Chapter 2: Modeling the Cold War

2.1 Introduction

The Cold War between the United States and the Soviet Union was the defining factor of the international order of the second half of the 20th century, helping to drive to everything from local insurgencies to space exploration. The conflict was rooted both in ideological differences and a desire for security. Indeed, both sides possessed not only powerful conventional militaries with global reach, but arsenals of nuclear weapons capable of guaranteeing so-called Mutually Assured Destruction. Each side posed an existential risk not only to the other's system of government, but to the lives of a large proportion of their citizens. Both sides also had stable blocs of aligned countries; in Europe this divide was geographic, with the continent divided by an 'Iron Curtain' between the west and east. The apparent parity of the two sides, and the stability of their alliances, led to an expectation that the cold war between them would persist indefinitely, terminated perhaps only in apocalyptic conflict.

Yet it was not the world but the Cold War, and the USSR itself, which ended – "Not with a bang but with a whimper," as Dobson and Marsh (2007); Prados (2011) and others quote T. S. Eliot. The end of the Cold War took policymakers and scholars both by surprise. In the aftermath, many political scientists (Gaddis, 1992; Lebow and Risse-Kappen, 1995) suggested that this was evidence of shortcomings in the state of international relations theory and methodology. Bueno de Mesquita (1998) offered an alternative explanation: that the end of the Cold War *was* predictable by contemporary international relations theory, but that it was an emergent phenomena: an unexpected higher-level outcome generated from the combination of multiple lower-level interactions happening in accordance with theoretical expectations. Indeed, complex systems do often exhibit such emergent phenomena, which often appear unpredictable or unexpected.

Bueno de Mesquita (1998, 2002) set out to test this claim by simulating the Cold War using the Expected Utility model. Computational modeling is uniquely suited to generating macro-level emergent phenomena from micro-level interactions; thus, if the model is well-grounded in theory and gives rise to the end of the Cold War, this is evidence that the theory was not mistaken, but rather that it was not property applied, or its implications not fully understood. Furthermore, if the model can outperform experts and predict a major historic outcome from relatively minimal data, this would be strong evidence in favor of the model's utility and predictive power. Indeed, the paper argues that the model can predict the end of the Cold War, with a victory for the US and its allies.

In the previous chapter, I presented my own reproduction and extension of the Expected Utility model that Bueno de Mesquita (1998) uses. In this chapter, I attempt to apply these models to the same Cold War case study. This serves several purposes. It attempt to replicate the previous study, while also providing an additional test of my reproduction: if my model produces similar outcomes to the original, this is evidence for a correct replication. Additionally, as with the original model, if my models are capable of successfully predicting the end of the Cold War from historic data, this is evidence if the models' usefulness. As I discuss in the previous chapter, my own models allow me to go further than the original model, generating not just outcomes but key events that lead up to them. Comparing these events – and in particular, conflicts – to historic data provides an additional external test of the models' validity. Finally, by comparing the predictive power of different model variants, I can attempt to locate the most predictive one.

The rest of the paper is organized as follows. I describe the input data, and how I use it to instantiate runs of two model variants, as well as the data I will compare the model outputs to. I then present the results of four experiments, each one a combination of an input dataset with a model variant. I describe the results of each experiment qualitatively, and test their output's power as a predictor of observed conflicts and post-Cold War alliances. Finally, I discuss the results more broadly, highlighting and attempting to explain the models' strenghts and weaknessses.

2.2 Modeling Methodology

The pipeline for this analysis follows a similar process to that described in the previous chapter: I start with a data source and use it to generate a list of actors, and the position and capability values for each. These actors then serve as inputs into the model variants, each of which is instantiated and run multiple times, generating collections of outputs. Finally, I analyze these outcomes and compare them to real-world data.

2.2.1 Input Data

To instantiate a model, we require several things: a list of agents, which in this case are countries active in a international system during the Cold War; and for each agent, an initial position on the one-dimensional position space and a capability value. There is no need here to attempt to estimate salience, since this will be one of the independent values which will be randomized for all agents, as described below.

Bueno de Mesquita (1998) generates the model inputs from data collected by the Correlates of War (COW) project. The list of actors, and their capabilities, are drawn from the National Material Capabilities dataset (), with the model using the 36 most-powerful actors in the starting year of 1948. Their positions are derived from the COW alliance dataset (), measuring each agent's closeness to the security preferences of the United States or Soviet Union. This closeness is defined as the Tau-B measure of similarity in alliance portfolios (Bueno de Mesquita, 1975), normalized such that a value of +100 is full match with the United States, and -100 is a full match with the Soviet Union. In order to make this data compatible with my implementation of the model, I simply rescale the starting positions so that they fall on the $[0, 1]$ range, with 0 being the Soviet Union's starting position and 1 being that of the United States. The original model inputs are reported in Bueno de Mesquita (1998, 2002) and reproduced in Table 2.1. I term this the Original input data.

I then attempt to reproduce the input data from the current, up-to-date COW datasets (Correlates of War Project, 2010; Gibler, 2009). Following the procedure descibed in the

Country	Capability	Position	Country	Capability	Position
Argentina	0.972	0.948	Italy	2.426	0.507
Australia	0.889	0.507	Mexico	0.774	0.948
Belgium	1.182	0.514	Norway	0.23	0.507
Brazil	0.993	0.948	Netherlands	0.836	0.514
Bulgaria	0.345	0.000	Pakistan	1.485	0.507
Canada	1.61	0.781	Philippines	0.408	0.507
China	11.941	0.507	Poland	3.273	0.045
Czechoslovakia	1.401	0.045	Romania	0.606	0.045
Denmark	0.24	0.507	Saudi Arabia	0.125	0.514
Egypt	0.408	0.516	South Africa	0.68	0.507
England	7.863	0.518	USSR	18.256	0.000
France	3.597	0.514	Spain	1.683	0.507
Greece	0.418	0.507	Sweden	0.648	0.507
Hungary	0.45	0.045	Syria	0.104	0.514
India	2.468	0.507	Thailand	0.414	0.507
Iran	0.491	0.512	Turkey	1.347	0.512
Iraq	0.157	0.519	USA	29.956	1.000
Israel	0.125	0.507	Yugoslavia	0.891	0.000

Table 2.1: Original Inputs (Bueno de Mesquita, 1998) with positions rescaled

original paper did not replicate the original reported inputs; however, it does produce a new input dataset, which I term the Updated input data. As shown in Table 2.3, the updated ranking of countries' power in 1948, as measured by the Combined National Material Capabilities Index, is different from that reported in Bueno de Mesquita (1998). This is likely due to updated data and methodologies, but already indicates a difficulty in replicating the previous study.

Next, we must assign each agent to a position. To do so, I construct an undirected network from the most recent COW alliance data for 1948, where the edge between each pair of countries is coded based on the alliance type, as summarized in Table 2.2. We can then calculate similarity scores between the adjacency matrix rows or columns of the actors, each of which represents the actor's alliance portfolio. To maintain correspondence with the previous data, I use the Tau-b similarity score between portfolios.

