer forecast horizon.

Violin plots are presented in Fig. 8 for assessing the sharpness. Each column corresponds to a forecast horizon, and the rows of plots correspond to the forecasted variables. Each violin represents the forecast density for a given estimator, forecast horizon and variable. The scales of the y-axes *are not the same in each row or across the columns*. All of the forecast models produce forecasts that are generally quite close to the observed value, since the

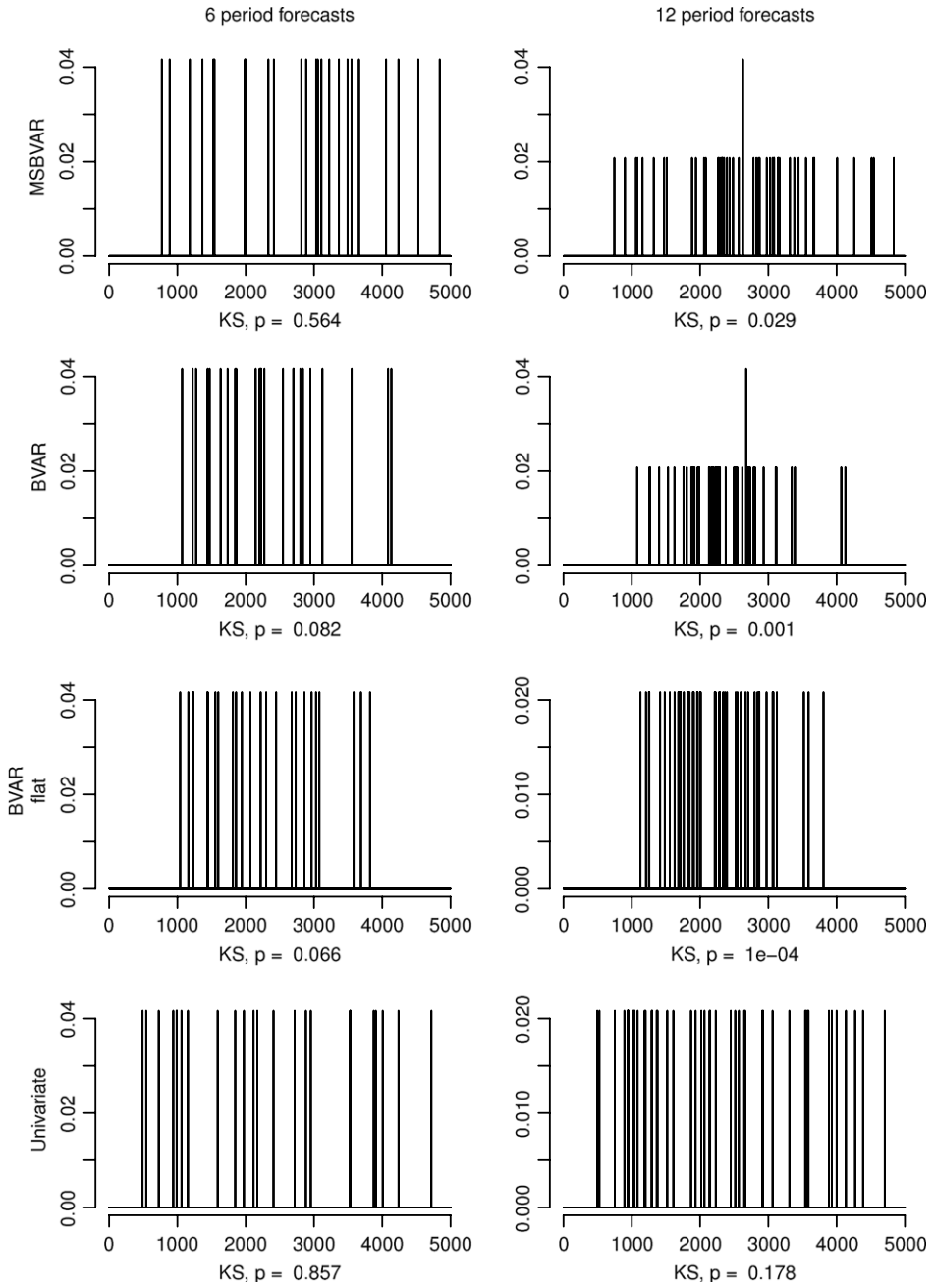

Fig. 7. Cross-Straits verification rank histogram plots. The *p*-values for the Kolmogorov–Smirnov tests for uniformity are given below each plot.

