Estimating the performance of international regulatory regimes: Methodology and empirical application to international water management in the Naryn/Syr Darya basin

Tobias Siegfried[1] and Thomas Bernauer[2]

Received 14 November 2006; revised 21 June 2007; accepted 18 July 2007; published 7 November 2007.

[1] We develop a methodology for estimating the performance of international regulatory regimes in the context of transboundary surface waters. Our performance metric relies on assessments, over time, of actual performance, counterfactual performance, and optimal performance. The metric is of practical relevance as a diagnostic tool for policy evaluation. Thus it provides a starting point for policy improvement. To demonstrate the empirical relevance of this methodology we examine international water management in the Naryn/Syr Darya basin, a major international river system in Central Asia. The emphasis is on the Toktogul reservoir, the main reservoir in the Naryn/Syr Darya basin, and its downstream effects. The biggest policy challenge in this case has been to design and implement international trade-offs among water releases for upstream hydropower production in winter and water releases for downstream irrigation in summer. We find that the international regime in place since 1998 is characterized by low average performance and high variability.

Citation: Siegfried, T., and T. Bernauer (2007), Estimating the performance of international regulatory regimes: Methodology and empirical application to international water management in the Naryn/Syr Darya basin, *Water Resour. Res.*, *43*, W11406, doi:10.1029/2006WR005738.

1. Introduction

[2] Even though there is widespread agreement among experts that transboundary waters should be managed in an integrated fashion at the basin scale by robust institutions that involve all riparian countries [*Saleth*, 2004, and references therein] this objective is rarely met. One of the principal obstacles appears to be that, owing to the unequal distributions of hydrological, economic and other endowments, upstream-downstream conflicts are prevalent and cooperation is notoriously difficult to achieve. As an example, the riparian states of the river Rhine struggled over several decades before transboundary water pollution could effectively be reduced [*Bernauer*, 1996]. If upstream-downstream cooperation is very difficult among highly developed, democratic countries, one should expect even greater difficulties in achieving similar levels of cooperation among countries at lower levels of income, with greater levels of domestic political instability, and weaker political and economic ties among each other. The Nile, Euphrates, and Tigris rivers are examples.

[3] Strong variation in the willingness and ability of riparian countries around the world to cooperate in transboundary water resources management has spurred much research in the social, natural and engineering sciences on the conditions that promote or hinder successful cooperation [*Bernauer*, 2002; *Dinar*, 2004]. The existing literature in this area is dominated by many qualitative case studies and a few comparative, quantitative analyses. While the case study approach allows for more nuanced assessments of how effective cooperation is, comparisons over time or across cases remain difficult because assessments are largely qualitative and idiosyncratic [e.g., *Bernauer*, 1996; *Marty*, 2001]. Conversely, some large data sets on transboundary water management have been developed in recent years [*Brochmann and Gleditsch*, 2006; *Wolf*, 1999; *Wolf et al.*, 2005; *Yoffe et al.*, 2004]. Although these data sets allow for statistical comparative analysis they often rely on simplistic measurements of the outcome to be explained, for example, binary measurement of cooperation by existence or absence of or membership in international agreements, treaties, organizations or regimes.

[4] The existing literature offers only very limited concepts for estimating quantitatively the performance of international regulatory regimes in ways that produce results that are easily comparable across cases, that is, basins, and time. Most assessments of performance rely on noncausal criteria. A common approach is to describe the development of a particular problem targeted by a regulatory regime (e.g., pollution) over time and to assess compliance with international obligations in this respect [*Chayes and Handler Chayes*, 1993]. Economic studies define performance chiefly in terms of efficiency related to maximization (e.g., monetary benefits) or minimization criteria (e.g., costs and environmental damages) respectively [e.g., *Cai et al.*, 2003]. In the local or national context, policy performance is sometimes assessed through quasi-experimental research designs and statistical analysis of differences among "treatment" and "nontreatment"

[1]Earth Institute, Columbia University, New York, New York, USA.
[2]Center for Comparative and International Studies, ETH Zurich, Zurich, Switzerland.

0043-1397/07/2006WR005738$09.00

groups [e.g., *Bennear and Coglianese*, 2005; *Greenstone*, 2004]. Such studies require a wealth of data that often does not exist in the international realm. Moreover, the statistical approach to performance measurement is usually not based on a clear notion of what outcomes would be desirable.

