Predicting International Conflict and Revision

Daniel Kent[*]

June 17, 2020

Abstract

When do revisionist states become revisionists? Most responses to this question fall under one of three categories: 1) differential growth rates, 2) domestic political changes, or 3) international dissatisfaction. While these arguments are all based on rich research traditions, it is an open question as to which theory best predicts when a state will turn to international revision. In order to provide such an empirical comparison, I build a series of machine learning ensembles that predict interstate conflict onset and vary only in the included features. With feature sets corresponding to separate theories, I evaluate each theory on identical empirical terms, comparing only their corresponding models' predictive accuracy in test sets. While the models for each of the three theories unsurprisingly demonstrate meaningful predictive capacity, the ensemble based on measures of international dissatisfaction is more accurate than counterparts based on differential growth rates and domestic political changes. The paper's results call for a greater focus on a state's standing within the broader international system when attempting to predict when states will or will not become revisionists. Revision does not just follow changes from within the revisionst, but also changes in its external relations and the willingness of other states to form strategically valuable bonds with it.

[*] kent.249@osu.edu; Ph.D. Candidate, Department of Political Science, The Ohio State University

1 Introduction

Political decision-makers face a choice at all times between pushing their country in a direction that supports or seeks to disrupt the international status quo. This fundamental choice between international revision and status quo seeking is widely understood to be highly consequential.[1] The world would be vastly different if Germany had been satisfied with the European order before either of the World Wars or if Revolutionary France did not turn its guns outward. In this sense, the typology of status quo and revisionist states is not just a conceptual exercise used to simplify a complicated world; it is a categorization that captures one of the defining variables in international politics. Whether it be Napoleonic France, Revolutionary Iran, Communist Russia, Imperial Japan, or 20th-century Germany, revisionist states have underscored the outbreak of some of history's bloodiest wars and costliest military competitions.

Indeed, one of the most discussed questions in modern international politics is whether China's long-term aims are likely to follow a revisionist or status quo path. (E.g. Christensen, 2006; Goh, 2013; Johnston and Ross, 1999; Johnston, 2003; Paul, 2016; Schweller, 1999) If China's strategic trajectory involves upending the current international order, then the prescriptions that follow diverge widely from those toward a China that intends to maintain the status quo which enabled its rise. Accurate predictions about which path China is most likely follow under current and alternative conditions are therefore a policy necessity. So when do countries – particularly revisionists – fall where they do on this spectrum? Put differently, when do countries sometimes attempt to forcefully undermine or alter the international system, rather than invest in and support it?

In this paper I present an empirical comparison of the three most prominent arguments about revision's origins. Generally speaking, arguments fall under one of three themes: 1) rising powers and differential growth rates, 2) domestic political shifts and revolutionary regimes, or 3) international dissatisfaction. Each of these arguments is backed by various historical cases

[1] The impact of this choice can be seen in the lengthy literature discussing the role of status-quo and revisionist states in international politics. (E.g., Carr (1946, 103-105), Morgenthau (1948, 40-74), Wolfers (1962, 81-102), Organski and Kugler (1981), Schweller (1994), Lyall (2005), and Goddard (2018b))

and is statistically associated with interstate conflict initiation. However, their relative ability to inform accurate policy predictions is an open question. Put differently, *it is unclear which theory provides the most predictive accuracy*, even though we have reason to expect that all do predict to some extent. Or, if we know about a state's domestic political changes, growth rates, and international dissatisfaction, then how much of a state's revisionist behavior can we predict with each variable and which variable informs the most accurate predictions?

Herein I compare the three theories empirically, building stacked ensembles of machine learning models that vary only in the features included for prediction. The target (outcome or dependent variable) is whether or not a state initiates a militarized interstate dispute in a given year. In order to compare each theory empirically, I measure and compare each ensemble's predictive accuracy in test sets which did not inform model training. Across almost all time periods, the ensemble based on international dissatisfaction predicts with the highest accuracy in test sets whose data did not inform model training. Notably, the other two ensembles also predict with reasonable accuracy in test sets, meaning the variables associated with domestic political changes and rising powers do provide meaningful, even if less, predictive capacity of a state's revisionist tendencies.

The remainder of the paper proceeds as follows. First, I overview the literature on revisionist states, discussing the three aforementioned classes of arguments in detail and setting up their empirical operationalizations. Next, I discuss how features were engineered from the available data and the methodological approach behind the machine learning models employed, emphasizing that this is a predictive, rather than a causal approach. This is followed by a presentation of model results, with an emphasis on moving from a "black-box" view of machine learning models to one where results can be interpreted in a substantively useful way, speaking to the theories that inform the variables chosen for each model. (Karpatne et al., 2017; Radford and Joseph, 2020) I close with a discussion of implications that follow from the paper's results regarding China and other great powers in the current era. Given the paper's results, I also propose potential avenues for future research – emphasizing an understanding of why some states find themselves relatively excluded from valued international institutions and communities while others are able to deeply embed themselves with general ease. The paper's predictive focus leaves it agnostic on the causal question of what leads to exclusion and the resulting dissatisfaction. However, dissatisfaction's predictive accuracy when it comes to interstate conflict emphasizes the need for future work to investigate why dissatisfaction occurs.

2 Theories of International Revision and Conflict

Prominent arguments about revisionist states tend to fall under one of three categories: rising powers, domestic political shocks, and international dissatisfaction. In other words, discussions of revision tend to stem from, respectively, the threat of a rising power, states with a drastically changing domestic political landscape, or an international system whose characteristics are viewed as untenable by certain key members. While these themes are investigated and argued to impact states in numerous ways, few, if any, arguments regarding international revision fall outside of the three. Below, I start with the logic of rising powers and international revision before turning to discussions of domestic politics and then closing with international dissatisfaction.

Starting with Thucydides[2] and generally discussed in terms of Organski and Kugler (1981), power transition theory links international revision to rising powers who threaten using their newfound or forthcoming capabilities to forcefully reshape the international system.[3] As a rising power grows stronger it threatens to overtake an established power's position as the systemic or regional hegemon. A commitment problem is found at the heart of power transition theory and is the reason why power transitions are argued to produce war (Fearon, 1995).

[2] See the recently re-popularized discussion of the "Thucydides Trap" in Allison (2017), where the core claim being applied to the US-China relationship is: "It was the rise of Athens, and the fear that this instilled in Sparta, that made war inevitable."

[3] Additional treatments include, but are not limited to: Christensen (2006); Duffy Toft (2007); Lebow and Valentino (2009); Kennedy (2010); Mearsheimer (2014); Dafoe et al. (2014).

Rising powers cannot credibly commit to using their new capabilities in a future manner that is acceptable for the established powers, because the state's future intentions and leadership are largely unknown, especially as one's strategic time horizon's extend further into the future (Edelstein, 2017; Tingley, 2011). Even if relations with the rising power are currently healthy for many states, the future is largely unknown and could very plausibly change radically to more conflictual relationships.

When this uncertainty about future intentions and a state's inability to commit indefinitely to any course of action are coupled with an impending power transition, then states find themselves in a uniquely precarious situation. Established powers may make a strategic calculus that starting a war under current conditions – where they are still more powerful than the rising power – is considered a better option than living in future subservience after a transition, despite the unavoidable costs in blood and treasure that come with war. Accordingly, a war-avoiding bargain between established and rising powers cannot be reached. Moreover, while this process is generally discussed in the context of power transitions – where the stakes of the commitment problem are most severe – the logic of rising powers applies to any situation where one state's growth rate is greater than its actual or potential competitors' growth rates. Even without a power-transition, if a state is growing rapidly and expects to continue growing, then it can use its newfound capabilities to reshape its environment in a way that puts it in a position of taking further advantage of other states in the future. Subsequent work has delved into the conditions under which these commitment problems are more or less severe, with an emphasis on: the number of other competitor states to consider (Shifrinson, 2018; Snidal, 1991), the compatibility of interests between actors (Schake, 2017), the strategic time horizons of key leadership (Edelstein, 2017; Tingley, 2011), and the relative plausibility of an actual power transition (Beckley, 2012, 2018; Kadera, 2001).

Secondarily, much of the relevant literature focuses on the domestic actors who decide to pursue a strategy of international revision. Revisionist policies are debated, decided upon, and implemented from within a country. So it follows to ask why decision-makers within certain domestic political systems and regimes may be more likely than their counterparts in other countries to pursue international revision. On this point, many famous revisionist states are inescapably linked to their unique domestic political environment at the time and are rarely discussed without mention of the domestic politics and leadership at the time of revision. Napoleonic France, Hitlerite Germany, Revolutionary Iran, Imperial Japan, and Communist Russia are core examples of revisionist states and almost always referred to in the context of these domestic labels. For these examples, their revisionist behavior can be viewed as intertwined with the radical domestic changes that preceded a decision to attempt widespread revision of their international environments. Under a counterfactual hypothetical where the international environment is the same but domestic leadership differs for these countries, revision appears far less likely.

Beyond the correlation between revision and major domestic changes in these historical cases, conceptually it is an appealing argument that political elites who are seeking radical political change will not just stop at home, but see similar grievances and opportunities abroad. Once major political changes have been profitable for instrumental elite actors at home, the next step may very well be to look for opportunities for further similar changes abroad. However, in response, outside actors may also see radical domestic changes within their neighboring states and grow fearful, creating an increasingly tense international environment which is ripe to spiral into interstate conflict. (Walt, 1996) Taken as a whole, whether it be the rise of militaristic regimes and leadership[4], the prominence of overexpansive and snowballing political projects,[5], or domestic revolutions and civil wars turning their focus abroad[6], various mechanisms have been theorized to link sharp domestic political changes to subsequent international revision.[7]

[4] Horowitz et al. (2015); Lemke and Reed (1996); Schweller (1994, 1999, 2015); Snyder (1984); Van Evera (1984); Weeks (2008)

[5] Davidson (2006); Lyall (2005); Snyder (1991)

[6] Lawson (2015); Schroeder (1994); Walt (1996)

[7] On the China debate, given the prominence of the CCP and Xi Jinping's leadership in political discussions, domestic theories of international conflict and revision are potentially of primary importance for current policy options.

Lastly, by definition, for revision to occur, there must be an environment to revise. The final class of arguments focuses on the structural makeup of each state's international environment. Here, the question is: what structural conditions are most prone to push states toward attempting to revise the international system for their benefit, despite the massive risks? These arguments can be generally cast as discussions of international dissatisfaction, where a state pursues revision because it views some aspect(s) of the international system as stifling and unacceptable. Notably, this class of arguments is amenable to both rationalist and more constructivist or sociological explanations. Across the literature, approaches span from identity (Thies and Nieman, 2017), linguistic (Goddard, 2018a), norms (Finnemore and Sikkink, 1998), relational (Jackson and Nexon, 1999; Qin, 2016; MacDonald, 2014), and status concerns[8] to more calculated cost-benefit analyses.[9] For this literature the greatest concern should be directed toward states with the capacity to coercively create widespread international change – whether or not they are concurrently a rising power – with longstanding international grievances and/or reasons to think that its environment can feasibly be reshaped in a way that better suits its interests. In other words, large dissatisfied states are the most dangerous and likely candidates for deciding they want to pursue a strategy of widespread international revision and conflict.

Across all of the aforementioned theories and arguments, statistically significant associations and detailed case-studies are abundant. In this sense, all variables of interest likely matter and are present in seminal cases to some degree. However, saying that many variables have some importance and can predict revisionist states is not the same as saying their predictive capacities are of equal magnitude. While that may be the case, it seems more likely that the mechanisms precede revision at differential rates. Furthermore, when it comes to projections and policy prescriptions about current and future potential revisionists, knowing which theoretical pathway has the greatest empirical support is vital. In the following sections I provide an empirical composition and comparison of each. After presenting empirical operationaliza-

[8] Chan (2004); Deng (2008); Duque (2018); Goh (2013); Larson and Shevchenko (2010); Paul et al. (2014); Renshon (2016, 2017); Ward (2017); Wolf (2011)

[9] Braumoeller (2013); Holsti et al. (2019); Gilpin (1983); Goddard (2018b); MacDonald and Parent (2018); Montgomery (2016); Morgenthau (1948); Carr (1946); Wolfers (1962); Trachtenberg (2012); Lipscy (2017)

tions of the general arguments, I evaluate each theory by comparing predictive accuracy across identical statistical frameworks and data.

3 Data

The outcome of interest for this study is whether or not a country initiates a reciprocated militarized interstate dispute (MID) in a given year. I treat whether or not a state initiates a MID in a given year as the dependent variable because a defining characteristic of revisionist states is that they are especially conflict-prone, relative to other states. This is not to say that all interstate conflicts are revisionist in nature,[10] but models meant to capture the underlying causes of revision should predict conflict onset well. Indeed, if variables meant to captures revision's origins do not accurately predict interstate conflict when modelled statistically, then this would put arguments that those variables are behind most cases of revision on shaky footing. Moreover, in the subsequent section with model results, when a model is fit with all possible features, controlling for other sources of conflict, international dissatisfaction is given primary importance by the model – generally providing the greatest increase in predictive accuracy across algorithmic iterations. Accordingly, my focus is on finding which variable both best predicts conflict onset when modelled on its own *and* alongside other variables also theorized to precede international revision. A variable which best predicts interstate conflict both on its own and alongside other theoretically-informed variables is likely going to best predict the emergence of revisionist states.[11]

However, I do not use all MIDs as my dependent variable. Rather, per Braumoeller (2019), employing reciprocated MIDs strikes a useful balance between modeling all MIDs and only looking at cases which resulted in a fatality. Because the MID dataset (Palmer et al., 2019)

[10] On this note, the subsequent models only predicting a portion of all MIDs, which we should expect because the following models are not meant to be the models of conflict. Rather, we are asking: For the variables that are understood to often precede revision, which best predicts conflict?