median forecasts (the white dots) are close to the observed values at each time point. The aspect that separates the forecast performance is the size of the credible intervals across the models (left to right within a plot) and forecast horizons (left to right across the columns). The tightest or narrowest forecast densities are those for the C2T-M and T2C-M series using the MS-BVAR model. The MS-BVAR model produces larger intervals for the verbal conflict series. The BVAR models have forecast densities that are many times larger (consistent with the RMSE, MAE, and CRPS results reported above). The univariate AR models do well for forecasting the material conflict-cooperation series, but have wide forecast densities for the verbal conflict-cooperation series (the bottom two rows in the figure).

However, note that, even in the forecast intervals in the sharpness diagram, the eye is drawn toward the extremes, or the whiskers on each plot. Looking at the densities — the grey parts — the densities are better centered and more concentrated overall for the MS-BVAR model results.

The Cross-Straits example results are similar to those from our simulation. First, using the RMSE or MAE criteria

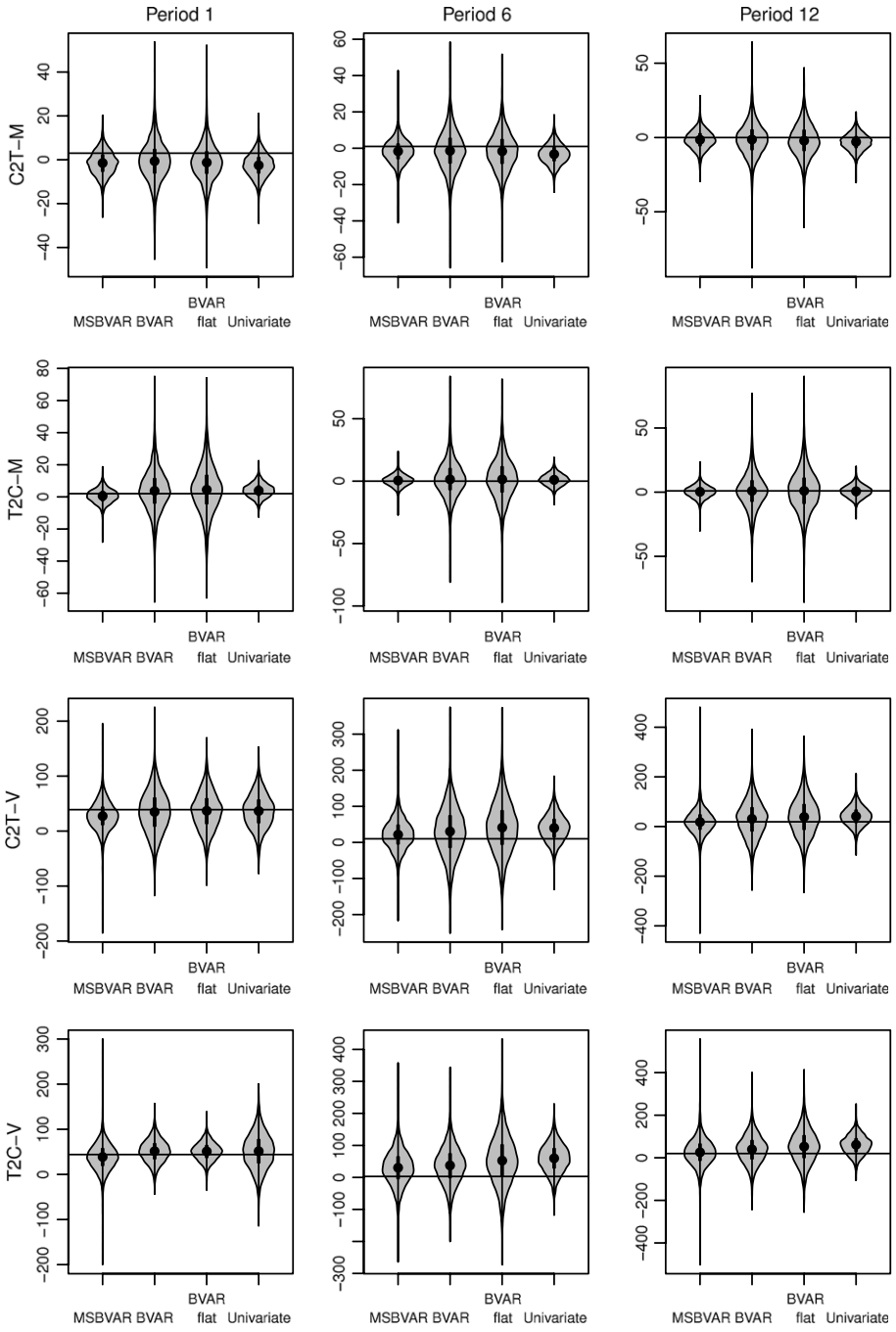

Fig. 8. Sharpness violin plots. The estimators are labeled at the bottom of each plot, while the columns are for the forecast periods and the rows are for the variables. The *y*-axes are the forecast values. The horizontal line in each plot is the observed data value for that period. The black dot in each violin is the median, and the black bar is the inter-quartile range.

alone does not assess the performance of the forecast models properly. Second, the PIT histograms only tell part of the story. Unlike our stylized simulation, the Cross-Straits data show that the way in which we marginalize the PIT across variables and time does not make it clear *which* forecast in an ensemble is performing poorly. While we know from Fig. 5 which models predict which series best over each horizon, even a seemingly holistic assessment can confound the forecast performance. Finally, as the VRH plots illustrate, we can find some evidence that will let us handicap the winners across forecasting models. The performance of the MS-BVAR model is potentially superior to that of the univariate models. This suite of tools gives us a clear route to improving the forecast performance that was not present in the comparisons of RMSE and MAE alone.

4. Conclusion

Because conflict forecasting analyzes the behavior of complex, open stochastic systems, we often make predictions about distributions rather than points.[22] This makes the problem of evaluating a prediction more difficult than the predictions of deterministic systems — for example, predictions of the locations of a comet and a spacecraft which is attempting to rendezvous with it — where simpler point predictions are adequate. Despite this fact, there has been a tendency in political forecasting to use methods such as RMSE and MAE. We have demonstrated that these metrics can be misleading in conflict forecasting.

Thanks to work in economics and meteorology, there exists a rich set of tools for dealing with this problem. We discussed a wide variety of them, along with some of their advantages and disadvantages. As is clear from our discussion and illustrations, no one tool is sufficient for assessing model performance: a suite of tools is needed. In addition, our illustrations show how the evaluation of calibration and sharpness can provide guidance on how the models might be refined further, for example by showing the existence of systematic bias and underdispersion in the estimates.

Acknowledgments

Earlier versions of this paper were presented at the Annual Meeting of the Midwest Political Science Association, Chicago, April 2012, and at the summer meeting of the Society for Political Methodology, Princeton University, July 2011. Drafts of this paper were also presented at colloquia at Princeton University, Texas A & M University, University of Pittsburgh, and University of Rochester. We thank the members of the Society for Political Methodology, participants at the colloquia, and especially Robert Erikson, Eleonora Mattiacci, Xun Pang, and Michael D. Ward for comments. This research is supported by the U.S. National Science Foundation, award numbers SES-0921051, SES-0921018, and SES-1004414. The authors are responsible for the contents.