[5] Recent work on the effectiveness of international environmental institutions suggests that a sophisticated quantitative approach is feasible [*Helm and Sprinz*, 2000; *Underdal*, 1992; *Young*, 2001]. Such an approach could help in substantively measuring and comparing success or failure in international cooperation in a transparent way. Hence it could provide a more solid foundation for explaining variation in success or failure in this regard. Moreover, it would be of practical relevance for policy evaluation and thus provide a starting point for policy improvement.

[6] In the first part of this paper we develop a methodology for estimating the performance (or effectiveness) of international regulatory regimes, building on work by *Helm and Sprinz* [2000], *Hovi et al.* [2003b], *Sprinz and Helm* [2000], and *Underdal* [1992]. Our policy performance metric is a function of the outcome that should ideally be reached, the actual performance of a given policy at the time of measurement and the outcome that would have occurred in the absence of this policy, that is, the counterfactual performance.

[7] The advantages of this measurement concept are as follows: first, it makes explicit reference to optimal performance and thus problem solving; second, it focuses explicitly on the causal relationship between international policies and outcomes; third, it can be used to assess international policy performance at specific points in time in contexts marked by very little data, but also to assess performance dynamics over time in contexts where more data exist; fourth, cooperative efforts can be disaggregated with reference to particular objectives, policy performance can then be measured for these objectives and aggregated or not.

[8] To demonstrate that our measurement concept is empirically applicable and can provide policy-relevant insights, we examine international water management in the Naryn/Syr Darya basin, a major international river system in Central Asia. The analysis focuses on the management of the Toktogul reservoir, the main reservoir in the Naryn/Syr Darya basin, and its downstream effects. The principal policy challenge in this case has been to design and implement international trade-offs among reservoir discharges for upstream hydropower production in winter and water releases for downstream irrigation in summer.

[9] Our analysis shows that the international management system in place since 1998 is characterized by low average performance and high variability. The principal policy implication is that, even though compliance is high, the management system in place is in urgent need of repair.

2. Problems of the Basic Measurement Concept

[10] The international regimes performance metric as proposed by *Sprinz and Helm* [2000] and *Underdal* [1992] is defined as

$$\Pi = \frac{a - c}{o - c}, \tag{1}$$

where a is an actual, observable variable affected by an international regime among n actors. c is the corresponding counterfactual variable and o is the optimal level of the variable under consideration. This approach to measuring the performance (i.e., effectiveness) of international regulatory regimes is referred to by the authors as the "Oslo-Potsdam Solution." Π can be estimated with relation to any public demand addressed by a public policy. In international water management, for example, such performance assessment may relate to hydropower production, irrigation water provision from transboundary groundwater and surface water resources as well as water quality.

[11] Equation (1) captures the extent to which a given problem has actually been solved ($a - c$) relative to the problem solving potential ($o - c$). The first difference alone only gives an indication that the relevant policy or regime has had some effect. Only by adding the second difference (and o in particular) information on the extent to which the problem has been solved is gained. Π is a standardized measure that allows comparisons across policies within and across policy domains as well as between different measures.

[12] The analysis of Π is difficult since at any given moment we are only able to study one reality, that is, variable a, while equation (1) calls for having observed two alternative realities o and c.

[13] By using the observed world, the problem thus pertains to how to make inferences about hypothetical ones. In other words, to quantitatively assess the performance of international regulatory regimes we need to conjure what has not in fact happened, but might, could or would, under different conditions. Hence a proper understanding of optimality and counterfactual analysis is necessary. Without it, the task of comparative politics seems overly heroic [see also *Przeworski*, 2004].