Alliance Code	Type
1	Entente: "one or both states in the dyad had an understanding that consultations with the other state in the dyad would take place if a crisis occurred."
2	Non-Aggression: "one or both states in the dyad had a non-aggression pact with the other state in the dyad."
3	Neutrality: "one or both states in the dyad had a neutrality pact with the other state in the dyad."
4	Defense: "one or both states in the dyad had a defense pact with the other state in the dyad."

Table 2.2: Alliance Coding (descriptions quoted from Gibler (2013))

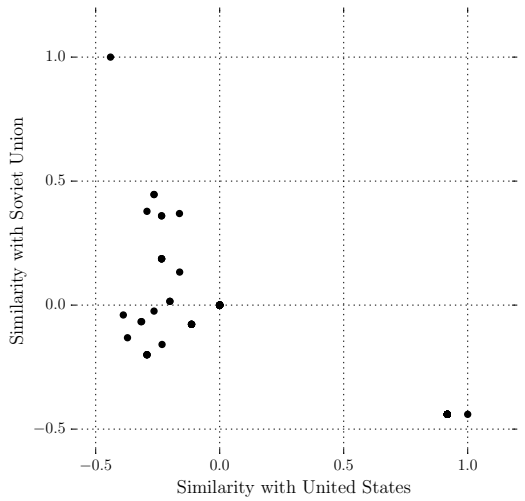


Figure 2.1: τ similarities of actors to United States and Soviet Union

It is here that we run into a question unaddressed in Bueno de Mesquita (1998). Comparing each country's alliance portfolio to those of the United States and the Soviet Union produces not a single value per country, but two: one for each superpower being compared to. As shown in Figure 2.1, the relationship between the similarities is not completely linear, meaning we need to choose a way to reduce the two-dimensional security position space to one dimension. Since the two dimensions are highly correlated (correlation coefficient -0.81), I use principal component analysis (Tipping and Bishop, 1999) to project the two-dimensional data onto a one-dimensional space, which I then rescale so that the values are between 0 and 1. The resuling positions for the Updated Inputs are reported in Table 2.3. This input set includes all the countries for which data is available in 1948. Like the original input data, it excludes Japan and West and East Germany, which had not yet become sovereign at that point.

Country	Capability	Position	Country	Capability	Position	Country	Capability	Position
Afghanistan	0.0023342	0.58	Haiti	0.0004093	0.98	Paraguay	0.0004179	0.98
Albania	0.0008609	0.36	Honduras	0.0001638	0.98	Peru	0.0017403	0.98
Argentina	0.0082709	0.98	Hungary	0.0041569	0.3	Philippines	0.0033156	0.56
Australia	0.0083341	0.57	Iceland	1.26E-05	0.56	Poland	0.0298539	0.3
Belgium	0.0109084	0.42	India	0.0524505	0.56	Portugal	0.0026259	0.46
Bolivia	0.0006092	0.98	Iran	0.0042677	0.51	Romania	0.0054891	0.34
Brazil	0.0121925	0.98	Iraq	0.001297	0.49	Russia	0.1639996	0.00
Bulgaria	0.0031112	0.42	Ireland	0.0009568	0.56	Saudi Arabia	0.0011293	0.59
Canada	0.0125928	0.56	Israel	0.0014135	0.56	South Africa	0.0052241	0.56
Chile	0.0020532	0.98	Italy	0.0192622	0.56	Spain	0.0148603	0.57
China	0.1150552	0.57	Jordan	0.000219	0.52	Sri Lanka	0.0008769	0.56
Colombia	0.002628	0.98	Lebanon	0.0003211	0.59	Sweden	0.0059829	0.56
Costa Rica	8.90E-05	0.98	Liberia	0.0001402	0.56	Switzerland	0.0019677	0.56
Cuba	0.0017359	0.98	Luxembourg	0.0029497	0.42	Syria	0.000855	0.59
Czechoslovakia	0.0126014	0.34	Mexico	0.0066099	0.98	Thailand	0.0033879	0.56
Denmark	0.0020571	0.56	Mongolia	0.0003635	0.57	Turkey	0.0117226	0.51
Dominican Republic	0.0005684	0.98	Myanmar	0.0022862	0.56	United Kingdom	0.075426	0.54
Ecuador	0.0008852	0.98	Nepal	0.0010616	0.56	United States of America	0.2946597	1.00
Egypt	0.0053934	0.52	Netherlands	0.0083195	0.42	Uruguay	0.000819	0.98
El Salvador	0.0002847	0.98	New Zealand	0.0008759	0.57	Venezuela	0.0013442	0.98
Ethiopia	0.0022585	0.56	Nicaragua	0.000143	0.98	Yemen Arab Republic	0.0005736	0.59
Finland	0.001916	0.57	North Korea	0.0036134	0.56	Yugoslavia	0.0080174	0.32
France	0.0325886	0.51	Norway	0.0015807	0.56			
Greece	0.0035732	0.56	Pakistan	0.0118022	0.56			
Guatemala	0.0006688	0.98	Panama	7.81E-05	0.98			

Table 2.3: New Inputs

	Baseline Model	Updated Model
Original Data	Experiment 1	Experiment 2
Updated Data	Experiment 3	Experiment 4

Table 2.4: Experiment Summary

2.2.2 Model Instantiations

I use two agent behavior variants, which are described in more detail in the previous chapter. They are:

Baseline is the attempt to reproduce the original model, from descriptions in previous papers.

Updated is my own modifications to the model. This variant differs from the baseline in two key ways: Risk acceptance is computed from incoming probabilities of defeat, rather than expected utilities; and agents do not renege on accepted offers if another agent accepts their own offer.

Taken together, the two model variants and two input datasets yield four experiments in all, as shown in Table 2.4.

As Bueno de Mesquita (1998) notes, there are countless unpredictable elements affecting how states will perceive, and act on, their international security interests in the context of the Cold War. Rather than extend the model to explicitly capture more and more of these factors, we implicitly incorporate them by randomizing agent salience values. At the beginning of each step of the model, each agent's salience is randomly drawn from a uniform distribution over the $[0, 1]$ range. Since capability is always multiplied by salience, this effectively adds a noise factor to the agents' strength as well. While agents with greater capabilities are likely to remain stronger than ones with lower capabilities, there exist configuration of salience values under which the generally weaker agent temporarily has an advantage over a stronger agent.

For all experiments, I instantiate all agents with parameters $Q = 0.5$ and $T = 0.5$ indicating maximum uncertainty about the world. Each model is run for 25 steps. Bueno de Mesquita (1998) hypothesizes that this is roughly equivalent to 50 years of history, meaning that agents experience salience shocks, engage in conflicts, and change position approximately once every two years. With 1948 as the starting year, the models will generate data corresponding approximately to 1998 as the end year.

Due to the large number of independent random variables driving the model, I run 1,000 instantiations of each experiment. This volume appears to be sufficiently large to sample the the many different possible configurations of saliences and conflict outcomes; excursions using 10,000 instantiations did not produce meaningfully different outcome distributions. Furthermore, since we are concerned with commonly-emerging outcomes and behaviors, particular phenomena occurring less frequently than in 0.1% of model runs are unlikely to be of practical interest.

2.2.3 Output Data

Each model run generates several outputs. At the beginning of each step of the run, it records the position currently held by each agent, as well as the current median voter position, and capability-weighted mean position. During each step, the model also logs all conflicts which occur between agents. I collect and aggregate these datasets across all 1,000 runs of each experiment.