Appendix A. MS-BVAR model

A.1. Basic model

The VAR and BVAR models used here are covered extensively by Hamilton (1994); Lütkepohl (2007), and Brandt and Freeman (2009). The MS-BVAR model is less well known, and versions of it are provided by Krolzig (1997) and Sims, Waggoner, and Zha (2008). This model is for multivariate time series where a latent indicator variable, governed by a first order Markov process, allows switching from one VAR specification to another.

Let y_t be an observed m-dimensional time series for $t = 1, \ldots, T$. Define $s_t = k$ for $k \in \{1, 2, \ldots, h\}$ as the unobserved state or regime variable that classifies the observations into the regimes where h is known. Finally, assume that there are regime-specific parameters, so that $y_t|s_t \sim f(y_t, s_t|\theta_k)\, \theta_k \in \Theta$. The latent s_t follow a first order Markov process with Q as the $h \times h$ transition matrix of regime transition probabilities. Thus, q_{ij}, as elements of Q, are the probabilities of transitioning from regime $s_{t-1} = j$ to $s_t = i$, namely $\Pr(s_t = i|s_{t-1} = j)$.

The resulting joint posterior of the data and the regime classifications is

$$\Pr(y_t, s_t|\Theta, Q) = \prod_{t=1}^{T}\left(\sum_{s_t=1}^{h} \underbrace{\Pr(y_t|Y_{t-1}, \Theta, s_t)}_{\text{Regime Likelihood}} \times \underbrace{\Pr(s_t|Y_{t-1}, \Theta, Q)}_{\text{Regime probability}}\right), \quad (8)$$

where Y_{t-1} is the history of y_t, for a given transition matrix Q. Note that this formulation is general for any $\Pr(y_t|s_t)$, as per Frühwirth-Schnatter (2008).

If s_t is known (or simulated as below), this means that one can carry out the regime-specific vector autoregressions:

$$y_t = c(s_t) + \sum_{\ell=1}^{p} y_{t-\ell} B_\ell(s_t) + \epsilon(s_t), \quad (9)$$

$$\epsilon_t(s_t) \sim N(0, \Sigma(s_t)) \quad t = 1, 2, \ldots, T, \quad (10)$$

where $c(s_t)$, $B_\ell(s_t)$, $\epsilon(s_t)$, and $\Sigma(s_t)$ are the regime specific constants, autoregressive coefficient matrices, residuals, and error covariance for a VAR(p) model for regime s_t. The parameters in Eqs. (9) and (10) for a given regime k make up the parameter space $\theta_k \in \Theta$.

A.2. Priors and estimation

The prior used has the form:

$$\Pr(\Theta, Q, S_T) = \underbrace{\Pr(\Theta)}_{\text{Sims–Zha BVAR}} \underbrace{\Pr(Q)}_{\text{Dirichlet}} \Pr(s_0|\Theta, Q) \times \underbrace{\prod_{t-1}^{T} \Pr(s_t|\Theta, Q, S_{t-1})}_{\text{Markov process}}, \quad (11)$$

where s_0 is the initial state with prior $\frac{1}{h}$, s_t are the regime indices, and S_{t-1} is the previous state path. Here the joint prior over (Θ, S_T, Q) has been partitioned into a conditional prior where the regime-specific VAR parameters have a Sims–Zha prior (Sims et al., 2008).[23]

[22] In terms of game theory, the prediction is a single outcome when players have dominant strategies, and a distribution of outcomes when players use mixed strategies or when players are boundedly rational (e.g., quantal response equilibrium).