[11] I make this claim recognizing that interstate conflicts also occur for a variety of non-revisionist reasons. Seeing dissatisfaction's subsequent predictive capacity for conflict, despite including an array of other features for prediction, provides greater confidence in the results speaking directly to international revision.

includes cases that range from low-level threats of force to interstate wars, it is easy to lump together non-threatening posturing with intense outbreaks of military conflicts. While fatal MIDs are likely too strict of a cutoff, eliminating cases that could have easily escalated further but fortunately did not, using all MIDs is likely too lenient of a standard, including low-level incidents such as fishing trawlers entering into territorial waters.[12] Instead, as Braumoeller (2019) points out, if the use of force is reciprocated, then both sides have demonstrated a willingness and ability to escalate.[13] In terms of revisionist states, as discussed above, while not all uses of force are revisionist in nature, revisionist states are especially prone to military action, so looking at a state's propensity to initiate reciprocated MIDs is an effective proxy metric. Looking at the data, of the 15925 total country-years, 3167 include the onset of a reciprocated MID. This leaves us with the common problem of class imbalance, where the dependent variable is mostly zeros – risking models that predict almost entirely zeros and returning often accurate but generally unhelpful predictions. While the following models are able to make accurate predictions without techniques such as downsampling or upsampling, I ultimately turn to downsampling because it provides the highest predictive accuracy across the entire training set, cross-validation folds, and test set.

Turning to features, each of the aforementioned theories is represented by a core variable, which are then each transformed in a set of identical ways. The three core variables are a state's international dissatisfaction, expected international benefits, and its domestic political regime. The latter two variables are then first-differenced in order to represent rising powers and domestic political changes.[14] As I discuss subsequently, estimates for a state's international expectations are based on its observable capabilities and general reputation for effectively us-

[12] That said, Figure 9 in the appendix includes model results when fatal MIDs are used as the dependent variable and the results are consistent with reciprocated MIDs. Also in Figure 10 I show model results when the DV is made even more fine-grained and limited to wars – or MIDs with 1000+ casualties. For these models, dissatisfaction also demonstrates the generally strongest predictive capacity.

[13] Moreover, as Braumoeller (2019) also points out, interstate conflicts can escalate very quickly in dramatic fashion. So just because a reciprocated MID did not escalate beyond low-levels of force does not, by any means, suggest that further escalation was impossible.

[14] These variables all cover very different processes. Figure 1 displays the distributions of each component variable and their correlations, where the strongest correlation between any two features is below 0.5.

ing its capabilities, both of which then inform how much a state expects it should be able to influence and benefit from the international system. Taking the first difference of these expectations therefore captures whether a state is rising, declining, or staying at its current size. A state's international dissatisfaction in a given year is operationalized as the difference between its expected international benefits and actual international benefits. The more a state's actual access to valued international goods and social recognition falls short of what it believes it should be receiving based on its relative capabilities, then the more that state will be dissatisfied with the status quo. The more dissatisfied a state is, then the more we expect it to start interstate conflicts. Put generally, dissatisfaction for state i in year t is expressed as:

$$\text{Dissatisfaction}_{it} = \log(\text{Expected Benefits})_{it} - \log(\text{Actual Benefits})_{it}. \qquad (1)$$

Both expected and actual benefits are logged because the variables – which I subsequently break down – are heavily right skewed, so logging them provides us with approximately normal distributions.[15] Both a state's expected and actual benefits are variables that I construct and require feature engineering with multiple readily available datasets.

Starting with the measure of actual international benefits, the actual benefits for state i in year t can be expressed as:

$$\text{Actual Benefits}_{it} = \frac{1}{N}\sum_{n=1}^{N} \text{PageRank}_{int} \qquad (2)$$

with PageRank_{int} referring to state i's PageRank centrality in year t in a network n (including military alliances, interstate trade, shared diplomatic relations, and arms transfers), which is then averaged to provide a state's general access to valued international goods in a given year. The idea behind using network centrality to measure actual benefits is that prominence in these valued networks provides access to the social and material goods that constitute the potential profits from international politics. The PageRank centrality formula (Brin and Page, 1998, 2012) estimates a node's prominence in a network, which reflects the sum of a node's ties,

[15] This of course builds in an assumption that working with normally distributed data is preferable to skewed data and makes little to no difference in our final substantive interpretations.

with ties to other prominent nodes weighted more heavily than ties to less prominent nodes.[16]

$$x_i = \alpha_A \sum_j a_{ij} \frac{x_j}{g_j} + (1 - \alpha_A) \frac{1}{N} \tag{3}$$

where $g_j = \sum_i a_{ji}$ (node j's total number of ties), x_i is the PageRank centrality for node i, and x_j is the PageRank centrality for node j.[17] a_{ij} is equal to 1 if a tie exists between nodes i and j, but it is equal to 0 otherwise. α_A is a constant damping factor that weighs how much a node's ties matter, as opposed to treating each node as equally central (if $\alpha_A = 1$); in the context of Google's search engine the damping factor approximates the probability that a user will stop clicking links at any time.

A state's expected benefits are produced by multiplying its CINC score (the state's share of the entire globe's observable material capabilities) by its average predicted probability of winning a MID against other states (a proxy for the state's perceived capabilities):

$$\text{Expected Benefits}_{it} = \text{CINC}_{it} \times \frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J - 1}. \tag{4}$$

While CINC scores captures a state's percentage of the globe's material capabilities – such as iron and steel production and total population – the reputational sources of power are much more difficult to model. Per Gilpin (1983), I decompose power into observed material capabilities and a state's reputation for using those capabilities. In order to model the latter, I build on Carroll and Kenkel (2016) and train a machine learning model on MID outcomes, building a model that predicts which state wins interstate conflicts based on every involved state's observed CINC components. After accurately classifying observed MID outcomes, I then apply the model to make predictions about which state would win in all possible MIDs. A state's reputation for power – $\frac{\sum_{j=1}^{J} \Pr(\text{Win}_j)}{J-1}$ – is then the sum of a state j's predicted probabilities of winning

[16] The original idea was that the internet can be viewed as a network where nodes are sites and ties are links to one site that are included on another site. Google's search engine then ranks results based upon a version of PageRank centrality, where a site's ranking on the search algorithm is based upon its PageRank centrality.

[17] If one carefully reads the formula, then they will see that calculating a node's PageRank centrality requires having already estimated every other node's PageRank centrality. This produces a chicken-egg problem where there is no obvious answer about how to estimate initial values. For a mathematically detailed explanation of how this problem is overcome see: http://www.ams.org/publicoutreach/feature-column/fcarc-pagerank.

across all possible MIDs, divided by the number of other states in the international system in a given year.

Combining these two variables produces a measure of expected international benefits which weighs a state's observed capabilities by its reputation for the capacity and willingness to use its capabilities. This combined measure represents how powerful a state is, relative to its peers, which then drives how much that state expects it should be receiving from the international system at any time. The more powerful a state is, then the more it should expect the system will be designed in a way that benefits it. Inversely, the less powerful a state is, then the lower its expectations will be, because it lacks the coercive capacity to meaningfully shape its international environment. Taken in comparison to its actual international benefits, the difference between a state's expected and actual international benefits therefore represents the extent to which a state believes the international system is currently depriving it of benefits that it could otherwise gain through a coercive bargaining scenario. This expectations-reality gap then creates an incentive for using military force to remedy these perceived systemic shortcomings, which otherwise appear to be persistent.

Moreover, because a state's expected international benefits represent its relative power – or share of global capabilities – the measure maps on cleanly to theories of rising powers. Because rising powers occur when a state is growing faster than its peers, if a state's share of global capabilities grows, then it by definition is growing faster than its peers and is a rising power. Accordingly, our measure of expected benefits is well-suited as a starting point for operationalizing whether or not a state is rising and the extent of its differential growth rate. More formally, we can capture differential growth rates by first-differencing a state's expected benefits – calculating the difference between a state's expected benefits in year t and year $t-1$. If the difference is positive, then the state is rising by that extent. Admittedly, differential growth rates are of most concern for international revision when the great powers are growing and this empirical measure only compares growth at time t and $t-1$, regardless of a state's initial size. While this is conceptually a concern, in terms of this measure – where a state's size is represented in terms of its share of the entire international system's capabilities – large growth rates only tend to occur within states that are already larger than most and have the capacity to expand relative to other states. This point is demonstrated in Figure 10 in the appendix. Returning to operationalizing growth rates, if the difference is negative, then that state is shrinking relative to its peers. While rising powers are often discussed as a binary category, this approach gives us a finer-grained continuous measure of differential growth rates:

$$\text{Differential Growth}_{it} = \text{Expected Benefits}_{it} - \text{Expected Benefits}_{it-1}. \tag{5}$$

Lastly, I consider three variables for relevant theories of domestic politics and international revision, where the primary variable is a state's polity score (Marshall et al., 2002), because of its conceptual utility and availability for all states in all years. Polity scores are used to capture regime type and span from fully autocratic (-10) to fully democratic (10). Similar to rising powers, domestic theories of revision are generally linked to drastic changes to a state's domestic political environment, so like rising powers a use a first-differencing strategy to capture changes in a state's regime type from one year to the next. However, adding some conceptual richness, drawing on theories of revolution and war, I include variables for whether or not a civil war began or if a civil war is ongoing in a country in a given year. While all of our other data starts in 1816, data on civil wars starts in 1945 and is subsequently only considered in models military conflict after World War II.

In terms of our dependent variable, class imbalance (Japkowicz and Stephen, 2002) is a concern, where most country-years do not include any interstate conflict onset. This is a common concern across many domains, where highly consequential events are relatively rare. In situations of class-imbalance a model with high accuracy is likely to incorrectly classify the minority class (here interstate conflict onset), which actually has more substantive importance. This data is no different, though the imbalance is not as intense as if this were a dyadic study of conflict. The imbalance is visualized in Figure 2, with conflict onset being unbalanced not just in the aggregate, but across all time periods. In the following section I discuss sampling

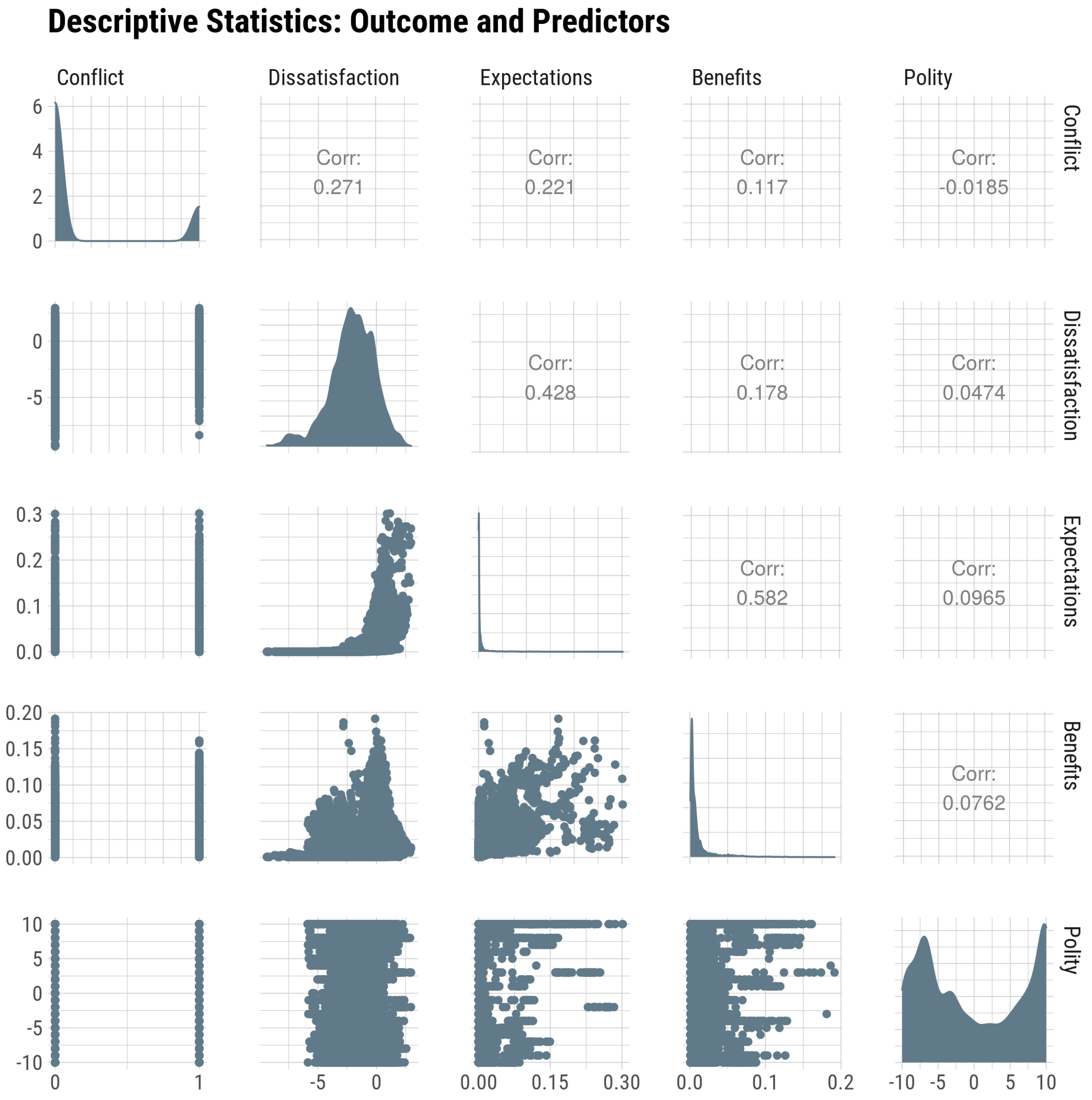

Figure 1: Distribution of and correlation between outcome and core features. Diagonal elements are the distribution for each variable. Elements to the left of the diagonal are scatterplots of each variable against another. Elements to the right of the diagonal list the correlation between each variable.

techniques used to create balance during model training.[18]

[18] Sampling methods – such as upsampling, downsampling, or SMOTE – are not the only way to deal with class imbalance, with zero-inflated models, and additional data gathering as alternative approaches. In our case, we have exhausted all available data and sampling methods proved more effective than zero-inflated regression models for this data.