[14] Surprisingly, the concept of optimality has not yet been adequately dealt with in the relevant literature. For example, Young argues that definitions of the optimum with reference to which regime performance is assessed must not necessarily be based on objective notions, but can depend on understandings of the nature of the problem and the options available for solving the problem [*Young*, 2001]. To address this issue, we submit that o is a particular, commonly agreed upon Pareto-optimal resource allocation out of the set of Pareto-optimal tradeoff allocations Ω that could be arrived at in the presence of a hypothetical, ideal regime. In other words, o is a specific cooperative allocation outcome given perfect actor compliance. In the following, we assume that $o \in \Omega$ is the allocation that provides the greatest benefit (e.g., minimal cost or maximal utility) to each of the n actors; that is, $d_o = |\mathbf{0} - \mathbf{f}(o)|$, where $\{\mathbf{0}, \mathbf{f}(o)\} \in \mathbb{R}^n$ are the origin and the n-dimensional vector valued benefit function, respectively. Finally, d_0 is the minimal or maximal distance correspondingly in the Euclidean space $\mathbb{R}^n$. o thus becomes the most incentive-compatible solution with regard to cooperation [*Binmore*, 1994] (see also *Wu and Whittington* [2006] on incentive compatibility in international river basin management).

[15] Realistically, real world regimes are not optimally designed for various reasons. These can include imperfect knowledge about the state of natural systems as well as their dynamics and inefficient institutional design that is based on compromises rather than objective, scientifically derived

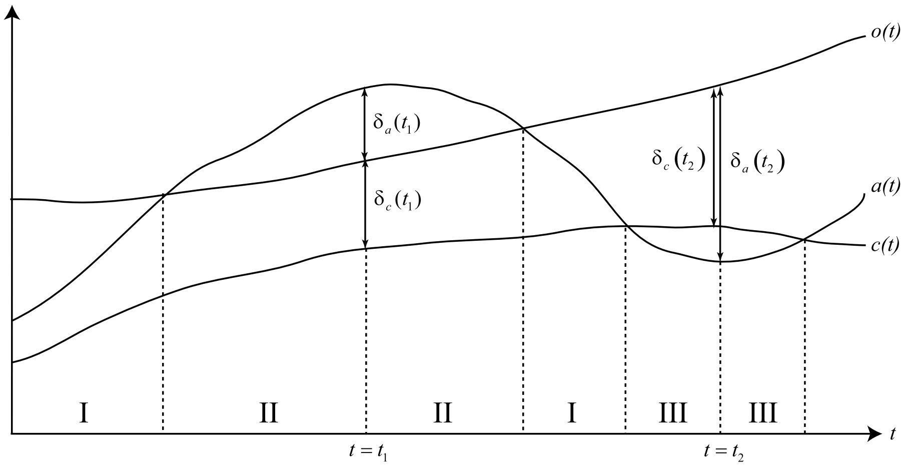

Figure 1. Stylized development of $a(t)$, $c(t)$, and $o(t)$ over time. Values δ_a and δ_c as defined in equation (3) are shown at different times t_1 and t_2. Note that $\delta_a(t) = |\Delta_a(t)|$ and $\delta_c(t) = |\Delta_c(t)|$.

optimality criteria. Consequently, o is rarely directly observable. In such circumstances, Ω can be approximated by multi-objective optimization methods and o subsequently selected [*Siegfried and Kinzelbach*, 2006].

[16] Similarly, a counterfactual assertion might well be understood as conveying a set of predictions under a well defined set of conditions. The noncooperative Nash-equilibrium allocation concept is an appropriate equilibrium approach for estimating c where a regulatory regime is entirely absent and actors individually optimize irrespective of others given above assumptions about underlying relationships. Normally, a finite number of Nash equilibria exist of which normally not all are self-enforcing, that is, stable in repeated interactions [*Van Damme and Weibull*, 1995]. In this case, c can be computed by application of the theory of equilibrium selection by *Harsanyi and Selten* [1988].

[17] Such an enterprise requires the appropriate modeling of existing structural relationships in the observable world. These relationships are assumed to be known and invariant in the systems under consideration (e.g., water, soil, vegetation, and economic arrangement). Furthermore, this approach necessitates known preference orderings of actors as well as known boundary and initial conditions [*Pearl*, 2001]. Given that the aforementioned issues can be addressed, computational modeling allows the study of different hypothetical institutions that would operate under the same conditions.