[23] Using the parameterization for the Sims–Zha prior, we set $\lambda_0 = 0.8$, $\lambda_1 = 0.15$, $\lambda_3 = 1$, $\lambda_4 = 0.2$, $\lambda_5 = 0$, $\mu_5 = 0$, and $\mu_6 = 0$. This is a weakly informed prior that is symmetric across the regimes. For $h = 2, 3$, the prior on Q_h is Dirichlet, or D_h:

$$D_2 = \begin{pmatrix} 10 & 2 \\ 2 & 5 \end{pmatrix}, \qquad D_3 = \begin{pmatrix} 10 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 2 \end{pmatrix}.$$

Table 7
Point metrics and scoring rules for competing forecasting models.

	Series 1			Series 2		
	MSBVAR	VAR	Normal	MSBVAR	VAR	Normal
RMSE	1.69	1.56	1.57	1.55	1.50	1.57
MAE	1.46	1.35	1.31	1.33	1.28	1.32
CRPS	0.59	0.58	0.64	0.59	0.58	1.11
IGN	1.52	1.45	1.52	1.52	1.48	2.25

Under this prior, the resulting conditional posteriors are (see Sims et al., 2008):

$$\Pr(S_T|Y_T, \Theta, Q) \propto \Pr(s_t|S_T) \quad \forall\, t \tag{12}$$

$$\Pr(Q|Y_T, \Theta, S_T) \propto \prod_{i=1}^{h} p_{i,j}^{n_{i,j}+\alpha_{i,j}} \tag{13}$$

$$\Pr(\Theta|Y_T, S_T, Q) \propto N(\tilde{\Theta}, \tilde{\Sigma}), \tag{14}$$

where α_{ij} is the Dirichlet prior on transitions from regime j to regime i.

A.3. MS-BVAR forecasting and uncertainty

To construct forecasts that will account for uncertainty from the (1) data, (2) parameters, (3) regime classification, and (4) forecast errors, the methods of Waggoner and Zha (1999) are employed. Thus, when forecasting τ periods ahead from period T, the conditional posterior of the last section is modified as follows.

For the first forecast sample:

1. Using the last estimate of Q, project forward the S_T for $S_{T+1}, S_{T+2}, \ldots, S_{t+\tau}$.
2. Then, based on the S_t classifications, forecast $y_{T+1}, y_{T+2}, \ldots, y_{T+\tau}$. Use the values from $Y = (y_1, \ldots, y_{T+\tau})$ as a data augmented forecast, and sample for future iterations.
3. Update the estimate of Θ for the sample, inclusive of the forecasts.
4. Update the estimates of Q conditional on the $S_1, \ldots, S_{T+\tau}$.

These iterations are repeated N times, after a burn-in period, in order to generate the relevant posterior forecasts for the MS-BVAR model.

Appendix B. Additional Monte Carlo for a linear VAR DGP

To show that the simulation results in Section 3.1 are not an artifact of using the MS-VAR DGP, we simulated a non-switching VAR process as per the Regime 1 specification. We then generated 1000 forecasts from these new data from each of the models we compare in that section (MSBVAR, VAR, Normal). Table 7 shows the different forecast evaluation metrics for each series and estimator.

Here, as we would expect, the MS-BVAR models are rejected based on the MSE and MAE comparisons. The CRPS and IGN score results also point in the same direction. Thus, for this simpler, linear DGP, the comparisons turn out as expected. However, RMSE and MAE will not be as useful when the DGP is of the non-linear, switching type—the type that international relations theory predicts.

References

Alker, H. R., Gurr, T. R., & Rupesinghe, K. (2001). *Journeys through conflict: narratives and lessons*. Rowman & Littlefield Publishers.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Azar, E. E. (1972). Conflict escalation and conflict reduction in an international crisis: Suez, 1956. *Journal of Conflict Resolution*, *16*(2), 183–201.

Bagozzi, B. E. (2014). *Forecasting civil conflict with zero-inflated count models*. University of Minnesota, Manuscript.

Beck, N., King, G., & Zeng, L. (2000). Improving quantitative studies of international conflict: a conjecture. *American Political Science Review*, *94*, 21–36.