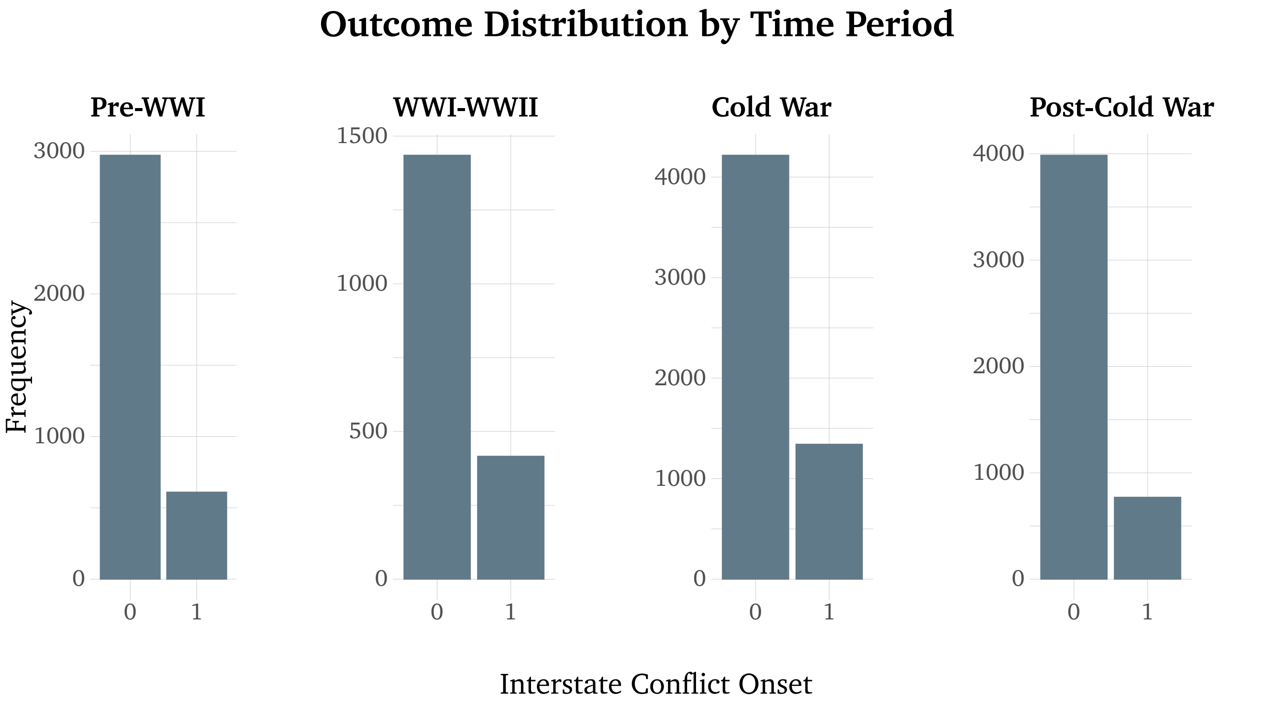

Figure 2: Outcome distribution across time periods. For each major era of international history interstate conflict onset is characterized by class imbalance, where most country-years are characterized by the absence of interstate conflict.

4 Methods

4.1 Comparing Predictive Accuracy

Rather than comparing regression coefficients or pursuing causal identification, I approach this data from a *predictive* perspective (Cranmer and Desmarais, 2017; Shmueli, 2010). In a traditional regression format, producing statistically significant coefficients for this data is a relatively straightforward task. However, testing for non-zero coefficients is not always equivalent to evaluating which variables have the strongest empirical link with conflict onset, especially when variables are of different scales and relationships are likely complex and non-linear. Instead, for each variable of interest I apply identical predictive models, feature engineering, cross-validation, and training and test sets. This leaves every model with identical modeling procedures but different predictors. Each variable is then evaluated through its predictive accuracy on identical test set data. While the purpose of this modeling approach is straightforward, in practice it presents an especially difficult standard. As one popular Bayesian statistics textbook puts it: “Fitting is easy; prediction is hard.”[19]

To set up a predictive architecture, I estimate a series of machine learning models, predicting whether or not a country starts a reciprocated militarized interstate dispute in a given year. These initial models serve as “base learners” whose predictions are then run through a stacked ensemble, or “metalearner”, in order to produce a final set of predictions. Each set of base learners is identical in format and data, varying only in the included predictors. The resulting product is three sets of predictions, each resulting from variables representative of a different theory. More formally, each theory is represented using identical feature engineering applied to a single starting variable – differential growth rates, yearly changes in a country’s regime type, or international dissatisfaction. For each of these predictors, let x_{it} represent a variable’s value for country i in year t, the general model is:

$$y_{it} = f\left(x_{it},\, x_{it-1},\, x_{it-2},\, (x_{it} - x_{it-1}),\, (x_{it} - x_{it-2}),\, x_{it}^2,\, x_{it}^3\right) + \varepsilon_{it}. \tag{6}$$

Each theory is therefore represented by a core variable, which is: lagged by 1 year, lagged by 2 years, first-differenced, second-differenced, squared, and cubed.[20]

As I discuss next, this predictive architecture is not set up to develop a new method or produce regression coefficients that can be tested in the null-hypothesis framework. Rather, the goal is to build a set of flexible machine learning models that draw on cutting-edge techniques from the statistical learning literature, allowing for accurate predictions of when interstate conflicts occur and investigating which variables contribute the most to accurate predictions.

[19] (McElreath, 2020, p12)

[20] The only exception is models of domestic changes from 1945 to the present, where binary variables are included for whether a civil war has begun or is ongoing. On this note, I fit multiple ensembles for each major period of international history, evaluating if the relative predictive accuracies are time-varying. This is driven by the fact that recent research suggests that the data-generating process for wars is time-varying (Anderson et al., 2016; Braumoeller, 2019; Jenke and Gelpi, 2017).

4.2 How This Approach Differs

In most Political Science papers, statistical results are communicated by fitting a generalized linear model to the entirety of one's data and presenting coefficient estimates in a table or dot-whisker plot. The approach outlined above differs from this practice in three important ways. First, while the traditional approach in Political Science evaluates variables by the statistical significance and direction of regression coefficients for the *entire dataset*, I instead measure a variable's statistical contributions by its predictive accuracy in a *test-set* – data which the model did not interact with when being fit. Model evaluation based on test-set predictive accuracy guards against inferring statistical quantities when one is actually overfitting and treating the case-specific nuances of various data points as if they were representative of the underlying characteristics of all data points. Moreover, if the model is able to accurately predict on data that did not contribute to training, then the model likely is accurately representing the true data-generating process.

Second, a particular benefit of the gradient boosting approach is that its training process increasingly weights observations that are difficult to classify. Each successive tree is fit to the last tree's residuals, meaning the goal is to predict observations which have previously been inaccurately modeled. This process is inherently well-suited for modeling rare events like interstate conflict because the first iteration of model-building will generally find that the highest-accuracy set of predictions is to just classify all cases as being peaceful. However, this leads to residuals being almost, if not entirely, cases of conflict onset, which the boosting machine will focus on more. With interstate conflict being notoriously difficult to predict and characterized by large class imbalance, this approach has a substantial advantage over straightforward generalized linear models, which treat all observations as equally important. Indeed, as Figure 10 demonstrates, the gradient boosting machine weighs a state's international dissatisfaction especially heavily in the training process. Considered in tandem with the gradient boosting machine's strengths for this data, this lends additional support for dissatisfaction, relative to the other variables of interest.

Third, both the random forest and gradient boosting machine are tree-based methods, which allows for non-linear relationships. Unlike a generalized linear model, where a specific linear form is assumed for a conditional expectation function and then parameters are estimated for that functional form, tree-based models are flexible and can accommodate many types of relationships.[21] This flexibility can admittedly lead to increasingly complex model formulations where increased fit comes with a tradeoff of decreased interpretability. However, as I demonstrate in the next section, techniques have been developed for gaining a general sense of which variables contribute the most to the final model and how predicted probabilities tend to vary across different values for one's predictor of interest.

4.3 Ensemble Component Models

In this section I present the methodology for a series of machine learning models (base learners) which are then run through a final stacked ensemble to classify interstate conflict initiation. In order to make my modeling choices as clear as possible, I go into a substantial level of detail for each technique. The three base learners are a regularized logistic regression, random forest, and gradient boosting machine.

First, the regularized regression is a LASSO, which estimates coefficients that minimize the negative log-likelihood for:[22]

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} \mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) + \lambda ||\beta||_1 \right\}. \tag{7}$$

Here, $\boldsymbol{y}$ is a N-length vector of outcomes, $\boldsymbol{X}$ is a $N \times p$ matrix of the predictors, and $\mathcal{L}$ is the log-likelihood function for a GLM (in our case a logistic regression). $\lambda||\beta||_1$ is referred to as the $\ell 1$ penalty, which is defined as $\lambda \sum_{j=1}^{p} |\beta_j|$, or the sum of the absolute value of all coefficients multiplied by a parameter, λ. Through a bias-variance tradeoff, increasing λ biases coefficients

[21] In terms of Breiman (2001b), one approach to statistics assumes a set data-generating process and estimates parameter values for that form, whereas the statistical learning approach treats the functional form for the data-generating process as unknown and to be estimated algorithmically from the available data.

[22] This notation is drawn from p.32 of Hastie et al. (2015). Hastie et al. (2009) and James et al. (2013) also discuss the LASSO in an acessable manner.

toward 0 but also decreases the model's variance, which increases accuracy when a model is otherwise prone to overfit. If $\lambda = 0$, then the model is equivalent to a traditional regression and if $\lambda = \infty$, then all coefficients will equal 0. While powerful and interpretable, the model does assume a linear relationship.

To relax this linearity assumption I next fit two tree-based classifiers, the first of which is a random forest. Rather than analytically solving for a set of coefficients, a random forest (Breiman, 2001a) builds a series of regression trees, where each tree provides a set of decision rules for predicting outcomes. The goal of a regression tree is to produce a set of non-overlapping regions, $R_1, R_2, ..., R_J$, with every observation that falls within a region, R_j, receiving the same prediction. These regions are decided upon through a partitioning process, where the model repeatedly evaluates which data point, s, within the available predictors, X_j, splits the data in a way that minimizes a within-group loss function. Random forests then build an ensemble of B regression trees – creating a forest – and average across each tree's predictions. The forest is built through a 'bagging' procedure (Breiman, 1996), where: bootstrapped samples are drawn (with replacement), a tree is built for each sample, and then the predictions for all trees are aggregated into a single final prediction. More formally:[23][24]

```
1. For $b = 1$ to $B$:
    (a) Sample some data with size $N$ from the training data, with replacement.
    (b) Fit a regression tree, $\hat{f}^b$, to the sampled data, where the subsequent steps are
        repeated until a pre-specified minimum node size, $n_{min}$, is reached.
          i. Randomly sample $m$ of the available $p$ features.
         ii. Find the split point that minimizes some loss function among the variables, $m$.
        iii. Split that node into two daughter nodes
2. Any tree $b$'s classification for data point $x$ is labelled $\hat{C}^b(x)$.
```

[23] See p588 and chapter 15 of Hastie et al. (2009) for this notation and a more detailed walkthrough.

[24] Generally for step ii, $m = \sqrt{p}$ or $m = \log_2 p$.

3. The random forest classifies data point x by evaluating:

$$\hat{C}^{B}_{rf}(x) = \textit{majority vote}\{\hat{C}^{b}(x)\}^{B}_{1} \tag{8}$$

where *majority vote* refers to the most frequent classification across all trees.

In a gradient boosting machine, rather than fitting a multitude of trees at once, trees are fit consecutively with the goal of predicting the error term from the past model. By repeatedly fitting a model to the previous residuals, misclassified observations receive additional focus. Indeed, the algorithm is designed so that the residuals for each model represent the difference between y_i and the sum of all previous models' predictions. Accordingly, once a tree correctly predicts the current residual, r_i, then the sum of the predictions across all models adds to y_i. Gradient boosting machines are powerful and generally more accurate than random forests. However, the process of iteratively converging on y_i also makes them prone to overfitting when y_i is widely dispersed around the target function $F(x)$. In order to reduce overfitting, two parameters are generally applied. First, λ, determines the learning rate, or how much each model's predictions are included. Second, like in a random forest, d determines how many splits are allowed for each tree – with fewer splits limiting the capacity of any one tree to overfit. Per James et al. (2013), the algorithm proceeds as follows:[25]

1. Set initial predictions as $f(x) = 0$ and initial residuals as $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, ..., B$, repeat:
   (a) Fit a regression tree, $\hat{f}^{b}$ with d splits ($d + 1$ terminal nodes) to the training data (X, r).

[25] p323

(b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{9}$$

where λ is a regularization parameter, minimizing the risk of overfitting.

(c) Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) \tag{10}$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{11}$$

4.4 Stacked Ensemble

Once each base learner is fit, I run them all through a stacked ensemble, or 'super learner'.[26] The goal of a stacked ensemble is to combine a collection of models and make predictions based on each model's strengths (Hastie et al., 2009).[27] This occurs by inserting the predictions from each base learner into a 'metalearner', which estimates how much each base learner is weighted in final predictions.[28] Rather than producing final predictions through an average of all base learner predictions, the stacked ensemble estimates a set of weights for the learner. Put in terms of our models, the stacked ensemble can be expressed as:[29]

$$\mathbb{P}(Y = 1|\hat{Y}_{LASSO}, \hat{Y}_{RF}, \hat{Y}_{GBM}) = \alpha_1 \hat{Y}_{LASSO} + \alpha_2 \hat{Y}_{RF} + \alpha_3 \hat{Y}_{GBM} \tag{12}$$

where $\alpha_1 \geq 0; \alpha_2 \geq 0; \alpha_3 \geq 0$ and $\sum_{k=1}^{3} \alpha_k = 1$.[30] The strength of the approach, as Van der

[26] Technically, the gradient boosting machine and random forest are also a type of ensemble method, because they aggregate predictions across multiple trees. But, the models are different from a stacked ensemble in that they weight all component learners equally.

[27] p605

[28] See Carroll and Kenkel (2016) for a recent application in Political Science.

[29] I use the H2O (The H2O.ai team, 2015) stacked ensemble functionality.

[30] Here I use the notation from Naimi and Balzer (2018), because it is easily interpreted. For a more thorough and formal presentation of the same process see Van der Laan et al. (2007).

Laan et al. (2007) and Polley and Van Der Laan (2010) demonstrate, is that the ensemble will perform *at least* as well as the "best" individual model within the ensemble. Across both stages of the stacking process – fitting component models and then estimating weights for those model predictions – all modelling concerns and recommendations about best-practices for cross-validation, loss functions, and parameter tuning apply. Indeed, a popular technique for estimating α_k is through a LASSO, so that regularization lowers the risk of overfitting.