I will then compare these outputs to more data from the Correlates of War project: alliance data, and militarized interstate disputes (MIDs), for the 1948-1998 timeframe. The MID dataset is particularly important, as it provides an expert-coded listing of all cases where one state threatened or used force against another (Jones et al., 1996). While there are many forms of conflict, militarized disputes, which can escalate into wars, are particularly important to the international system, and thus of particular interest for forecasting. I will attempt to see whether the number of conflict events between pairs of actors generated by the model runs is a useful predictor of the number of conflicts between states in the relevant

time period.

Similarly, the model runs produce traces of agent positions, with agents with closer positions more likely to support one another more strongly in conflict with other agents. In other words, agents with closer positions are more likely to be allied with one another. Thus, I will also test whether the distance betwee agents at the end of model runs is a predictor of alliances in the model end year.

2.3 Results

2.3.1 Overview

The initial question posed by the original paper was whether the end of the Cold War – that is, a substantial shift in the median international security position to the advantage of the United States – could be generated from the model. Figure 2.2 shows the cumulative distribution of model end positions reported by that paper. Note that the largest volume of outcomes appears very close to 0 (equivalent to 0.5, in the variant used here), indicating that the most common type of outcomes is one where the Cold War persists at the end of the model run, with a median position between the starting position of both superpowers and no clear advantage to either side. However, the distribution is also clearly non-symmetric, with more outcomes favoring the United States (closer to +100) than the USSR. The conclusion of the original paper was thus that the end of the Cold War, with an American victory, was not as unforeseeable as others had argued, and in fact could be emerged from the model.

Figure 2.3 shows the same cumulative distributions across the four experiments. Immediately, we note that the overall shape is similar, though not identical. The bulk of outcomes are very close to the center, with substantially more US-victory outcomes than USSR-victory ones. In particular, Experiment 1, which attempts to replicate the original model, has almost no USSR-victory outcomes, and more outcomes where the median position does not indicate a victory to either side.

Let us take a closer look at the behaviors these models are generating. Figure 2.4 shows

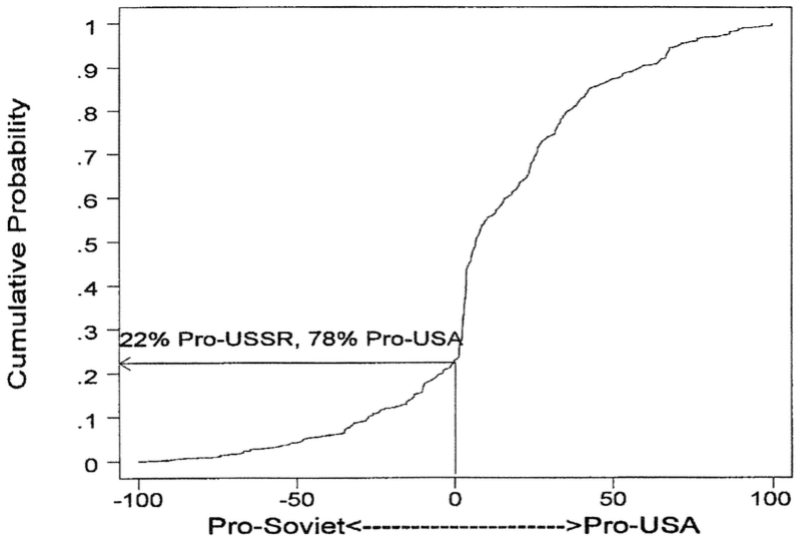


Figure 2.2: Model outcome cumulative distribution from Bueno de Mesquita (1998)

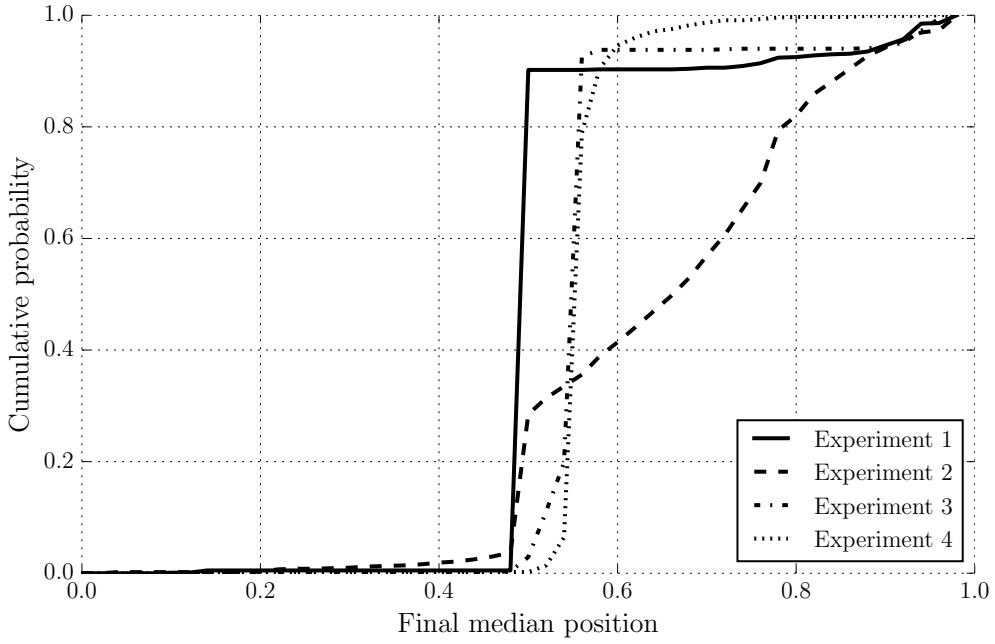
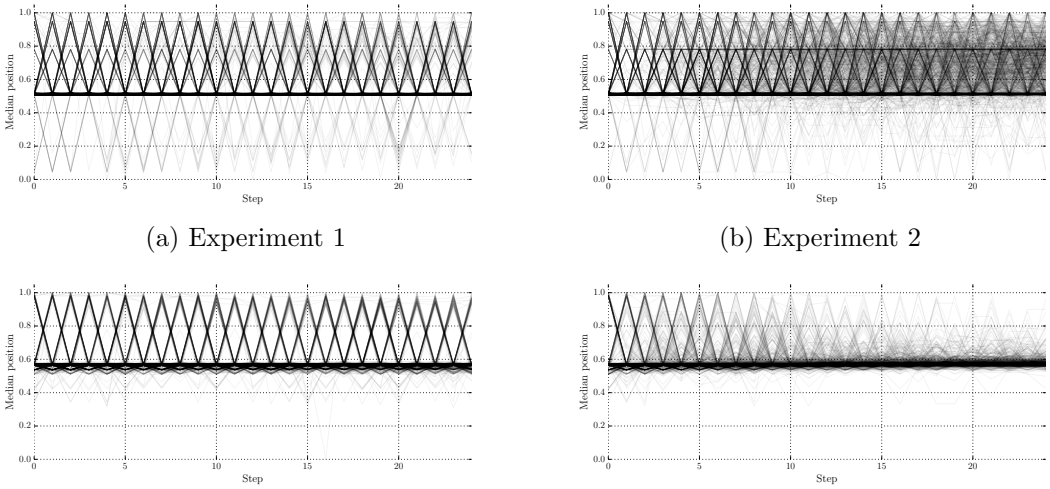