Bennett, P., & McQuade, P. (1996). Experimental dramas: prototyping a multiuser negotiation simulation. *Group Decision and Negotiation*, *5*, 119–136.

Bloomfield, L. P., & Moulton, A. (1989). *CASCON III: computer-aided system for analysis of local conflicts*. Tech. rep., MIT Center for International Studies, Cambridge, MA.

Bloomfield, L. P., & Moulton, A. (1997). *Managing international conflict: from theory to policy*. New York, NY: St. Martin's Press.

Brams, S. J., & Togman, J. M. (2000). Agreement through threats: the north Ireland case. In N. Miroslav, & J. Lepgold (Eds.), *Being useful: policy relevance and international relations theory* (pp. 325–342). Ann Arbor: University of Michigan Press.

Brandt, P. T., Colaresi, M., & Freeman, J. R. (2008). The dynamics of reciprocity, accountability, and credibility. *Journal of Conflict Resolution*, *52*, 343–374.

Brandt, P. T., & Freeman, J. R. (2006). Advances in Bayesian time series modeling and the study of politics: theory testing, forecasting, and policy analysis. *Political Analysis*, *14*(1), 1–36.

Brandt, P. T., & Freeman, J. R. (2009). Modeling macro political dynamics. *Political Analysis*, *17*(2), 113–142.

Brandt, P.T., Freeman, J.R., Lin, T., & Schrodt, P.A. (2013). *A Bayesian time series approach to the comparison of conflict dynamics*. Manuscript.

Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (2011). Real time, time series forecasting of political conflict. *Conflict Management and Peace Science*, *28*, 41–64.

Butterworth, R. L., & Scranton, M. E. (1976). *Managing interstate conflict, 1945–1974: data with synopses*. Pittsburgh, PA: University of Pittsburgh.

Chiozza, G., & Goemans, H. (2011). *Leaders and international conflict*. New York: Cambridge University Press.

Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic time series*. New York: Cambridge University Press.

Clements, M. P., & Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: applications to output growth and employment. *Journal of Forecasting*, *19*, 255–276.

Dawid, A. P. (1984). Statistical theory: a prequential approach. *Journal of the Royal Statistical Society, Series A*, *147*, 278–292.

de Mesquita, B. B. (2002). *Predicting politics*. Columbus, OH: Ohio State University Press.

de Mesquita, B. B. (2011). A new model for predicting policy choices: preliminary tests. *Conflict Management and Peace Science*, *28*(1), 65–87.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883.

Diebold, F. X., Hahn, J., & Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics*, *81*, 661–673.