4.5 Evaluating and Comparing the Stacked Ensembles

After fitting a stacked ensemble for each theory of revision, I compare the final models with precision recall (PR) and receiver operator characteristic (ROC) curves. The area under both curves is calculated for *test* sets, with an emphasis on the former because conflict onset is skewed toward zero (Cranmer and Desmarais, 2016; Davis and Goadrich, 2006). The set of variables with the highest test-set AUC score then represents the most empirically-supported theory. Example curves are visualized below in Figure 3. Per Davis and Goadrich (2006), the relevant quantities behind the curves are:

- True Positive Rate: $\frac{TP}{TP+FN}$
- False Positive Rate: $\frac{FP}{FP+TN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$

The PR-AUC and ROC-AUC curves then compare relative model scores for each statistic across different classification thresholds (where a case is classified as 1 at predicted probabilities ranging from 0.0 to 1.0.), which are denoted by the diagonal lines in Figure 3. Ultimately, the closer either AUC statistic is to 1, then the better model performance is, but for this data the PR-AUC is a stronger and more difficult benchmark.

Lastly, because of imbalance in the outcome variable – with most country-years lacking any conflict onset – I employ downsampling. Rather than selecting a completely random set of observations to train a model on, where that sample is also characterized by class imbalance, downsampling intentionally selects a lower proportion of the minority class. This creates a

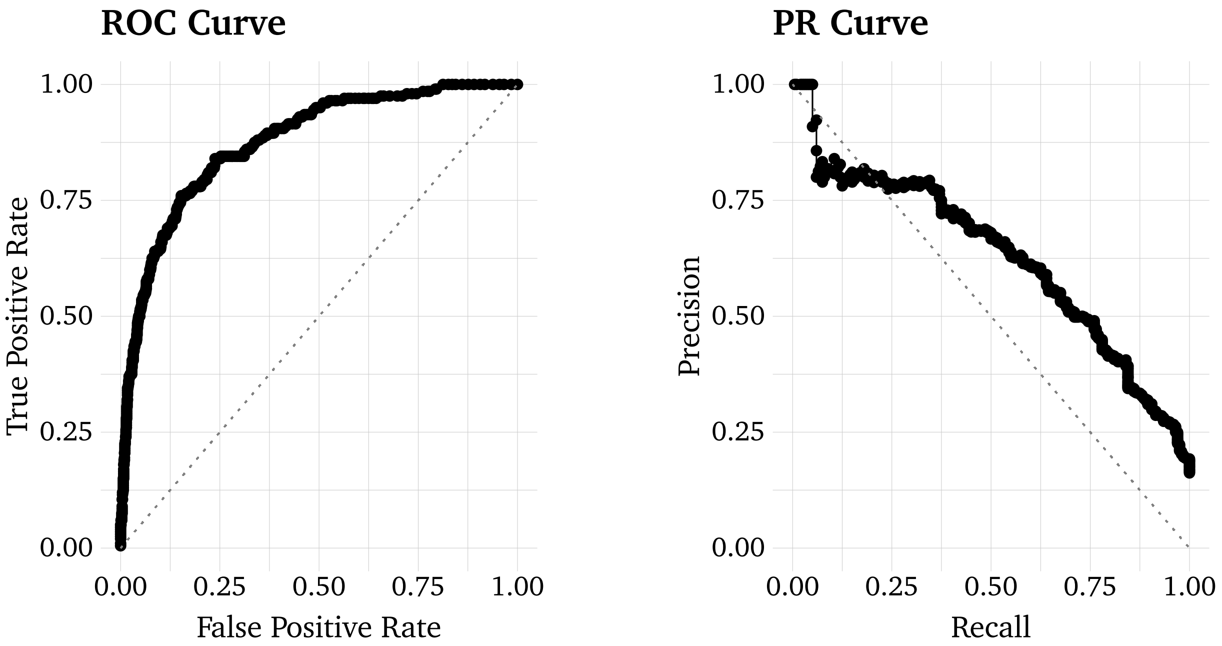

Figure 3: Example precision and recall curves for stacked ensemble trained on all possible features post-1945. Notably, while the ROC-AUC is 0.875, the PR-AUC is lower at 0.622. The drop in model accuracy is reflective of the class imbalance in conflict outcomes, where a high ROC-AUC is easy to achieve by predicting most cases as never having any conflict. Visually, a model with high predictive accuracy will curve into the top-left corner of the ROC-AUC plot and the top-right corner of the PR-AUC plot.

training set where the distribution of both classes is even. While this does induce some sampling bias – as the training sample now looks intentionally different from the full dataset – if one has a substantive reason for emphasizing the minority class, then that bias is often worth the increase in predictive accuracy toward the minority class.[31] In our case, a useful model needs to be able to predict cases of onset, because a naive simple model will tend to classify all years as peaceful and be correct most of the time. However, given the importance of preventing conflict, accurate predictions of when conflict does occur are of greater importance.

[31] In the case of these models, while downsampling lowers the test set ROC-AUC it increases the test set PR-AUC, because the latter emphasizes accuracy in the minority class.

5 Results

Before outlining the data and results, we should be careful to note that these ensembles are not meant to be *the one true model of international conflict*. Indeed, a careful read of the results will certainly raise questions of whether the models should have higher accuracy in test sets. But this paper does not intend to capture all of the variance and nuances behind interstate conflict. Rather, the goal is to compare how much revisionist behavior can be predicted, given certain information about a state – whether that be changes in their domestic political environment, differential growth rates, and/or dissatisfaction with the international system. On that point, as Figure 4 demonstrates, test set accuracy is the strongest for each time period when every possible variable is included in model training. Much like the R^2 in a linear regression always increasing with more variables, including additional variables adds predictive value. But the figure also makes it clear that not all variables contribute equally, which is our primary interest.

The test-set results for each are displayed below in Figure 4. Because the model trained on all possible features always returns the highest predictive accuracy in the test set, I display every other model's predictive *in comparison to the corresponding model trained with all features*. The left-hand plot compares each by their PR-AUC scores while the right-hand plot compares by ROC-AUC scores. For each plot the x-axis represents the time periods considered. The y-axis is then the difference in predictive accuracy for a model trained with variables corresponding to a single comparison, relative to a model trained on that time period's data with all possible features. The closer a model's value is to zero on the y-axis, then the better that model performs because that model is closer to the best-case model, given the available features. Models within each time period are differentiated by their color, with: orange for a model trained on features representing international dissatisfaction, blue for domestic variables, and grey for differential growth rates. Lastly, the first column for both plots includes model results when all years are considered.

The tables in Figure 4 show models trained on international dissatisfaction providing generally the highest test set predictive accuracy, with 8 out of the 10 AUC scores having the highest

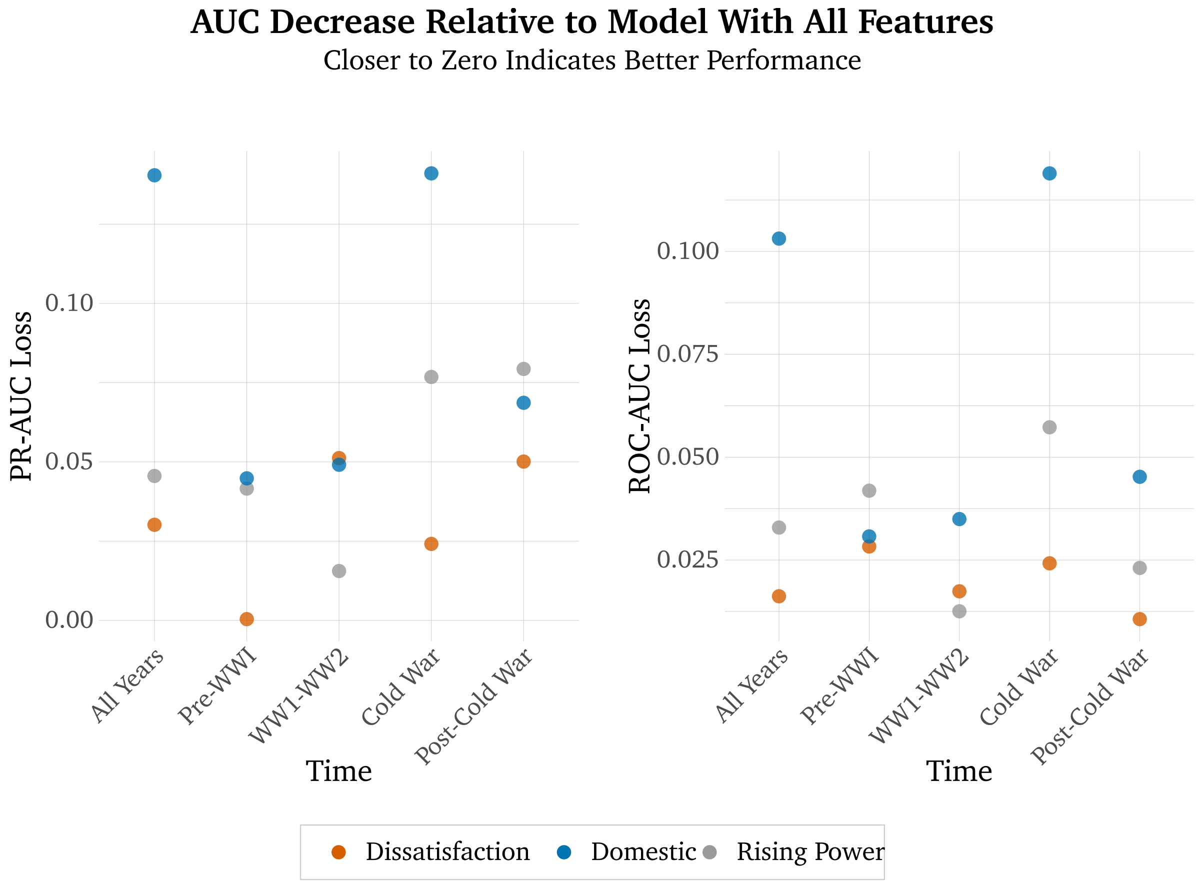

Figure 4: Across all time periods, the stacked ensemble trained with all possible features has the highest predictive accuracy in the test set. This plot compares models trained on a single set of features corresponding to each theory to the model trained with all possible features. The x-axis corresponds to the time period. The y-axis corresponds to how close the model of interest is to the model trained with all features. The left-side figure includes model PR-AUC scores and the right-side figure includes model ROC-AUC scores. Dissatisfaction returns the highest test-set accuracy expect for during the World Wars. This represents the fact that while also dissatisfied, Germany was a quintessential rising power before both conflicts and initiated the majority of the observed conflicts. However, the decrease in relative predictive accuracy for rising powers across all other time periods captures the danger of extrapolating Germany's behavior during the two wars and mapping it onto all other cases and time periods.

value for models trained on international dissatisfaction.[32] The one exception for both PR-AUC and ROC-AUC scores is between World Wars I and II, where models of rising powers return the highest score. In a sense, this is not surprising, because most interstate conflict during the time periods was initiated by Germany, which for both wars is often referred to as the quintessential

[32] This is not including the models trained on all variables, which unsurprising are the most accurate.

rising power. These lessons of both World Wars are understandably applied to many contexts across space and time, though this figure suggests that one should be hesitant to do so. Indeed, dissatisfaction returns a much higher predictive accuracy than rising powers in all other time periods.

One immediate concern when seeing these figures is that when considered separately, dissatisfaction may tend to provide greater predictive accuracy than differential growth rates and domestic political changes, but the most accurate model – including all variables – may rank the variables in an entirely different manner. In order to investigate whether or not models trained on all possible features weighs the features in the same manner as Figure 4 would suggest, I include variable importance plots. Fortunately, when we start to open the hood and investigate how the aggregated model processes the data, then we see that each of the machine learning models within the stacked ensemble gives dissatisfaction primary importance. Figure 10 plots out the top-10 variables in terms of importance for each model when trained on all possible years.[33] Variable importance plots capture a variable's prominence within a model and how much a variable's inclusion tends to reduce the model's mean squared error (or other loss function) on training data. The x-axis in Figure 10 is a variable's importance scaled between 0 and 1 and the variables are ranked along the y-axis in descending order from most important and down.

The variable importance plots all emphasize dissatisfaction's prominence in model-training. For the random forest, dissatisfaction and its transformations are the most important variables. For the gradient boosting machine, dissatisfaction lagged by two years has the most importance (by quite a bit) and for the LASSO dissatisfaction is similarly prominent across the top variables. Coupled with models trained on dissatisfaction returning the generally highest AUC scores in test sets, dissatisfaction's importance in models trained on all features gives it strong empirical support. However, neither AUC curves nor variable importance plots tell us anything about the estimated *directionality* of the relationship between dissatisfaction and conflict onset and

[33] The rankings are consistent when mdoels are trained on the different time periods.