[18] Π as defined in equation (1) is a strictly increasing or decreasing function of a, depending on the sign of the difference $(o - c)$. This can lead to critical misconceptions if Π is not properly defined as an interval measure. A simple example shall demonstrate this. Imagine that Π is assessed with regard to public demand coverage. Let us further assume that o corresponds to freshwater demand of a particular economic sector. If the actual performance of the international water management regime is suboptimal, that is, $a < o$, we obtain $\Pi < 1$. However, if too much water is allocated to a particular sector and hence wasted, that is, $a > o$, we obtain $\Pi > 1$. This result suggests that wasting resources in allocating ''too much'' is preferable over the allocation of ''too little.'' Both conditions are clearly undesirable from the point of view of economic efficiency. Similar arguments could be made in regard to policy performance in other areas where policies may oversupply public (or collective) goods. By investigating the European regime for Transboundary Air Pollution, Sprinz and Helm provide an example where the effectiveness measure Π as defined in equation (1) fails [see *Sprinz and Helm*, 2000, Table 1].

[19] A correct real-world application of the effectiveness measure Π would thus necessitate an arbitrary scaling of observed values to an ordinal interval scale where, for example, $o = 10$ denotes maximum performance and $c = 0$ is counterfactual performance with $c < a < o$ [e.g., *Rieckermann et al.*, 2007]. Certainly and because of an ad hoc assignment and scaling of a, c, and o, the arbitrariness in such a subjective approach introduces unwanted randomness.

[20] An additional problem is that the basic measurement concept as defined by equation (1) may lead to ad hoc integral assessments over time and thus to wrong conclusions (see Figures 1 and 2). Assume, for the moment, that the estimation of Π at time $t = t_1$ leads to the value $\Pi = \pi_1$, as highlighted in Figure 2. If Π is assessed at time $t = t_2$, performance $\Pi = \pi_2$ is obtained, which clearly differs from the performance value π_1. Inevitably, policy performance usually varies in time since public management efforts include time-varying state and demand variables. Unfortunately, equation (1) does not account for such variability in time. It could, however, be applied at various instances in time.

[21] Simply taking the mean over time, $\bar{\pi}_t$, $t = 0, 1, \ldots, \tau$, could lead to awkward assessments that charge up floods against droughts only to arrive at a balanced performance assessment. Imagine, as another example, that one tries to assess postimpoundment impacts of a large international dam project over a period of 50 years. Assume, furthermore, that the catchment initially benefits from the hydropower production resulting from the dam project. The negative downstream effects on soil and deltaic systems, however, accumulate in time and gradually appear only after some decades. If performance is viewed as a concept related to demand coverage, initial hydropower demand may have been fully met (i.e., $\Pi = 1$). However, subsequent

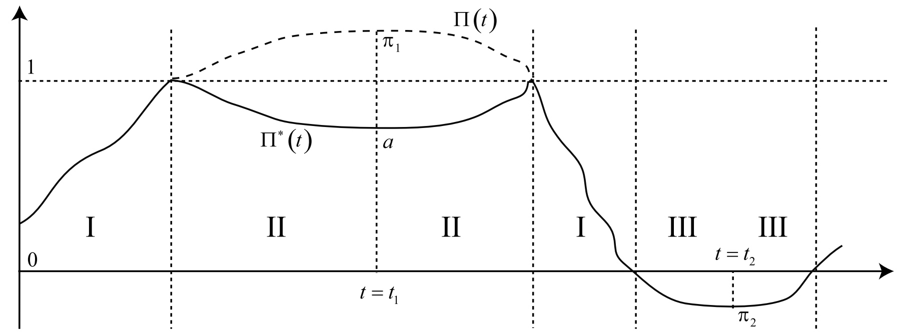

Figure 2. Stylized development of Π and Π* as a time-dependent function of the stochastic processes as depicted in Figure 1. Note that Π > 1 during a certain time interval, which would lead us to assume falsely that during such wasteful allocation the performance of the respective regulatory regime is highest.

demand coverage in respect to downstream environmental services experiences a dramatic decline. Any assessment of Π at a certain time thus provides only a partial picture of performance.

[22] Various scientific disciplines have come up with time-dependent measurement concepts. For example, the water engineering literature has defined several performance criteria (e.g., reliability, resilience, and vulnerability) that account for time dynamics [*Hashimoto et al.*, 1982; *Kjeldsen and Rosbjerg*, 2004]. Similarly, climate science has identified several concepts for assessing computational model forecasting quality [*Nurmi*, 2003]. The importance of accounting for time dependence is also emphasized by *Young* [2001], who argues that a static mode of reasoning leads to ad hoc assessments and introduces arbitrariness. We view the lively debate that followed Young's critique as an expression of the need for more research in this field [*Hovi et al.*, 2003a, 2003b; *Sprinz*, 2005; *Young*, 2003].