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. In G. S. Maddala, & C. R. Rao (Eds.), *Handbook of statistics 14: statistical methods in finance*. Amsterdam: North-Holland.
Diehl, P. F. (2006). Just a phase? Integrating conflict dynamics over time. *Conflict Management and Peace Science*, *23*(2), 199–210.
Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985–987.
Esty, D. C., Goldstone, J., Gurr, T. R., Harff, B., Surko, P. T., Unger, A. N., et al. (1998). The state failure project: early warning research for U.S. foreign policy planning. In J. L. Davies, & T. R. Gurr (Eds.), *Preventive measures: building risk assessment and crisis early warning systems* (pp. 27–38). Lanham, Md: Rowman and Littlefield.
Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics*, *24*, 435–468.
Frühwirth-Schnatter, S. (2006). Finite mixture and Markov switching models. In *Springer series in statistics*. New York: Springer.
Frühwirth-Schnatter, S. (2008). Comment on article by Rydén. *Bayesian Analysis*, *3*(4), 689–698.
Gerner, D. J., Schrodt, P. A., & Yilmaz, Ö (2009). Conflict and mediation event observations (CAMEO): an event data framework for a post Cold War world. In J. Bercovitch, & S. Gartner (Eds.), *International conflict mediation: new approaches and findings* (pp. 287–304). New York: Routledge, Ch. 13.
Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B*, *69*, 243–268.
Gneiting, T., Larson, K., Westrick, K., Genton, M. G., & Aldrich, E. (2006). Calibrated probabilistic forecasting at the Stateline wind energy centre: the regime-switching space–time (RST) method. *Journal of the American Statistical Association*, *101*, 968–979.
Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.
Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*, 248–249.
Goldstein, J. S., & Freeman, J. R. (1990). *Three-way street: strategic reciprocity in world politics*. Chicago: University of Chicago Press.
Goldstein, J. S., & Pevehouse, J. C. (1997). Reciprocity, bullying, and international cooperation: a time series analysis of the Bosnia conflict. *American Political Science Review*, *91*(3), 515–529.
Goldstein, J. S., Pevehouse, J. C., Gerner, D. J., & Telhami, S. (2001). Reciprocity, triangularity and cooperation in the Middle East, 1979–1997. *Journal of Conflict Resolution*, *45*(5), 594–620.
Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., et al. (2010). A global model for forecasting political instability. *American Journal of Political Science*, *54*(1), 190–208.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, *14*, 107–114.
Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*, 550–560.
Hamilton, J. D. (1994). *Time series analysis*. Cambridge University Press.
Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., & Urdal, H. (2012). Predicting armed conflict, 2010–2050. *International Studies Quarterly*, *1*, 1–21.
Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*, 559–570.
Huth, P. K., & Allee, T. L. (2002). *The democratic peace and territorial conflict in the twentieth century*. Ann Arbor, MI: University of Michigan Press.
King, G., & Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, *53*, 623–658.
Krolzig, H.-M. (1997). *Markov-switching vector autoregressions: modeling, statistical inference, and application to business cycle analysis*. Berlin: Springer.
Lai, B., & Slater, D. (2006). Institutions of the offensive: domestic sources of dispute initiation in authoritarian regimes, 1950–1992. *American Journal of Political Science*, *50*(1), 113–126.
Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Cambridge University Press.
Moore, W. (1998). Repression and dissent: substitution, context, and timing. *American Journal of Political Science*, *42*(3), 851–873.
Morris, B. (1999). *Righteous victims: a history of the Zionist-Arab conflict, 1881–1999*. Vintage Books.
O'Brien, S. P. (2010). Crisis early warning and decision support: contemporary approaches and thoughts on future research. *International Studies Review*, *12*(1), 87–104.