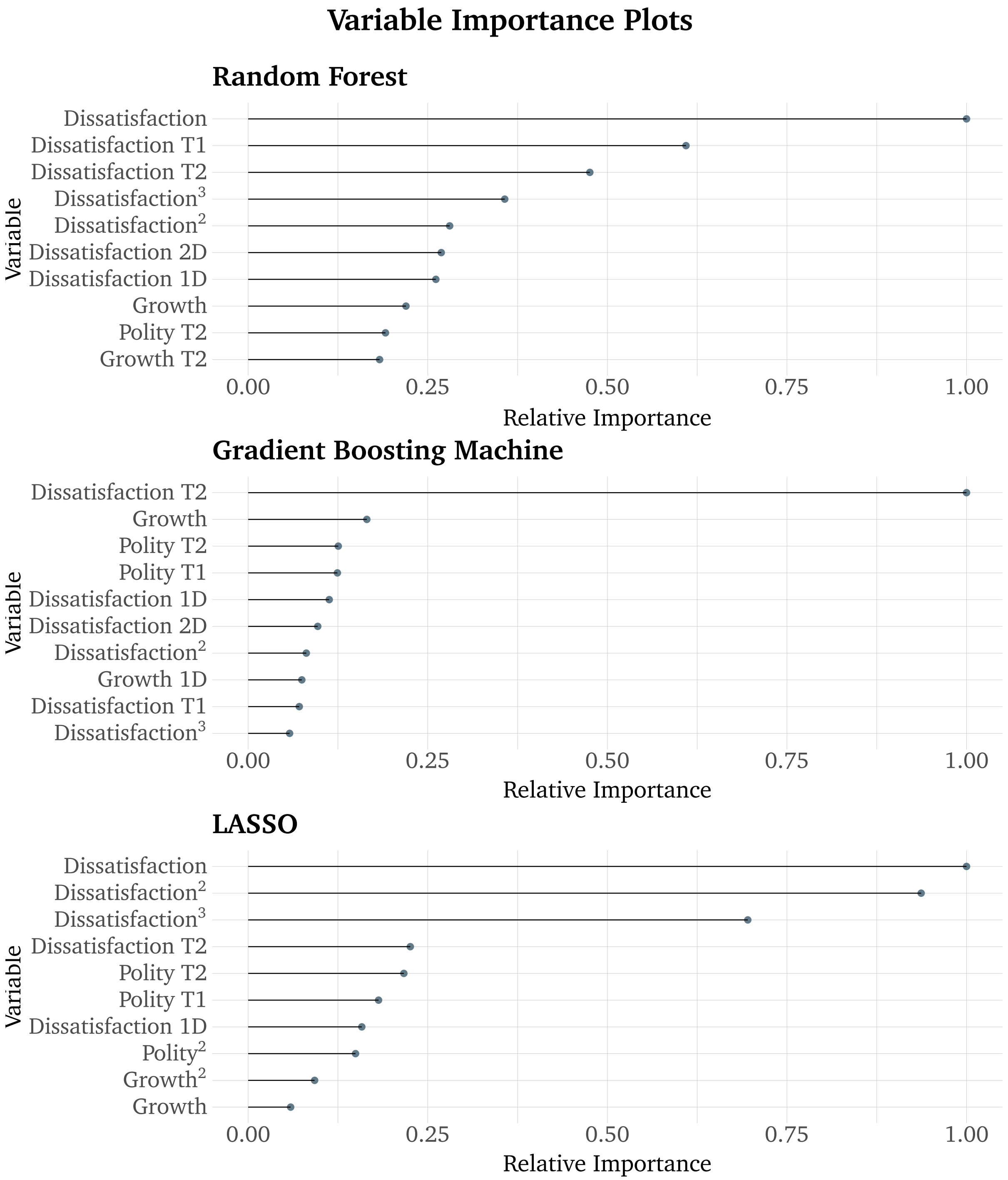

Figure 5: Top ten features per model, *in training* when each model is trained on all possible features for all possible years. Relative importance presents a standardized measure of how much a model's loss-function tends to decrease when a variable is included into a model.

that relationship's relative linearity. Fortunately, one straightforward way to capture this is to produce predicted probabilities across a reasonable range of values for all other features and to see how the predicted probability of conflict onset varies across possible values for dissatisfaction.

For complex non-linear machine learning models, partial dependence plots are a useful solution to challenges of interpretability (Greenwell, 2017). A partial dependence plot (PDP) holds all observations at their observed value and records the predicted value for each observation, varying one feature across a predefined range. Here, I take the model fit with all features in the Post-Cold War era, holding all features at their observed value except for dissatisfaction in the current year. Then for each observation the predicted probability of conflict is recorded at each value of dissatisfaction.[34] While the complexity of these machine learning models is a barrier to comprehensively mapping predicted probabilities across all possible situations, partial dependence plots do give us a fairly representative interpretation of how the model of interest is actually employing dissatisfaction in the actual data at hand, since predictions for each observation are included.[35]

Figure 6 provides two PDPs. The top figure is solely the average predicted probability of conflict onset for all observations, at each possible value of dissatisfaction. As we can see, while non-linear, the relationship is positive, confirming that the stacked ensemble does indeed tend to associate dissatisfaction positively with the probability of conflict onset. Next, visualizing how this average relationship is decided upon, the bottom plot provides a fairly comprehensive summary of the data-generating process for interstate conflict. The average association is positive, but we also see the model capturing a substantial amount of variation across observations, with a highly-variant intercept term and slopes. While most observations are deemed low-risk of conflict onset regardless of their dissatisfaction – capturing the class imbalance in conflict onset – we see the proportion of cases that are considered high-risk starting near a predicted probability of conflict at 0.5, even for the minimum possible value of dissatisfaction.

[34] I define the possible range of values for dissatisfaction as the range between the minimum and maximum observed measure of dissatisfaction for any country in the years under consideration.

[35] The danger of these plots is that they are easily discussed in a causal manner, but for the models at hand we are maximizing predictive accuracy, which is often not the same as a research design for reaching causal identification. Nonetheless, we do receive a measure of model-based association for any variable of interest.

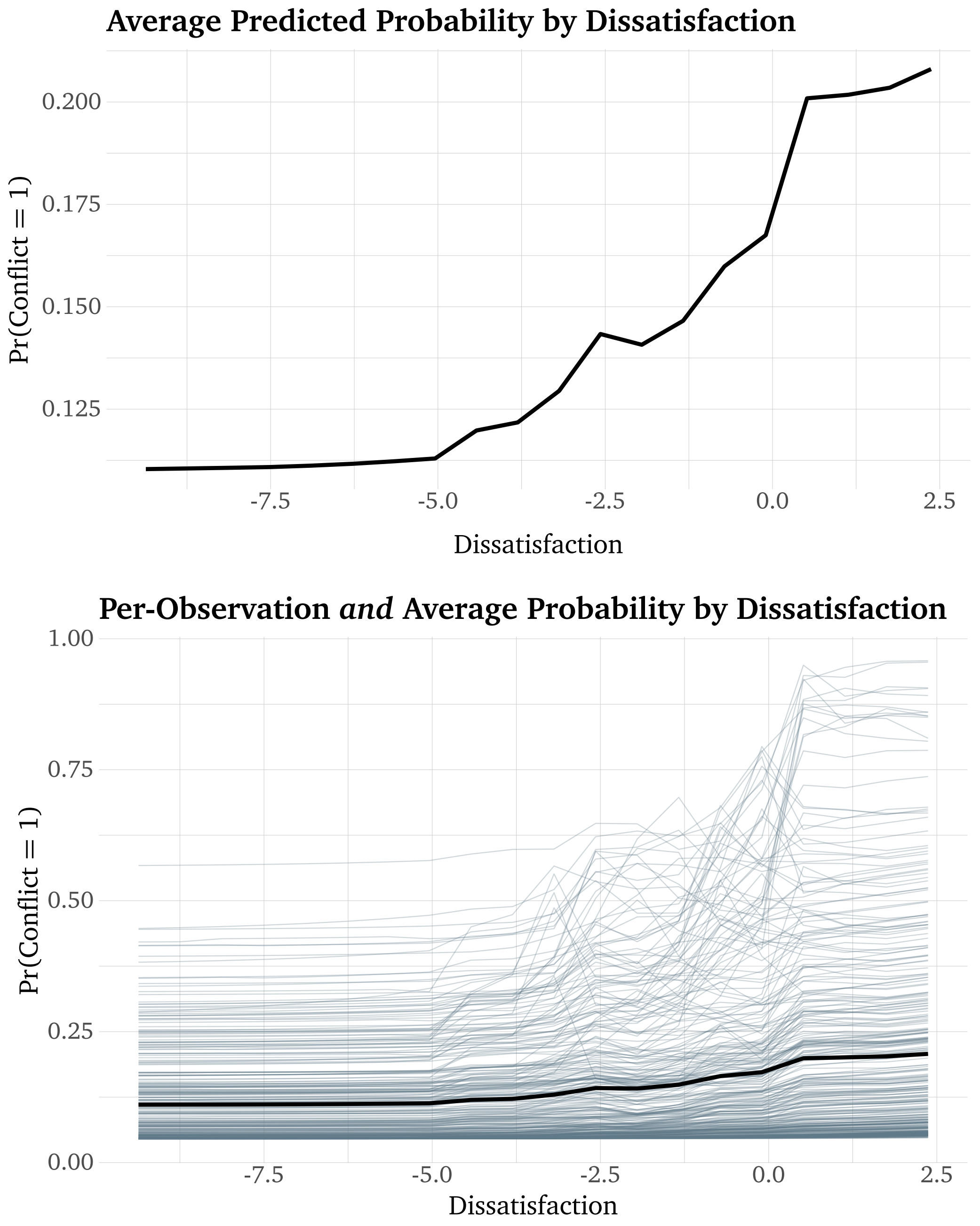

Figure 6: Predicted probability of conflict onset by dissatisfaction. The top plot includes the average predicted probability at each value of dissatisfaction. The bottom plot includes the predicted probabilities of conflict onset *for each observation* at each level of dissatisfaction.

These high risk cases also have a more severe slope, quickly ramping up to a probability of conflict onset near 1 as dissatisfaction increases – which we can see in the plot's top-right corner.

Understood in terms of interstate conflict, most states in most years do not consider starting a military conflict as a possibility, but for a small handful of states the prospects are very real and quickly can escalate.

Lastly, unpacking the machine learning model's complexity, I also consider how the model treats observations which are both highly dissatisfied and a rising power. In Figure 7, I calculate the average predicted probability of conflict onset for all observations, varying both dissatisfaction and differential growth rates. While the probability of conflict onset clearly varies the most around dissatisfaction, with the average predicted probability not varying by differential growth rates unless dissatisfaction is sufficiently high, we do see that the highest-risk cases on average (those in the top-right corner) are states which are both intensely dissatisfied and rising quickly. However, even if a state is not growing or shrinking, the average predicted probability of conflict is relatively high when a state is sufficiently dissatisfied. This figure captures the important point that dissatisfaction is not the *only* predictor of conflict, it just tends to provide greater predictive capacity than the other variables under consideration.

6 Implications

What are the policy implications of these findings? The predictive nature of these results provides particularly useful insights into current and future trends for great power competition, revision, and conflict. The dissatisfaction measure reveals European powers that are unsurprisingly satisfied with a system that has been immensely beneficial to them since the end of the Cold War. These states are therefore unlikely to become conflictual revisionists. Yet, the remaining non-European global great powers have consistently been dissatisfied with their benefits from the international system, given their power-based expectations. Indeed, even the United States is estimated to have reason to think that the international system could be better designed to reflect its preferences and therefore is likely to continue to engage in some conflictual revisionist behavior, despite its hegemony. Moreover, for United States foreign pol-

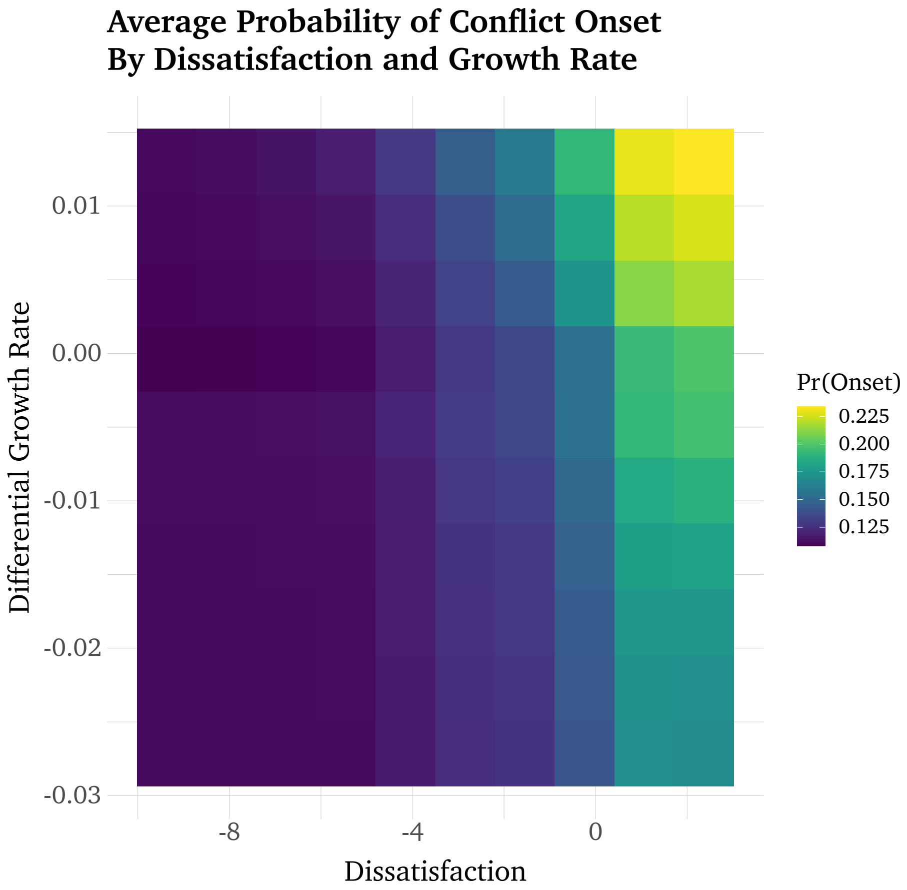

Figure 7: Average predicted probability of conflict onset for all states if each observation is held at its actual values except for the denoted dissatisfaction and growth rate. International dissatisfaction is included on the x-axis and differential growth rates on the y-axis. The lighter colored a grid square is, then the *greater* the predicted probability of conflict onset is. We see the top-right corner with highly dissatisfied and quickly growing states being the highest-risk scenario.

icy, as debates move from counter-terrorism to great power competition, then we can see that China and Iran have been especially dissatisfied, relative to their peers, informing predictions that they are both similarly likely to pursue conflictual revision and be continual hotspots over their peers.

Given that these estimates stop in 2012, we now know that both China and Iran have been frequently active in military affairs throughout their regions and their revisionist tendencies are now at the forefront of policy discussions and debates. Yet, while dissatisfaction estimates cannot be produced for more recent years – due to the component datasets not being available past 2012 – these trends are likely to be more troubling in 2020 for both countries. Iran's nuclear program has advanced – further alienating the country and producing increasingly stifling sanctions – and China is increasingly seen as an adversary, not a partner. The estimated measure and predictive exercise therefore highlight China and Iran as being the highest-risk of revisionist behavior. On the other hand, while Russia is also more dissatisfied than other large states and therefore likely to subsequently initiate conflict – which we retrospectively saw in Ukraine – the measure suggests that China and Iran merit greater focus and expectations for subsequent revisionist behavior than Russia.