Figure 2.3: Model outcomes cumulative distributions

the superimposed traces of the median position for all the model runs of all experiments. The opacity of each trace line is low, so that the darker a line appears, the more traces exhibited that particular transition. While it is impossible to discern individual runs here, these visualizations serve to highlight recurring patterns. Two patterns are obvious across all four experiments. The first is that, as cumulative distributions Figure 2.3 suggests, the most common positions for the model at any given step are at or close to the starting median – by and large, the median does not change very often. The next key pattern is that, across all the experiments, the median position exhibits a sawtooth pattern, diverging from the 0.5 position towards one extreme or the other and immediately returning. In other words, the advantage swings towards one side or another (though more frequently towards the United States) and then quickly returns to the status quo. This sawtooth pattern is particularly distinct and uniform in the Baseline model used in Experiments 1 and 3. Experiment 2, and to a lesser extent Experiment 4, show a wider variety of traces. In particular, close examination of the Experiment 2 traces indicates more cases where changes in median are not one-round spikes, but transitions to a new stable position. Finally, Experiment 4 shows a decline in movement as time advances, suggesting that the agents are sorting themselves into a more stable configuration around the same median.

The sawtooth pattern raises a question: do the model runs ending in a US (or Soviet, for that matter) victory represent a true long-term shift in the status quo, or are they simply cases where a temporary spike towards one side or the other coincides with the final step of the model? In fact, a close examination of the model runs suggests that the latter is exactly what is happening in Experiments 1 and 3. Figure 2.5 shows example traces from randomly-selected runs which end with a median position of 0.7 or greater, indicating a US victory. In the traces from Experiments 1 and 3, the median position exhibits several short-lived excursions away from the baseline position, quickly reverting after each. There is no reason to suppose that the spike preceding the end of the run represents a true change from this pattern. The trace from Experiment 2, in contrast, shows that the median position had gone up prior to the final step of the run, and remained relatively stable at its new position.



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

(d) Experiment 4

Figure 2.4: Median position traces

The Experiment 4 trace also shows the run achieving a new stable position, with smaller spikes away from it, including one coinciding with the terminal step. These patterns recur across the runs in each experiment that are not visualized here, including those not ending in a US victory.

A close examination of model runs suggests that the spikes – including the terminal ones – are driven less by changes in agent positions than by the changes of agent salience values. Importantly, these changes are not accompanied by a similar change in the weighted mean position – which is to say, not driven by similar changes in the positions of the agents themselves. Recall that the median is computed by checking which agent position has the highest probability of defeating all other agent positions in bilateral conflicts; thus, they are affected by the stochastic salience changes. In contrast, the mean position is weighted by (fixed) capabilities, and thus is not directly affected by salience variations.

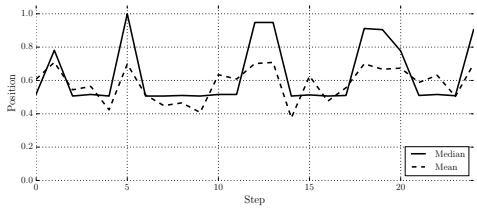
This phenomenon is illustrated in Figure 2.6, showing the individual agent positions over time for the same runs as in Figure 2.5. In this figure, each line corresponds to an agent: the y axis indicates each agent's position at each step, shown on the x axis; line widths correspond to the agents' capabilities. The top-most line is always the position of the United

States agent and the bottom-most line is the USSR agent. The traces for all the experiments show that the radical, temporary spikes in the median position do not correspond to similar sharp shifts in any agent position. In Experiment 1, the agents exhibit a slight movement towards the center of the position space. Note, however, that the movement by the USSR agent and its allies (at the bottom of the graph) are more substantial than those of the US and its aligned agents at the top. In Experiment 3, however, very little movement occurs at all. In Experiments 2 and 4, in contrast, the agents exhibit sharp, more substantial changes in their positions, which correspond to the changes in the median, highlighting that the median change is not being driven solely here by stochastic changes in salience.

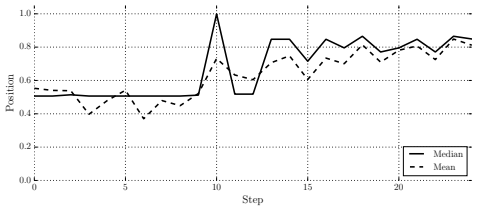
In fact, in the Experiment 2 and 4 traces, we can see a particularly interesting story emerge. Early on, several agents whose positions are closer to the median are drawn closer to the position of the United States; this is followed by several Soviet-aligned agents being drawn 'upward' towards the United States's preferred position across multiple steps. The Soviet Union itself lags behind these agents for a number of turns. However, eventually it too is driven to adjust its position upwards, in the Experiment 4 trace even eventually joining the US-aligned cluster. In effect, this model generates a recognizable notional history of the Cold War leading to a US-dominated unipolar world.

2.3.2 Predicted Conflicts

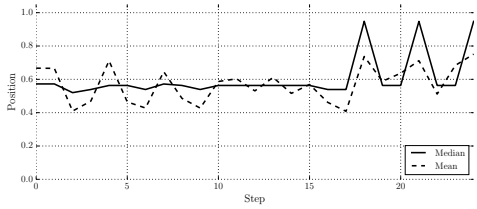
The models do not only generate overall positional traces, but simulated events, and conflicts in particular. These conflicts are an important part of the simulated history, not only because they can drive substantial changes in agent positions but because they provide useful anchor-points for comparing model runs to real history. We have robust historic data on wars and lesser militarized interstate disputes, and a well-developed understanding of international partnerships and rivalries, which allow us to assess the plausibility of the conflicts the model produces. Furthermore, conflicts are an important area for prediction and forecasting, and it is valuable to assess whether these models can be used for such prediction.



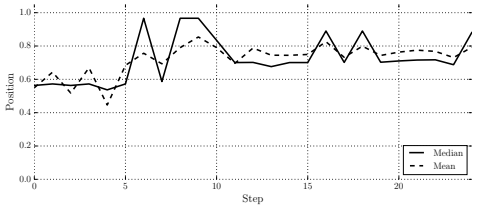
(a) Experiment 1



(b) Experiment 2

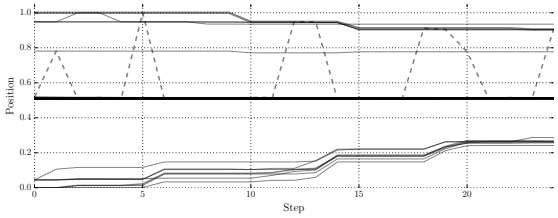


(c) Experiment 3

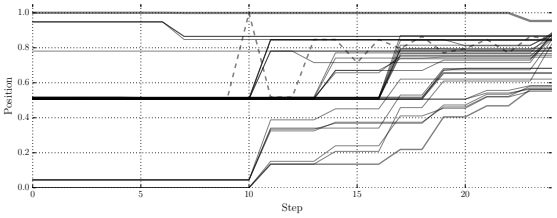


(d) Experiment 4

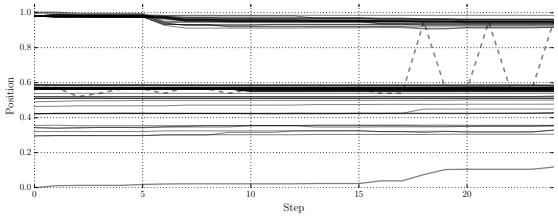
Figure 2.5: Example model traces - median positions