Organski, A. F. K., & Lust-Okar, E. (1997). The tug of war over Jerusalem: leaders, strategies and outcomes. *International Interactions*, *23*(4), 333–350.
Pevehouse, J. C., & Goldstein, J. S. (1999). Serbian compliance or defiance in Kosovo? Statistical analysis and real-time predictions. *Journal of Conflict Resolution*, *43*(4), 538–546.
Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.
Ross, R. (2000). The 1995–96 Taiwan Strait confrontation: coercion, credibility, and the use of force. *International Security*, *25*(2), 87–123.
Rost, N., Schneider, G., & Kleibl, J. (2009). A global risk assessment model for civil wars. *Social Science Research*, *38*(4), 921–933.
Schrodt, P.A. (2000). Forecasting conflict in the Balkans using hidden Markov models. *Paper presented at the annual meeting of the American Political Science Association*. Washington, D.C.
Schrodt, P. A., & Gerner, D. (2000). Cluster based early warning indicators for political change in the contemporary Levant. *American Political Science Review*, *94*(4), 803–818.
Senese, P. D., & Vasquez, J. A. (2008). *The steps to war: an empirical study*. Princeton, N.J.: Princeton University Press.
Shellman, S., Reeves, A., & Stewart, B. (2007). Fair & balanced or fit to print? In *The effects of media sources on statistical inferences*. University of Georgia.
Sherman, F. L. (1994). SHERFACS: a cross-paradigm, hierarchical, and contextually sensitive conflict management data set. *International Interactions*, *20*(1–2), 79–100.
Sherman, F. L., & Neack, L. (1993). Imagining the possibilities: the prospects of isolating the genome of international conflict from the SHERFACS dataset. In R. L. Merritt, R. G. Muncaster, & D. A. Zinnes (Eds.), *International event-data developments: DDIR Phase II* (pp. 87–112). Ann Arbor, MI: University of Michigan Press.
Sims, C. A., Waggoner, D. F., & Zha, T. (2008). Methods for inference in large multiple-equation Markov-switching models. *Journal of Econometrics*, *146*(2), 255–274.
Sims, C. A., & Zha, T. A. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, *39*(4), 949–968.
Tay, A. S., & Wallis, K. F. (2000). Density forecasting: a survey. *Journal of Forecasting*, *19*, 235–254.
Tetlock, P. E. (2005). *Expert political judgement*. Princeton: Princeton University Press.
Timmerman, A. (2000). Editorial: Density forecasting in economics and finance. *Journal of Forecasting*, *19*, 231–234.
Tung, C. (2003). Cross-strait economic relations: China's leverage and Taiwan's vulnerability. *Issues and Studies*, *39*(3), 137–176.
Vasquez, J. A., Johnson, J. T., Jaffe, S., & Stamato, L. (1995). *Beyond confrontation: learning conflict resolution in the post-Cold War era*. Cambridge University Press.
Waggoner, D. F., & Zha, T. A. (1999). Conditional forecasts in dynamic multivariate models. *Review of Economics and Statistics*, *81*(4), 639–651.
Ward, M. (1982). Cooperation and conflict in foreign policy behavior: reaction and memory. *International Studies Quarterly*, *26*(1), 87–126.
Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by *p*-value: predicting civil conflicts. *Journal of Peace Research*, *47*(4), 363–375.
Weeks, J. L. (2012). Strongmen and straw men: authoritarian regimes and the initiation of conflict. *American Political Science Review*, *106*(2), 326–347.
Weigend, A. S., & Shi, S. (2000). Predicting daily probability distributions of S&P returns. *Journal of Forecasting*, *19*, 375–392.
Winkler, R. L., & Murphy, A. H. (1968). 'Good' probability assessors. *Journal of Applied Meteorology*, *7*, 751–758.
Wolfers, J., & Zitzewitz, E. W. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126.
Wolford, S. (2007). The turnover trap: new leaders, reputations and international conflict. *The American Journal of Political Science*, *51*(4), 772–788.
Zeitzoff, T. (2011). Using social media to measure conflict dynamics. *Journal of Conflict Resolution*, *55*(6), 938–969.