Summarizing and discussing the dissatisfaction estimates in more detail, figure 8 includes the dissatisfaction estimates for a handful of great powers from 1991-2012. With the European powers below zero, they are estimated to be satisfied with the current state of affairs for each year. However, when it comes to the last year of data – 2012 – we see dissatisfaction from China, Russia, and the United States. Even more troubling – and likely a foreshadowing of the next 7 years – Iran is the most dissatisfied state by 2012. Indeed, in 2011 the United States congress passed legislation to begin sanctioning foreign banks processing transactions with Iran's Central Bank.[36] Put in this context, as Iran's nuclear program has developed, so too have its relations with many countries frayed. While Iran's estimated dissatisfaction provides an assurance of the measure's validity, it also emphasizes the security dilemma associated with developing a nuclear program. As Iran's program has advanced, so too have other countries grown more wary of Iran's position in the world, creating a greater sense of threat to Iran and perceived need for the nuclear program. Given the state of relations with Iran and intense sanctioning under the current U.S. administration, we should expect to see continued military conflict in the Persian Gulf, not just as a part of the ongoing catastrophic proxy war between Iran and Saudi Arabia in Yemen.

Turning to China, if its dissatisfaction is considered alongside this paper's results, then

[36] https://www.armscontrol.org/factsheets/Timeline-of-Nuclear-Diplomacy-With-Iran

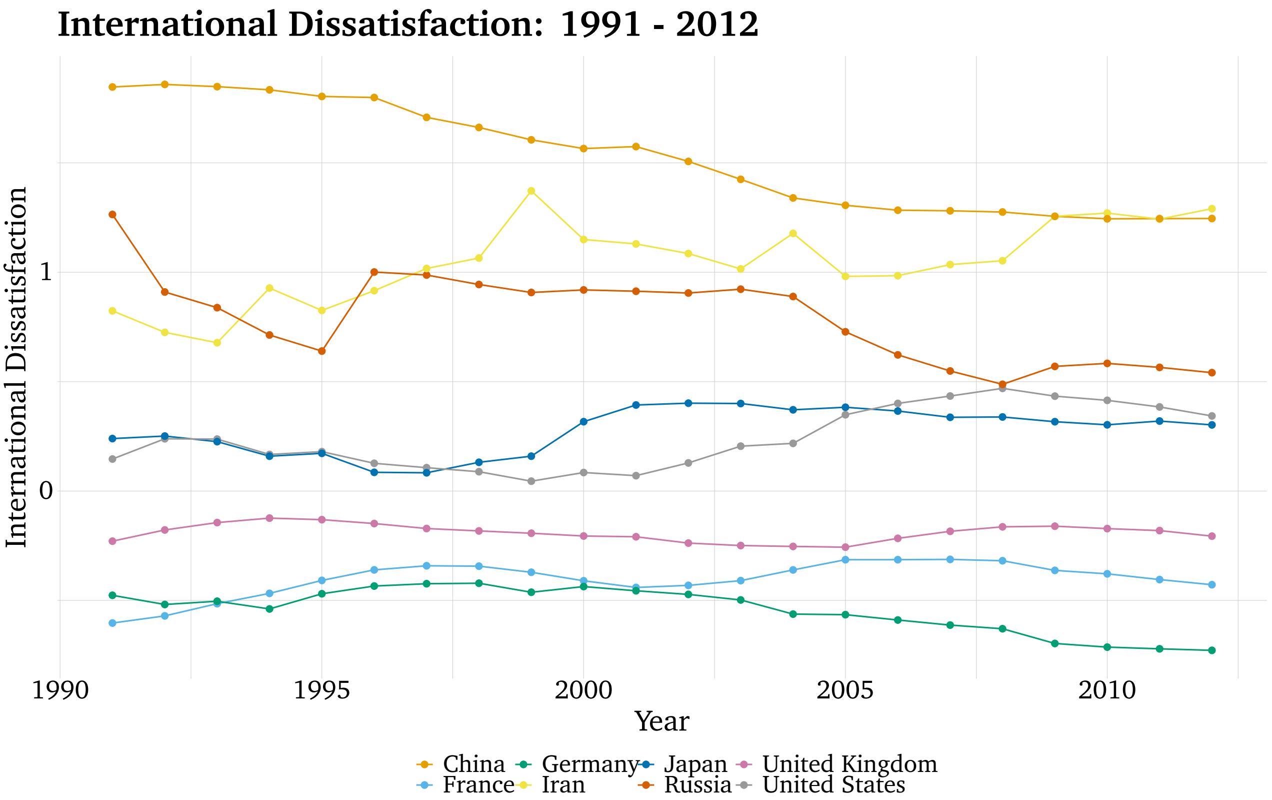

Figure 8: International dissatisfaction estimates for select great powers from 1991 to 2012. Values greater than zero represent increasing dissatisfaction and values below zero represent increasing satisfaction with the international status quo. While the European powers are consistently satisfied with the status quo, non-European states display a consistent desire of and expectation for greater benefits. Iran's international dissatisfaction is particularly striking, likely reflecting the growth of its nuclear program and the intense international sanctioning that has followed it.

greater focus should be directed to its relations with other states – or the lack thereof. While the dominant Chinese narrative is likely around rising powers and the 'Thucydides Trap' (Allison, 2017), whether or not China pursues a long-term strategy of revision (Johnston, 2003) may very well be decided more by whether or not other states are comfortable with deepening relations with China as it grows, not whether or not a power transition will occur or if Xi Jinping's leadership will remain uncontested. This is not to say that the latter two questions are unimportant, but this paper's results call for a focus on the rest of the international system and whether that system is likely to behave in a way that China would want to revise. In response to these concerns one may ask why China would seek to revise an international system which enabled its rise. However, a straightforward reply is that China has good reason to think it can and should be getting more from its international environment, based on its power position alone.

For example, as the Belt and Road Initiative (OBOR) expands and East Asian institutions like the Asian Infrastructre Investment Bank (AIIB) are spearheaded by China, should potential partners demonstrate a willingness to buy into those institutions, then large revisionist conflict will likely decrease in probability. On the other hand, if potential partners view these institutions as illegitimate or too high-risk for the expected payoffs, then China's expected international benefits will continue to outstrip its actual international benefits and radical change to the international system will gain appeal. The appeal of these institutions to potential members is an open, but potentially the most important, question when it comes to China's future foreign policy.

7 Conclusion

When do revisionist states become revisionists? This paper approaches the question of revision's origins through a predictive approach, comparing the predictive accuracy of machine learning models trained on different theoretically-informed variables when forecasting interstate conflict initiation. While models trained on differential growth rates and domestic political changes demonstrate meaningful predictive accuracy – with the former being the most accurate model during the World Wars, a state's international dissatisfaction presents the most predictive accuracy, in most years. These results do not invalidate theories of domestic politics and rising powers when considering international revision, but they do support concentrating one's focus on a state's relative standing and relations with other states when asking when international revision occurs. In other words, revision does not just follow from changes within the state, but in response to whether or not a state has reason to believe that the system can plausibly be altered to better suit that state's desires.

I also investigate the implications for current great powers relations. Both Iran and China are estimated to be the most dissatisfied major states during the last year of available data. While the two states' dissatisfaction, (alongside its predictive capacity) lend credence for the measure itself, these two states' strong dissatisfaction with the international system raises questions of foreign policy design and potential future outcomes. For both states, dissatisfaction stems from a lack of international standing and strong relations commensurate to their capabilities. This dissatisfaction over their international *relations* with other states means whether or not the two turn toward increased or decreased international conflict in the future will likely be very predictable by the decisions of other states and whether potential partners are willing to deepen relations with both states. In this context, questions of international revision and conflictual actors should be just as, if not more, focused on the makeup of the relations that constitute the international system as the potential belligerent actors themselves.

Lastly, this project has been relatively agnostic on the sources of international dissatisfaction. A logical next step for future research is to examine why some states become highly dissatisfied and socially excluded whereas others find themselves in a situation of general satisfaction and benefit with the status quo. While domestic politics and growth rates are certainly part of the picture – raising questions of a potentially spurious relationship – as we saw earlier in Figure 1, the correlation between dissatisfaction and growth rates is below 0.5 and the correlation with polity scores is almost zero. Put differently, the question is more than one of academic interest and nuance. While this study identifies dissatisfaction as an especially meaningful predictor of interstate conflict and revisionist states, it provides little insight on the best policy levers that can be pulled in order to peacefully bring states into the international fold – remedying intense dissatisfaction among great powers. Understanding why some international environments and states may be more prone than others to social exclusion may then provide a better understanding of which levers are best-suited for ameliorating these complicated but consequential dynamics.

References

G. Allison. *Destined for War: Can America and China Escape Thucydides's Trap?* Houghton Mifflin Harcourt, 2017.

C. C. Anderson, S. M. Mitchell, and E. U. Schilling. Kantian dynamics revisited: Time-varying analyses of dyadic igo-conflict relationships. *International Interactions*, 42(4):644–676, 2016.

M. Beckley. China's century? why america's edge will endure. *International Security*, 36(3): 41–78, 2012.

M. Beckley. *Unrivaled: Why America will remain the world's sole superpower*. Cornell University Press, 2018.

B. F. Braumoeller. *The great powers and the international system: systemic theory in empirical perspective*. Cambridge University Press, 2013.

B. F. Braumoeller. *Only the dead: the persistence of war in the modern age*. Oxford University Press, 2019.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

L. Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001b.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

E. H. Carr. The twenty years' crisis, 1919-1939: an introduction to the study of international relations. 1946.

R. J. Carroll and B. Kenkel. Prediction, proxies, and power. *American Journal of Political Science*, 2016.

S. Chan. Can't get no satisfaction? the recognition of revisionist states. *International relations of the Asia-pacific*, 4(2):207–238, 2004.

T. J. Christensen. Fostering stability or creating a monster? the rise of china and us policy toward east asia. *International security*, 31(1):81–126, 2006.

S. J. Cranmer and B. A. Desmarais. A critique of dyadic design. *International Studies Quarterly*, 60(2):355–362, 2016.

S. J. Cranmer and B. A. Desmarais. What can we learn from predictive modeling? *Political Analysis*, 25(2):145–166, 2017.

A. Dafoe, J. Renshon, and P. Huth. Reputation and status as motives for war. *Annual Review of Political Science*, 17:371–393, 2014.

J. Davidson. *The origins of revisionist and status-quo states*. Springer, 2006.

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Y. Deng. *China's struggle for status: the realignment of international relations*. Cambridge University Press, 2008.

M. Duffy Toft. Population shifts and civil war: A test of power transition theory. *International Interactions*, 33(3):243–269, 2007.

M. G. Duque. Recognizing international status: A relational approach. *International Studies Quarterly*, 2018.

D. M. Edelstein. *Over the Horizon: Time, Uncertainty, and the Rise of Great Powers*. Cornell University Press, 2017.

J. D. Fearon. Rationalist explanations for war. *International organization*, 49(3):379–414, 1995.

M. Finnemore and K. Sikkink. International norm dynamics and political change. *International organization*, 52(4):887–917, 1998.

R. Gilpin. *War and change in world politics*. Cambridge University Press, 1983.

S. E. Goddard. *When Right Makes Might: Rising Powers and World Order*. Cornell University Press, 2018a.

S. E. Goddard. Embedded revisionism: Networks, institutions, and challenges to world order. *International Organization*, pages 1–35, 2018b.

E. Goh. *The struggle for order: Hegemony, hierarchy, and transition in post-Cold War East Asia*. Oxford University Press, 2013.

B. M. Greenwell. pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1):421–436, 2017. doi: 10.32614/RJ-2017-016. URL https://doi.org/10.32614/RJ-2017-016.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

O. R. Holsti, R. M. Siverson, and A. L. George. *Change in the international system*. Routledge, 2019.

M. C. Horowitz, A. C. Stam, and C. M. Ellis. *Why leaders fight*. Cambridge University Press, 2015.

P. T. Jackson and D. H. Nexon. Relations before states: Substance, process and the study of world politics. *European journal of international relations*, 5(3):291–332, 1999.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

L. Jenke and C. Gelpi. Theme and variations: Historical contingencies in the causal model of interstate conflict. *Journal of Conflict Resolution*, 61(10):2262–2284, 2017.

A. I. Johnston. Is china a status quo power? *International security*, 27(4):5–56, 2003.

A. I. Johnston and R. S. Ross. *Engaging China: The management of an emerging power*, volume 24. Psychology Press, 1999.

K. Kadera. *The power-conflict story: A dynamic model of interstate rivalry*. University of Michigan Press, 2001.

A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

P. Kennedy. *The rise and fall of the great powers: economic change and military conflict from 1500 to 2000*. Vintage, 2010.

D. W. Larson and A. Shevchenko. Status seekers: Chinese and russian responses to us primacy. *International Security*, 34(4):63–95, 2010.

G. Lawson. Revolutions and the international. *Theory and Society*, 44(4):299–319, 2015.

R. N. Lebow and B. Valentino. Lost in transition: A critical analysis of power transition theory. *International Relations*, 23(3):389–410, 2009.

D. Lemke and W. Reed. Regime types and status quo evaluations: Power transition theory and the democratic peace. *International Interactions*, 22(2):143–164, 1996.

P. Y. Lipscy. *Renegotiating the World Order: Institutional Change in International Relations*. Cambridge University Press, 2017.

J. M. Lyall. *Paths of Ruin: Why Revisionist States Arise and Die in World Politics*. Cornell University, 2005.

P. K. MacDonald. *Networks of Domination: The Social Foundations of Peripheral Conquest in International Politics*. OUP Us, 2014.

P. K. MacDonald and J. M. Parent. *Twilight of the Titans: Great Power Decline and Retrenchment*. Cornell University Press, 2018.

M. G. Marshall, T. R. Gurr, C. Davenport, and K. Jaggers. Polity iv, 1800-1999: Comments on munck and verkuilen. *Comparative Political Studies*, 35(1):40–45, 2002.

R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

J. J. Mearsheimer. Can china rise peacefully? *The National Interest*, 25:23–37, 2014.

E. B. Montgomery. *In the Hegemon's Shadow: Leading States and the Rise of Regional Powers*. Cornell University Press, 2016.

H. Morgenthau. Politics among nations: The struggle for power and peace. *Nova York, Alfred Kopf*, 1948.

A. I. Naimi and L. B. Balzer. Stacked generalization: an introduction to super learning. *European journal of epidemiology*, 33(5):459–464, 2018.