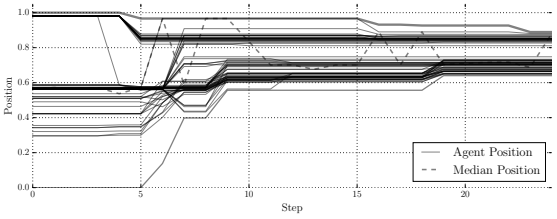
(a) Experiment 1



(b) Experiment 2



(c) Experiment 3



(d) Experiment 4

Figure 2.6: Example model traces - agent positions

For each model run in each experiment, I collect the conflict events generated across all run steps, computing the mean number of conflicts occurring between each dyad of agents across all models. These dyads are undirected, since in the model variants applied here both parties must mutually decide to start a conflict for it to occur. These simulated conflict

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Const.	1.043***	1.014***	0.392***	0.362***
	(0.183)	(0.173)	(0.049)	(0.044)
Model Conflicts	1.473	8.327***	17.409***	1015.746***
	(1.235)	(1.950)	(5.090)	(40.768)
Adjusted R^2	0.001	0.027	0.004	0.195

Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 2.5: Militarized Interstate Disputes – Linear Regressions

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Const.	-1.300***	-1.422***	-2.201***	-2.197***
	(0.101)	(0.111)	(0.067)	(0.066)
Model Conflicts	1.952***	27.658***	19.739***	654.501***
	(0.622)	(7.190)	(4.556)	(148.758)
McFadden's R^2	0.015	0.030	0.009	0.016

Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 2.6: Militarized Interstate Disputes – Logistic Regressions

counts are then merged with the count of militarized interstate disputes for each dyad, as described above.

The correlations between the predicted and observed number of conflicts are small. Most conflicts generated by the model do not reflect real conflicts, while most real conflicts are not mirrored in the simulated data. Nevertheless, we may still ask whether the model provides any statistically useful information. In order to do so, I run two regressions on each experiment output, with predicted conflicts as the independent variable: a linear regression, with observed MIDs as the dependent variable, shown in Table 2.5; and a logistic regression, with a dependent dummy variable for the presence of any MIDs on the dyad, in Table 2.6. The model generates relatively few conflicts – like real-world wars, these are rare events (**?**), meaning that the mean conflicts values are small, leading to large-magnitude coefficients.

First, let us examine the linear regressions. Note that we have no particular reason to believe that the relationship between the number of model-generated and observed conflicts

is linear. However, if a model were to predict reality perfectly, the relationship would be linear and follow the identity line. In fact, all the coefficients are positive and statistically significant, suggesting that the models are generating some useful information as to the number of observed conflicts.

Further evidence of this is visible in Table 2.6. These regressions test whether more model conflicts between dyads of actors predict *any* conflicts occurring in the relevant time period. In this case the coefficient values are positive and significant across all the models, indicating that when the model predicts more conflicts between a pair of countries, we are indeed more likely to have observed at least one conflict between them in the historic data.

While R^2 is an imperfect measure of the goodness-of-fit of a statistical model, it is an acceptable way of comparing models of the same type and the same dependent variable. The R^2 values across all models are low, indicating that they are explaining only a small fraction of the variance observed in the MID data. Nevertheless, we can see that the R^2 value on the Experiment 4 linear regression is an order of magnitude better than for the other experiments, and that this is not the case for the logistic regression. This suggests that the model is doing a better job at predicting the volume of conflicts, conditional on conflicts occuring at all. Across both types of regressions, furthermore, we can see that Experiment 3 has the least explanatory power.

It is also useful to look at the specific conflicts being predicted. Table 2.7 shows the top ten most frequently predicted conflict dyads for each experiment. One finding that stands out here is the difference between the Baseline and Updated models. The Baseline model generates no direct conflicts between the two rival superpowers in either Experiment 1 or 3, and very few even directly involving either. The conflicts are largely between lower-tier agents, with positions closer to the median. Belgium is especially over-represented in the conflicts generated in both Experiments 1 and 3. This appears to be driven by its combination of relative strenght and location close to the median (in Experiment 1, its own position is frequently the median position), making it a particularly influencial actor in the simulated international system. In contrast, the Updated model generates many conflicts

Experiment 1		Experiment 2	
England	Iraq	Bulgaria	USA
Egypt	England	USSR	USA
Belgium	France	USA	Yugoslavia
Belgium	Iran	China	USA
Argentina	USA	England	USSR
Belgium	Egypt	Australia	USA
Argentina	Brazil	China	England
Egypt	Iraq	Poland	USA
Belgium	Netherlands	Bulgaria	England
Argentina	Canada	Czechoslovakia	USA
Experiment 3		**Experiment 4**	
Egypt	Jordan	USA	USSR
France	Turkey	USA	United Kingdom
Belgium	Netherlands	USSR	Argentina
France	Iran	USA	India
Belgium	Bulgaria	USA	Yugoslavia
Afghanistan	Saudi Arabia		
Belgium	Luxembourg	USA	China
Czechoslovakia	Romania	USA	Czechoslovakia
Poland	Hungary	China	France
Turkey	Iran	USA	Portugal
		USA	Poland

Table 2.7: Top predicted conflict dyads

involving the USA and USSR agents, including frequent direct conflicts between the two.

Several of the predicted conflict dyads stand out as being particularly plausible, or implausible. The United States and Soviet Union are the most frequent dyad observed in the MID dataset for the 1948-98 period. However, these disputes never escalated to a full military conflict. Such a conflict, if it had occurred, would likely have had catastrophic consequences far outside the scope of this model. Furthermore, far from being in conflict with one another, Belgium, France, and the Netherlands were all parties to the Treaty of Brussels of 1948, which included a mutual defense guarantee (Belgium et al., 1948). Interestingly, the Baseline model generates several regional conflict dyads despite not incorporating geography, including Egypt and Iraq (Podeh, 1995) and Brazil-Argentina (Mariano, 2013).

2.3.3 Predicted Alliances

In addition to conflicts, the other detailed output the model produces is the position of each agent at each step. The space in which these positions exist is fairly abstract, particularly in this example. While in some cases positions can be assigned concrete meaning on a particular issue (e.g. number of years before a policy is adopted, as in Stokman and Bueno de Mesquita (1994)), in this case, the positions simply indicate alignment with the security interests of the United States or Soviet Union, as anchored in the starting year. The specific issues underlying these interests, however, obviously change substantially over the course of the time period we are concerned with, and assessing the position of each country across such issues is a substantial and ill-defined task in and of itself. Furthermore, both superpowers can and occasionally do change their own position as well, suggesting that they are compromising some interests in order to maximize their security with regard to the current state of the world.