Patrick T. Brandt is an Associate Professor in the School of Economic, Political and Policy Sciences at the University of Texas, Dallas, and a Faculty Associate at the University's Center for Global Collective Action. He regularly teaches Bayesian and time series analysis material at the Inter-University Consortium for Political and Social Research's (ICPSR) Summer Program at the University of Michigan. His research involves the development of time series models for application to political, economic, and social phenomena. This research has been supported repeatedly by the U.S. National Science Foundation.

John R. Freeman is the John Black Johnston Professor in the College of Liberal Arts at the University of Minnesota. He is a fellow of the American Academy of Arts and Sciences and of the American Political Science

Association's Political Methodology Society. He has taught at a number of institutions, including the Massachusetts Institute of Technology and the University of Michigan. His publications are in the fields of political economy, international relations, and methodology.

Phillip A. Schrodt is Professor of Political Science at Pennsylvania State University. Prior to joining Penn State in 2010, he taught for 21 years at the University of Kansas, and 11 years at Northwestern University. His areas of research are quantitative models of political conflict and computational methodology. His current research focuses on predicting political change using statistical and pattern recognition methods, and has been supported by the National Science Foundation, the Defense Advanced Research Projects Agency, and the U.S. government's Political Instability Task Force. He is a past president and current fellow of the Society for Political Methodology.