A. F. Organski and J. Kugler. *The war ledger*. University of Chicago Press, 1981.

G. Palmer, V. D'Orazio, M. R. Kenwick, and R. W. McManus. Updating the militarized interstate dispute data: A response to gibler, miller, and little. *International Studies Quarterly*, 2019.

T. Paul. *Accommodating rising powers: past, present, and future*. Cambridge University Press, 2016.

T. V. Paul, D. W. Larson, and W. C. Wohlforth. *Status in world politics*. Cambridge University Press, 2014.

E. C. Polley and M. J. Van Der Laan. Super learner in prediction. 2010.

Y. Qin. A relational theory of world politics. *International Studies Review*, 18(1):33–47, 2016.

J. Radford and K. Joseph. Theory in, theory out: The uses of social theory in machine learning for social science. *arXiv*, pages arXiv–2001, 2020.

J. Renshon. Status deficits and war. *International Organization*, 70(3):513–550, 2016.

J. Renshon. *Fighting for status: hierarchy and conflict in world politics*. Princeton University Press, 2017.

K. Schake. *Safe Passage: The Transition from British to American Hegemony*. Harvard University Press, 2017.

P. W. Schroeder. *The transformation of European politics, 1763-1848*. Oxford University Press, 1994.

R. L. Schweller. Bandwagoning for profit: Bringing the revisionist state back in. *International Security*, 19(1):72–107, 1994.

R. L. Schweller. Managing the rise of great powers: history and theory. *Engaging China: The management of an emerging power*, pages 1–31, 1999.

R. L. Schweller. Rising powers and revisionism in emerging international orders. *Russia in Global Affairs*, 7, 2015.

J. R. I. Shifrinson. *Rising Titans, Falling Giants: How Great Powers Exploit Power Shifts*. Cornell University Press, 2018.

G. Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

D. Snidal. Relative gains and the pattern of international cooperation. *American Political Science Review*, 85(3):701–726, 1991.

J. Snyder. Civil-military relations and the cult of the offensive, 1914 and 1984. *International Security*, 9(1):108–146, 1984.

J. Snyder. *Myths of empire: Domestic politics and international ambition*. Cornell University Press, 1991.

The H2O.ai team. *h2o: R Interface for H2O*, 2015. URL http://www.h2o.ai. R package version 3.1.0.99999.

C. G. Thies and M. D. Nieman. *Rising Powers and Foreign Policy Revisionism: Understanding BRICS Identity and Behavior Through Time*. University of Michigan Press, 2017.

D. H. Tingley. The dark side of the future: An experimental test of commitment problems in bargaining. *International Studies Quarterly*, 55(2):521–544, 2011.

M. Trachtenberg. Audience costs: An historical analysis. *Security Studies*, 21(1):3–42, 2012.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

S. Van Evera. The cult of the offensive and the origins of the first world war. *International Security*, 9(1):58–107, 1984.

S. M. Walt. *Revolution and war*. Cornell University Press, 1996.

S. Ward. *Status and the Challenge of Rising Powers*. Cambridge University Press, 2017.

J. L. Weeks. Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1):35–64, 2008.

R. Wolf. Respect and disrespect in international politics: the significance of status recognition. *International Theory*, 3(1):105–142, 2011.

A. Wolfers. Discord and collaboration: Essays on international politics, 1962.

Appendix

This figure includes variable importance plots for models predicting the onset of fatal MIDs.

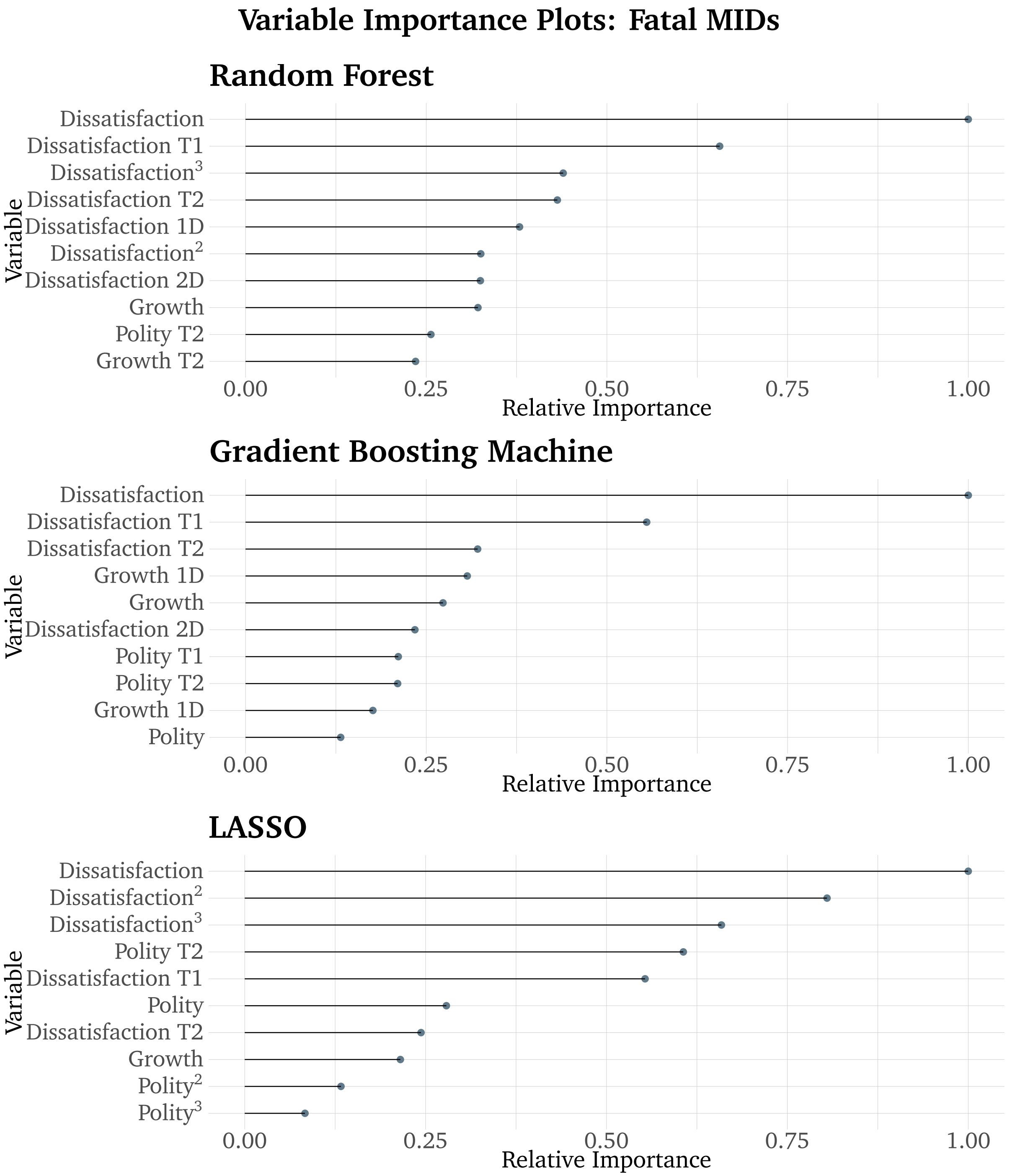

Figure 9: Top ten features per model, *in training* when each model is trained on all possible features for all possible years. Relative importance presents a standardized measure of how much a model's loss-function tends to decrease when a variable is included into a model.

When wars are the DV, then polity scores take primary importance for the LASSO, which is change from models with reciprocated or fatal MIDs as the outcome of interest. However, the LASSO also contributes the least to a stacked ensemble and is the only case where polity scores have such a high predictive importance.

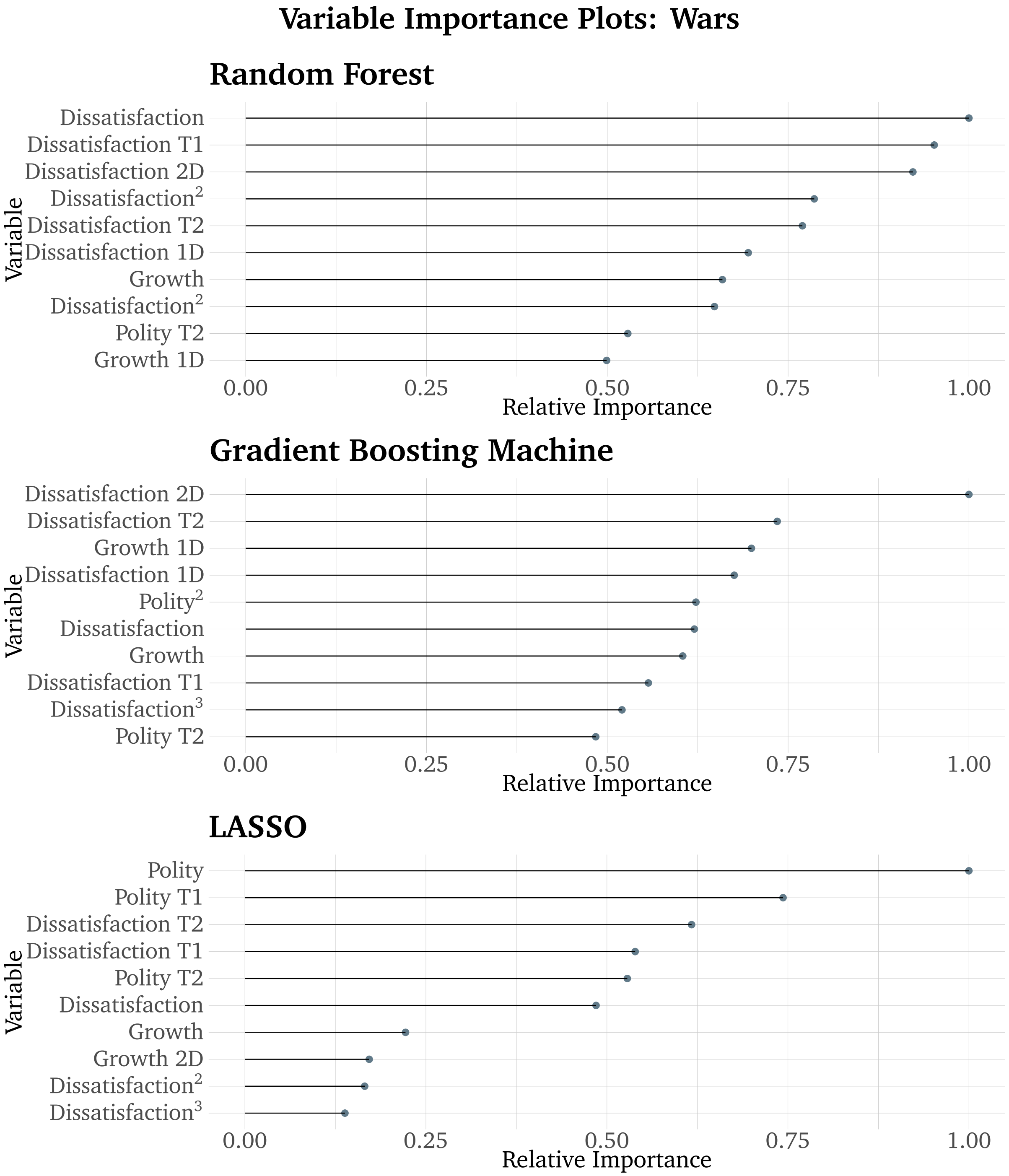

Figure 10: Top ten features per model, *in training* when each model is trained on all possible features for all possible years.

This scatterplot shows that the states that tend to grow the fastest tend to be the large states, not the small states.

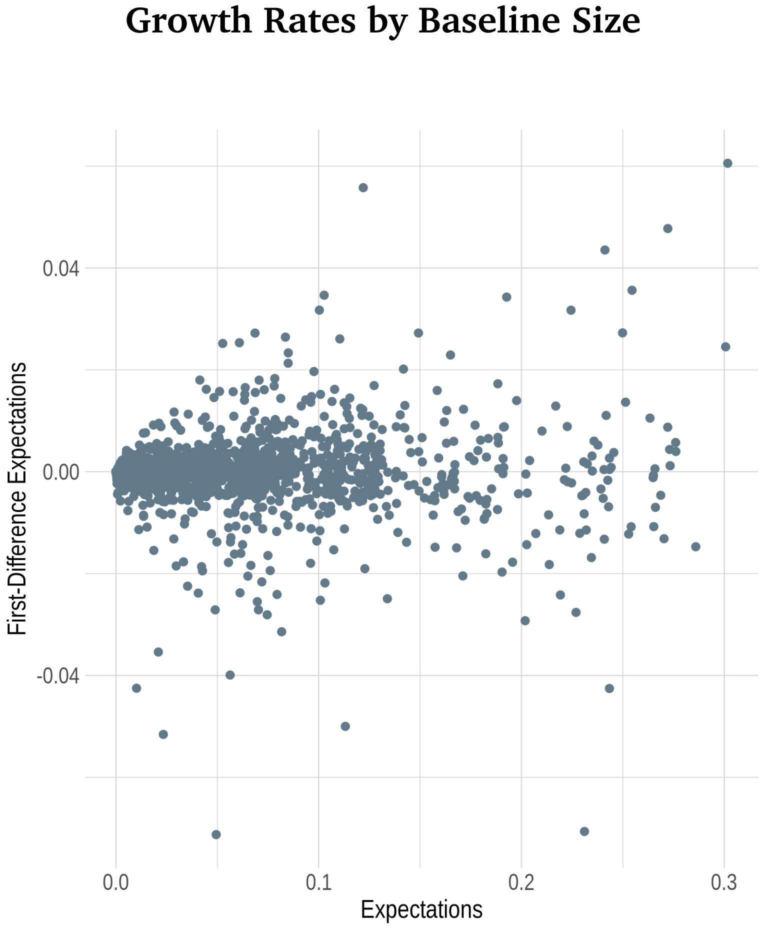

Figure 11: Growth rates compared to state size

Reformatting of model comparison by PR-AUC and ROC-AUC.

The following figures include estimates for a state's international dissatisfaction, benefits, and expectations for select great powers.