One way around needing to interpret the position space is to examine not the specific positions themselves, but the distance between the positions held by two agents. Within the model, when agents' positions are closer, they are more likely to assist each other more strongly in conflicts. Furthermore, this fact is public – all other agents know it, and explicitly take it into account when estimating the probabilities of winning or losing potential conflicts. This behavior bears a clear resemblance to alliances, an enduring institution in international relations (both in practice and theory). Indeed, the Realist school of thought often holds that alliances and treaties are not meaningful in and of themselves, but only inasmuch as they reflect an underlying alignment of states' interests (Walt, 1987). With this in mind, I hypothesize that if a model is capturing aspects of real state behavior, the closer two agents are, the more likely the relevant countries are to have some sort of alliance or other formal defense or security agreement.

In order to test this hypothesis, I use the alliance data provided by the Correlates of War project (Gibler, 2009, 2013), for alliances in effect in 1998. Recall that this dataset served to generate the model inputs as well; however, I will not use alliance data from the starting

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Alliance=1				
Const	-6.828***	-8.060**	-7.387***	-8.484***
	(1.917)	(2.429)	(1.601)	(1.459)
Mean Distance	2.540	9.158	2.586	50.793**
	(4.019)	(7.667)	(3.559)	(20.673)
Alliance=2				
Const	-4.902***	-5.003***	-5.660***	-5.211***
	(0.863)	(1.327)	(0.897)	(1.270)
Mean Distance	0.044	0.624	-1.554	-37.990
	(2.472)	(6.523)	(2.921)	(56.118)
Alliance=3				
Const	-4.043***	-4.332***	-4.073***	-3.915
	(0.544)	(0.763)	(0.417)	(0.607)
Mean Distance	0.893	2.843	-2.087	-29.748
	(1.393)	(3.386)	(1.439)	(25.521)
Alliance=4				
Const	-0.974***	-0.056	-0.080	-1.2184***
	(0.150)	(0.329)	(0.086)	(0.122)
Mean Distance	-3.286***	-9.905***	-7.792***	-5.716
	(0.660)	(2.231)	(0.529)	(4.214)

Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 2.8: Alliance Type - Multinomial Choice Logistic Regression

year of 1948 (which was used to generate the input data) but from 1998, corresponding approximately to end-year of the number of steps the model has run for. Furthermore, for experiments 1 and 2 I am using the data from the original paper; as noted in Section 2.1, this appears to be substantially different enough from the most recent COW alliance dataset to reduce the risk of circular prediction.

I compute the distance between the positions of each pair of agents at the end of each run, and then find the average of these distances across each experiment. I then merge the resulting data from each experiment with the dyadic alliance data, padded to include 0 values for each pair of states for which no agreement is recorded. With these merged datasets, I run two models: a multinomial logistic regression (McFadden, 1984), and an exponential random graph model (Robins et al., 2007).

The logistic regression treats each dyad as independent, and attempts to predict which category the relationship between the two states falls into. Across all four experiment results, shown in 2.8 the mean distance between agents is not a significant predictor of ententes, non-aggression or neutrality agreements (alliance categories $1-3$, as detailed in Table 2.2). However, in Experiments $1-3$, for defense pacts (category 4, the strongest type of relationship), the model mean distance is statistically significant, with a negative sign. This means that the smaller the distance between the agents (which is to say, the closer they are), the more likely an alliance becomes. This is not the case with Experiment 4. Though the coefficient is still negative, it is not significant.

A key assumption of a logistic regression is that the observations are independent. However, we know that this is not the case with regard to interstate relationships in general, and alliances in particular. Rather, the alliances form a network, and should be studied as such (Maoz, 2010). In order to test whether the model distances are a useful predictor of the alliance network, I use exponential random graph models, a methodology developed specifically to account for such networked interdependence. The independent variable here is the simplified alliance network, where the nodes are states, connected by an edge wherever a defense pact ($Alliance = 4$) relationship exists. The key independent variable is the generated Mean Distance values, as with the logistic regression. Following Cranmer et al. (2012), I include a *2-star* coefficient (the number of pairs of edges, or trios of connected nodes) in order to capture the structural effect whereby states with more allies tend to attract additional allies (which Cranmer et al. (2012) term the 'popularity effect').

The results of this model are shown in Table 2.9. The key finding here is that the Mean Distance between actors is still a significant, and negative, predictor of an alliance between them, even when taking the network structure into account. Note that unlike with the logistic regression, here Experiment 4 Mean Distances are significant as well. However, the Experiment 4 coefficient is larger in magnitude and negative; since ERGM coefficients reflect log odds ratios, this means that the Experiment 4 results are substantially weaker than those of the other experiments.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
2-star	-0.188^{***}	-0.171^{***}	-0.023^{***}	-0.031^{***}
	(0.017)	(0.017)	(0.003)	(0.003)
Mean Distance	-2.113^{***}	-5.613^{***}	-6.308^{***}	-25.981^{***}
	(0.607)	(1.245)	(0.615)	(4.589)

Standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 2.9: Alliance Network - Exponential Random Graph Models

2.4 Discussion

This chapter set out to accomplish several goals: to attempt to reproduce the results of Bueno de Mesquita (1998); to test the model more rigorously, by examining it not only qualitatively but by using its outputs as predictors of real data; and to compare several model variants against one another, to see where they differ, how, and why, and to test which has the most explanatory or predictive power.

I was unable to reproduce the model inputs from contemporary data. This highlights two issues. One is the quality of the input data: the alliance network which I used is an updated version of the one used to generate the original input data, and as such is intended to be more accurate and correct (Gibler, 2013). In order to use the methodology presented here for forecasting, we must explicitly be aware that input data sources may be incomplete or contain errors, and if possible account for that possibility explicitly (e.g. by performing sensitivity analyses on the input data). The other issue is the need for explicit clarification of the pre-processing steps needed to get from the raw input data (in this case, the alliance network) to the actual model input (in particular, the agent positions). As shown in Sub-Section 2.2.1, the method described for generating agent positions in fact produces a two-dimensional position space; the paper does not specify the heuristic or projection used to reduce this space to one dimension.

Despite being unable to reproduce the original inputs from raw data, I was able to instantiate my model runs using the input data provided in the original paper. The overall

distribution of outcomes produced by my reproduction of the original model is qualitatively similar to the output described in the original paper, though not identical. The original paper appears to generate a wider range of outcomes, and in particular more USSR-leaning ones than the reproduction model. Furthermore, the individual run traces do not appear to be a good match to the examples presented in the original paper. While the stochastic nature of the model makes it essentially impossible to reproduce specific model traces without access to the original source code and random seed, the example traces reported in the original appear to show more movement of individual agent positions than characterize my reproduction, while the median position does not have the saw-tooth pattern which characterizes all the experiments, but in particular Experiments 1 and 3. Experiment 3, using the updated input data, generates similar behavior of the median and individual agent positions to Experiment 1. This suggests that the lack of major, stable shifts in the median position is a feature of this particular model variant, rather than the particular input configuration of agents, positions, and capabilities. Recall that as demonstrated in the previous chapter, cascades of agents backing out of offers in response to other agents accepting their own offers leads to an overall reduction in agent movement.

As I discuss above and in the previous chapter, a close examination of my reproduction model suggests that the stability of the median across its runs is driven by the rule allowing agents to renege on an accepted offer if another agent has accepted their own offer. This behavior is one of the least-specified not only in Bueno de Mesquita (1998) itself but in the other descriptions of the model as well. There are two possibilities: one is that, despite my best efforts, I have not implemented this model behavior correctly as described. The second is that the descriptions of this behavior in the prior material do not accurately reflect the behavior of the underlying source code. As the original code is not available to inspect, it is impossible to verify directly which possibility is correct.

Nevertheless, there is at least anecdotal evidence that the flaw is not in my replication but in the original description. Private communications with scholars familiar with the original paper have suggested that the experiments it describes were implemented using the

commercial version of the model software, developed by BDM's consulting firm, Decision Insights, Incorporated (DII). This software was the subject of litigation, during which Gary Slack, a DII employee who is thanked for "invaluable programming assistance" in Bueno de Mesquita (1998), testified that the model contains proprietary methods not disclosed in the public literature (4th Circuit, 2011). This suggests, in turn, that the behavior described in the original paper, and the prior literature more broadly, is incorrect or incomplete. In this case, the difference between the described and observed model appears to be attributable to a specific model rule, I believe that rule is indeed a key place where the public and proprietary versions of the model diverge. In fact, the traces in Experiment 2, where this rule is not applied, appear to be qualitatively more similar to the traces reported in the original paper than Experiment 1.

While replication is certainly important, a more interesting question is how well the models reproduce the observed history of the Cold War and its end. Historically, while the end of the Cold War was sudden and unexpected, it was not a short-lived phenomenon, at least on the time-scale we are modeling here. In fact, it was followed by at least a decade of 'unipolarity' during which many countries formerly oriented towards the USSR, and often allied with it, rapidly shifted their international position towards the United States and its allies (Wohlforth, 1999). The most notable demonstration of the long-term transition in the international geopolitical orientation is was the countries who were formerly members of the Soviet-led Warsaw Pact alliance joining NATO, the US-led alliance which had been the Warsaw Pact's primary adversary (Waltz, 2000). Examination of the example model traces shown in Figure 2.6, and other traces not shown here, indicate that the replication model of Experiments 1 and 3 does not tend to produce such sharp, stable transitions. In contrast, Experiments 2 and 4 do feature relatively sharp changes, with Soviet-aligned agents moving towards the United States followed by the Soviet Union itself. Qualitatively, at least, these models appear to be generating behaviors which mirror the historic record. Note, however, that the majority of model runs across all experiments do not follow this behavior, and indeed do not generate a 'victory' to one superpower or the other.

None of the experiments presented here appear to be be powerful predictors of observed conflicts. While the coefficients are statistically significant, they are nevertheless only weak predictors of observed militarized disputes in the relevant historic period. There are several possible explanations here. One is that model conflicts do not play the same role as do MIDs in the international system, at least at the scale and with the issues being examined here. States use force to attempt to threaten one another to accede to particular demands, or to deter or prevent them from taking a particular action. A minority of conflicts do end with one side completely changing its orientation – for example, the Second World War led to Japan being coerced from being an enemy of the United States to its ally (Schaller, 1997), and Tanzania's invasion of Uganda in 1978-79 successfully replaced the regime of Idi Amin with one friendly to Tanzania's interests (Acheson-Brown, 2001). However, in the majority of cases, even if one side has successfully forced concessions from the other side (such as in the Falklands or Gulf Wars), the other side's overall geopolitical orientation remains largely unchanged; certainly, they do not become allies of the power which defeated them. Furthermore, the model's single dimension means it is most likely to generate conflicts driven by the overall 'issue' – in this case, the competition between the two superpowers. However, many conflicts arise due to issues particular to a state, dyad or region (Senese, 1996). If these issues are not embedded in the initial position-generating alliance network, the model will obviously be unable to account for them, and thus will not generate conflicts driven by them. Finally, the model lacks a spatial component. Many conflicts are between neighbors, and may be driven at least in part by disputes arising due to that relationship – with territory being the most obvious example, though not the only one. The ability of other states to assist one side or another in a conflict is also influenced by geography: in terms of sending military forces or assistance, but also in terms of 'softer' power. For example, nearby states trade more than far-away ones (Ramanarayanan, 2011), meaning that overall, a state's ability to affect another by increasing or decreasing trade and commerce is also tied to the distance between them.

Similarly, the models are significant but not powerful predictors of alliances. They

appear most capable of predicting defensive alliances, the strongest category: the closer the final positions of two agents are, the more likely they are to have a defensive pact between them, committing one to come to the other's aid if attacked. Of the alliance types coded in the COW data, defensive pacts most closely mirror the behavior coded in the model itself, with agents more likely to contribute more resources to a closer agent when that agent is involved in a conflict – and this fact, like defensive pacts, is public common knowledge. This correspondence suggest that the model is indeed capturing one dimension of states' decisionmaking. Importantly, the dyad-wise position distances are a stronger predictor in the Experiments 2 and 4, where the agents tend to change positions more than in Experiments 1 and 3. This indicates that the dynamics of the models are generating useful information not directly present in the initial data.

We have already touched on comparing the experiments to one another in the various sections above. A common thread is that the Updated model outperforms the Baseline model: the agent behaviors it generates are more qualitatively plausible, and the conflicts it generates are more predictive of real conflicts across both input sets. Note that the one exception is the prediction of alliances in Experiment 4. Though this experiment generates the best predictions of conflicts, it also produces the worst predictions of alliances. It is somewhat surprising that these two metrics do not go together, since we would hope that the most accurate model of the system would generate strong predictions for both. One possible explanation is suggested by the fact that Experiment 4 generates the fewest outcomes with one side or another gaining a decisive advantage. Nearly all of its outcomes involve the Cold War continuing indefinitely, and (as Figure 2.4 shows) becoming more stable. In contrast, the target alliance network is from 1998, reflecting a decisively post-Cold War world. Thus, the experiment may be predicting conflicts likely to occur in a Cold War context, which accounts for the bulk of the time (and the conflicts) being studied; however, since it does not predict an end to the Cold War, it fails to predict the alliance network which follows. This does not, however, explain the qualitatively different behavior between Experiments 2 and 4, using the same model with different input data. There does not appear to be a

simple, reduced explanation for this difference. Rather, the particular configuration of the two input data sets yield substantially different distributions of outcomes. This ought to raise some concerns as to these models as well, as it highlights that different input data sets attempting to capture the same underlying system can yield substantially different behaviors and results.

Finally, let us pull back to examine the original question which motivated Bueno de Mesquita (1998): whether the end of the Cold War could be explained by contemporary theories of international relations, as operationalized by the Expected Utility Model. The answer here appears to be a qualified 'yes'. Despite the differences in the behavior and dynamics of the different model variants and input data, all experiments generate an overall prediction of an advantage to the United States, despite the initial balance of power being nearly completely even. This suggests that the initial hypothesis is correct, and that the end of the Cold War was potentially predictable from much earlier data. In fact, this provides more evidence for the robustness of the US advantage, as it recurs across several model variants, indicating it is not simply an artifact of one particular model implementation. Nevertheless, these results also suggest the caution we must exercise in using this model for prediction. While the input data may have been knowable in the starting year, we necessarily used knowledge from future years to assess the model outputs, and determine that the conflicts were non-predictive while the positions were. More research is required to assess whether these are general features of the model, or outcomes specific to this particular case or system scale.