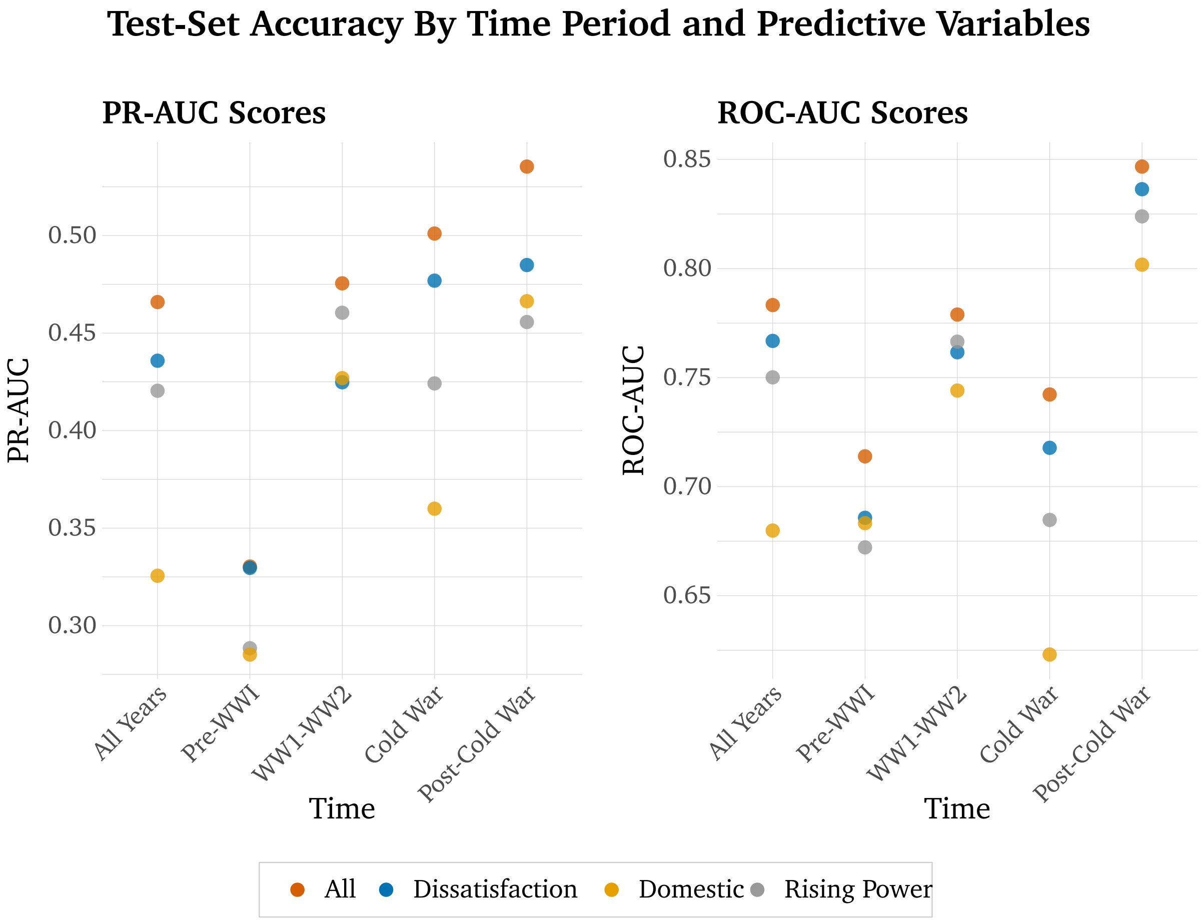

Figure 12: Test set PR-AUC and ROC-AUC for the stacked ensemble across time periods and features included. Higher values on the y-axis represent higher accuracy in test set predictions. Across all time periods for both AUC scores a model trained on all possible features unsurprisingly returns the greatest accuracy. However, among the models trained on a single variable with identical transformations, dissatisfaction returns the highest accuracy expect for during the World Wars. This represents the fact that while also dissatisfied, Germany was a quintessential rising power before both conflicts and initiated the majority of the observed conflicts. However, the decrease in relative predictive accuracy for rising powers across all other time periods captures the danger of extrapolating Germany's behavior during the two wars and mapping it onto all other cases and time periods.

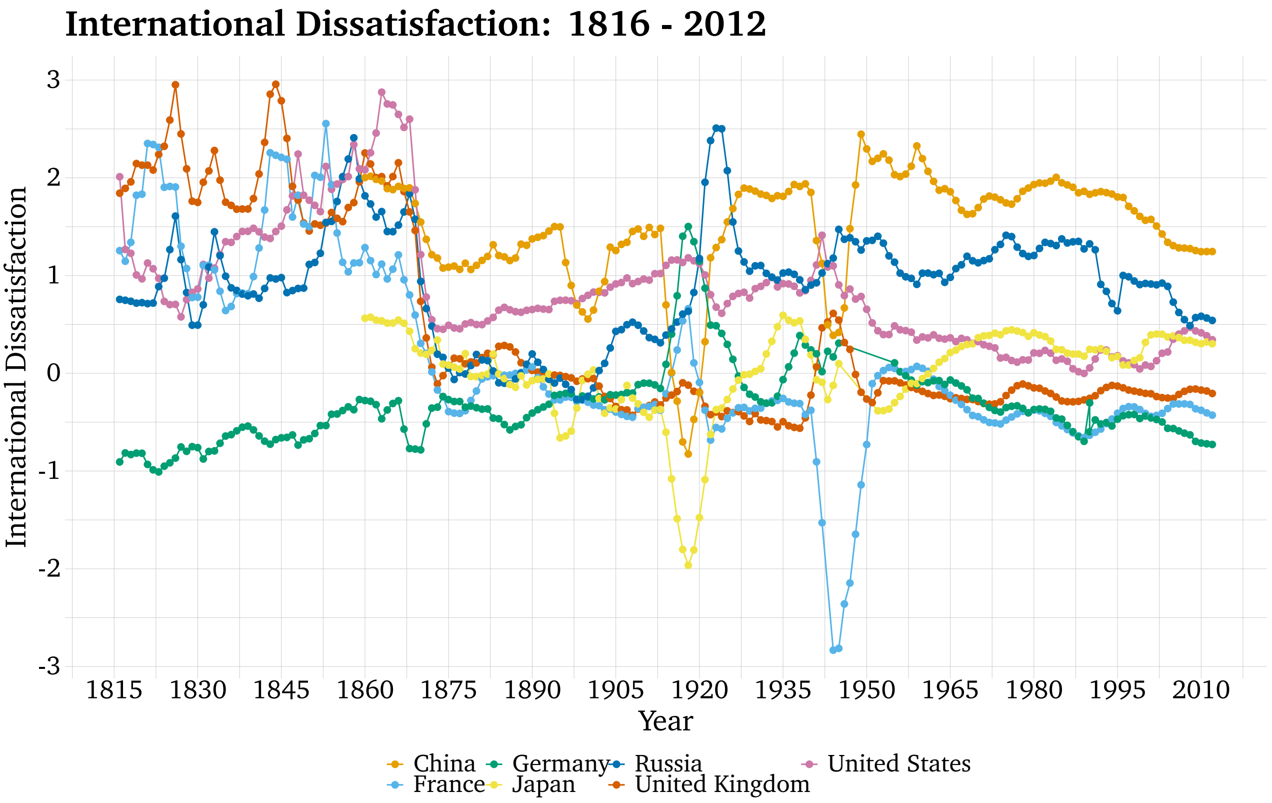

Figure 13: Estimates for a state's international dissatisfaction from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable. Colors correspond to the country. Positive values on the y-axis correspond to increasingly dissatisfied states. Negative values on the y-axis correspond to increasingly *satisfied* states. Note that before each World War Germany spikes in dissatisfaction.

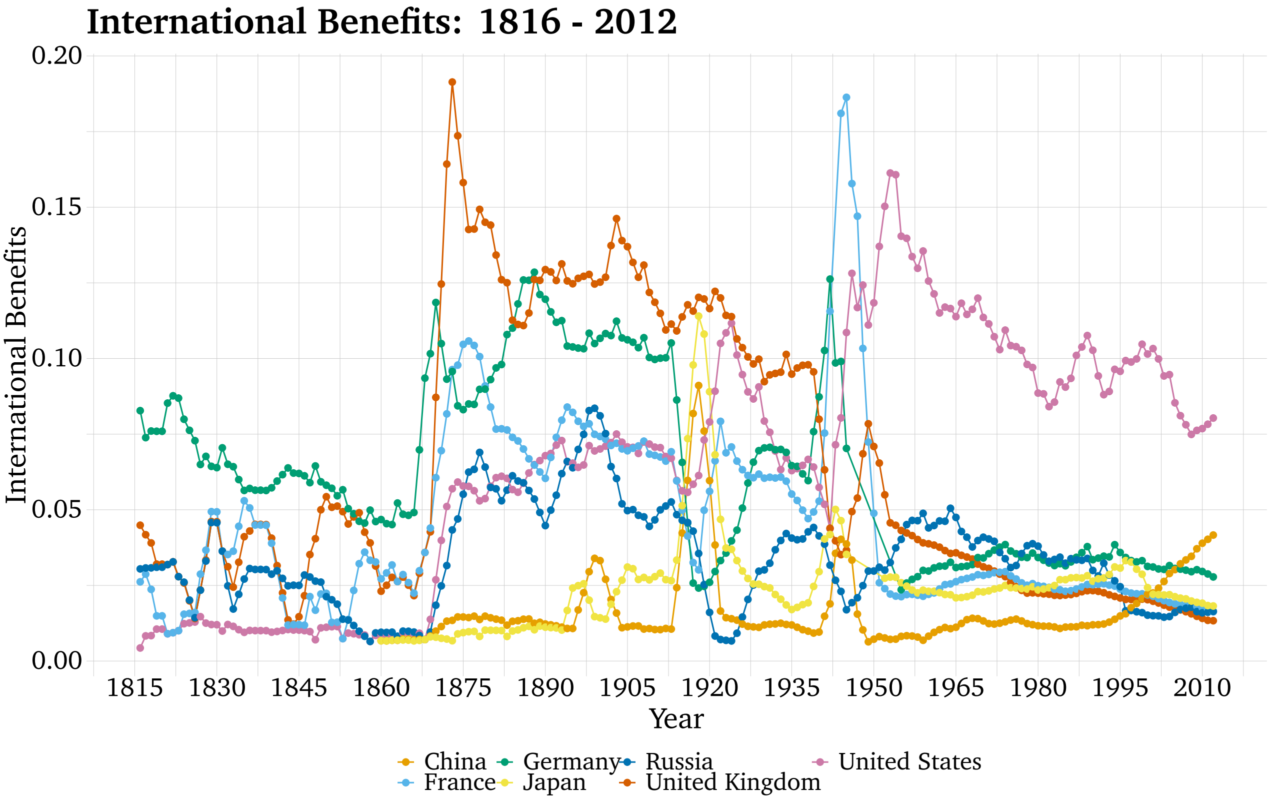

Figure 14: Estimates for a state's *actual* international benefits from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable.

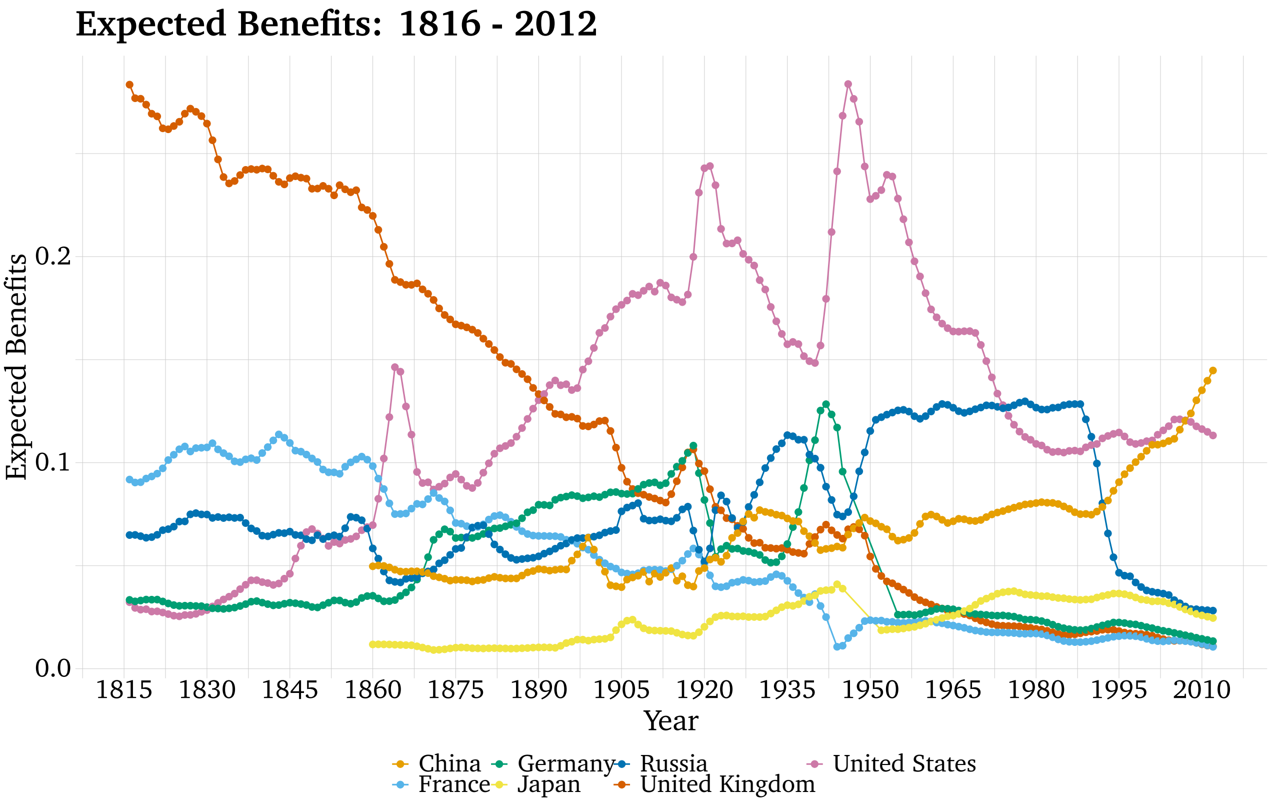

Figure 15: Estimates for a state's *expected* international benefits from 1816-2012. The y-axis is the estimate for each state and the x-axis is the estimate's year. Values are only included for a handful of great power, so that the plot is easily interpretable.