Bibliography

4th Circuit (2011, March). Decision Insights, Inc. v. Sentia Group, Inc., No 09-2300.

Acheson-Brown, D. G. (2001). The Tanzanian Invasion of Uganda: A Just war? *International Third World Studies Journal and Review 12*, 1–11.

Belgium, France, Luxembourg, The Netherlands, and United Kingdom (1948, March). Treaty of Economic, Social and Cultural Collaboration and Collective Self-Defence.

Black, D. (1948). On the rationale of group decision-making. *The Journal of Political Economy*, 23–34.

Boschee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward (2015). ICEWS Coded Event Data.

Bratton, P. C. (2005). When Is Coercion Successful? And Why Can't We Agree on It? Technical report, DTIC Document.

Bueno de Mesquita, B. (1975). Measuring Systemic Polarity. *The Journal of Conflict Resolution 19*(2), 187–216.

Bueno de Mesquita, B. (1984). Forecasting Policy Decisions: An Expected Utility Approach to Post-Khomeini Iran. *PS: Political Science & Politics 17*(02), 226–236.

Bueno De Mesquita, B. (1985, March). The War Trap Revisited: A Revised Expected Utility Model. *The American Political Science Review 79*(1), 156–177.

Bueno De Mesquita, B. (1994). Political forecasting: an expected utility method. In F. N. Stokman and B. Bueno De Mesquita (Eds.), *European Community Decision Making:*

Models, Applications and Comparisons, pp. 71–104. New Haven and London: Yale University Press.

Bueno de Mesquita, B. (1997, December). A decision making model: Its structure and form. *International Interactions 23*(3-4), 235–266.

Bueno de Mesquita, B. (1998). The End of the Cold War Predicting an Emergent Property. *Journal of Conflict Resolution 42*(2), 131–155.

Bueno de Mesquita, B. (2002, June). *Predicting Politics* (1st edition ed.). Columbus: The Ohio State University Press.

Bueno de Mesquita, B. (2010). *The Predictioneer's Game: Using the logic of brazen self-interest to see and shape the future*. Random House LLC.

Comer, K. W. (2014, October). *Who Goes First? An Examination of the Impact of Activation on Outcome Behavior in Agent-based Models*. Ph. D. thesis.

Correlates of War Project (2010). National Material Capabilties Dataset, v4.0.

Cranmer, S. J., B. A. Desmarais, and E. J. Menninga (2012). Complex dependencies in the alliance network. *Conflict Management and Peace Science 29*(3), 279–313.

Dobson, A. P. and S. Marsh (2007, January). *US Foreign Policy Since 1945*. Routledge.

Gaddis, J. L. (1992). International Relations Theory and the End of the Cold War. *International Security 17*(3), 5.

Gibler, D. M. (2009). *International military alliances, 1648-2008*. Correlates of war series. Washington, D.C: CQ Press.

Gibler, D. M. (2013, March). Release Notes for Version 4: Correlates of War Formal Interstate Alliance Dataset, 1816-2012.

Gibler, D. M. and M. R. Sarkees (2004). Measuring alliances: The correlates of war formal interstate alliance dataset, 18162000. *Journal of Peace Research 41*(2), 211–222.

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, and G. Huse (2006). A standard protocol for describing individual-based and agent-based models. *Ecological modelling 198*(1), 115–126.

Hendrix, C. S. and Idean Salehyan (2013). Social Conflict in Africa Database (SCAD).

Heuer, R. J. and Center for the Study of Intelligence (U.S.) (2001). *Psychology of intelligence analysis*. Wahington, D.C.: Center for the Study of Intelligence, Central Intelligence Agency.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association 97*(460), 1090–1098.

Jones, D. M., S. A. Bremer, and J. D. Singer (1996). Militarized interstate disputes, 18161992: Rationale, coding rules, and empirical patterns. *Conflict Management and Peace Science 15*(2), 163–213.

Kimbrough, S. O. (2014, October). Making Predictions: The Group Decision Forecasting (GDF) Problem. SSRN Scholarly Paper ID 2506969, Social Science Research Network, Rochester, NY.

Lebow, R. N. and T. Risse-Kappen (1995, January). *International Relations Theory and the End of the Cold War*. Columbia University Press.

Leng, R. J. and J. D. Singer (1988). Militarized interstate crises: The BCOW typology and its applications. *International Studies Quarterly*, 155–173.

Maoz, Z. (2010). *Networks of nations: The evolution, structure, and impact of international networks, 18162001*, Volume 32. Cambridge University Press.

Mariano, K. L. P. (2013). *Two to tango: an analysis Brazilian-Argentine relations*. SciELO Brasil.

McFadden, D. (1984). Econometric analysis of qualitative response models. Handbook of Econometrics, Elsevier.

McKibben-Sanders, J. (2014). Bdm scholz expected utility model.

Podeh, E. (1995). *The Quest for Hegemony in the Arab World: The Struggle Over the Baghdad Pact*. Brill.

Prados, J. (2011). *How the Cold War Ended: Debating and Doing History*. Potomac Books, Inc.

Ramanarayanan, A. (2011, July). Distance and the impact of gravity help explain patterns of international trade. *Economic Letter 6*(7).

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007, May). An introduction to exponential random graph (p*) models for social networks. *Social Networks 29*(2), 173–191.

Schaller, M. (1997). *Altered states: The United States and Japan since the occupation*. Oxford University Press.

Scholz, J. B., G. J. Calbert, and G. A. Smith (2011, October). Unravelling Bueno de Mesquitas group decision model. *Journal of Theoretical Politics 23*(4), 510–531.

Schrodt, P. A., E. M. Simpson, and D. J. Gerner (2001). Monitoring conflict using automated coding of newswire reports: a comparison of five geographical regions. In *Conference Identifying Wars: Systematic Conflict Research and its Utility in Conflict Resolution and Prevention, Uppsala*, pp. 8–9.

Senese, P. D. (1996). Geographical proximity and issue salience: Their effects on the escalation of militarized interstate conflict. *Conflict Management and Peace Science 15*(2), 133–161.

Signorino, C. S. and J. M. Ritter (1999, March). Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions. *International Studies Quarterly 43*(1), 115–144.

Stokman, F. N. and B. Bueno de Mesquita (Eds.) (1994). *European Community Decision Making: Models*. Applications, and Comparisons: Yale University Press.

Tipping, M. E. and C. M. Bishop (1999). Mixtures of probabilistic principal component analyzers. *Neural computation 11*(2), 443–482.

Walt, S. M. (1987, November). *The Origins of Alliances*. Ithaca: Cornell Univ Pr.

Waltz, K. N. (2000, August). NATO expansion: A realist's view. *Contemporary Security Policy 21*(2), 23–38.

Ward, M. D., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford (2013). Comparing GDELT and ICEWS event data. *Analysis 21*, 267–297.

Wohlforth, W. C. (1999, July). The Stability of a Unipolar World. *International Security 24*(1), 5–41.

Yonamine, J. E. (2011). Working with event data: A guide to aggregation choices. *Penn State University: Working Paper*.