3. Upgraded Policy Performance Concept

3.1. Definition

[23] To address the problems discussed in section 2, we propose the definition of a new performance measure as given by

$$\Pi^*\,(t) = 1 - \left|\frac{a(t) - o(t)}{c(t) - o(t)}\right|, \tag{2}$$

where $\Pi^*\,(t)$ is a measure of policy performance at a certain time t. $\Pi^*\,(t)$ measures performance relative to optimal performance $o(t)$ at a specific observation time t. If we use the notation $\delta_a(t) = |\Delta_a(t)| = |a(t) - o(t)|$ and $\delta_c(t) = |\Delta_c(t)| = |c(t) - o(t)|$, then equation (2) becomes

$$\Pi^*\,(t) = 1 - \frac{\delta_a(t)}{\delta_c(t)} \tag{3}$$

by the definition of the absolute value and its properties. If $c(t) < a(t) < o(t), \forall a(t)$ or $c(t) > a(t) > o(t), \forall a(t)$, that is, $c(t)$ and $o(t)$ effectively bound the domain of feasible $a(t)$, it is easy to see that the two performance measures as defined by equations (1) and (2) are equal; that is, $\Pi^*\,(t) = \Pi\,(t)$ (see domains I and III in Figures 1 and 2). Note that $\Pi^*\,(t)$ is symmetric around $o(t)$ and that, according to equation (3), $\Pi^*\,(t)$ is defined as long as $\delta_c(t) \neq 0$ (see also Figure 2). If the performance of an international regulatory regime is optimal at a certain time τ, that is, $a(\tau) = o(\tau)$, and by using equation (3), we obtain

$$\lim_{\delta_a(t)\to 0}\left(1 - \frac{\delta_a(t)}{\delta_c(t)}\right) = 1. \tag{4}$$

Converse to that, if performance is nil, that is, $a(t) = c(t)$, the limit is simply

$$\lim_{\delta_a(t)\to \delta_c(t)}\left(1 - \frac{\delta_a(t)}{\delta_c(t)}\right) = 0. \tag{5}$$

Where $a(t) < c(t) < o(t)$, for example, domain III in Figures 1 and 2, actual regime performance produces outcomes that are less favorable compared to the counterfactual performance and hence $\Pi^* < 0$. If, on the contrary, $c(t) < o(t) < a(t)$, actual performance exceeds optimal levels which describes an inefficient situation as described in section 2 (domain II in Figures 1 and 2). Note that in this case we again have $\Pi^* < 1$.

3.2. Accounting for Temporal Development and Variation

[24] Successive observations in time series data are usually not independent of each other. Effectively, each observation for the measured variable is a bivariate observation with time as the second variable. Variation in time can for example be caused by seasonal variation, trends and irregular fluctuations, or a combination of the above. Most series are stochastic in that future values are only partly determined by past time series values. Simple examples include stochastic rainfall, recharge and runoff processes as well as future per capita and sectoral demand developments.

[25] In our context, we regard the time series $a(t)$, $c(t)$ and $o(t)$ as well as the derived $\delta_a(t)$ and $\delta_c(t)$ as finite realizations of underlying stochastic processes. In the subsequent analysis, we restrict our focus to stationary processes. A process is stationary if the properties of the underlying model do not change. Note, however, that, for example, precipitation patterns need not be particular realizations of stationary processes since climate change can affect the underlying model. However, if the time horizon for performance assessment is short compared to such model changes nonstationarity in this context can be neglected [*Jenkins et al.*, 1994].

[26] The goal is to provide a general and straightforward approach to the characterization of policy performance over a certain period of time by making use of basic concepts and definitions of probability theory and statistics. This approach assumes neither knowledge of the underlying probability distribution functions, nor of the stochastic processes that eventually produce $a(t)$, $c(t)$ and $o(t)$. We submit that the expected value $\langle \Pi^* \rangle$ as well as the variance $\sigma^2_{\Pi^*}$ of $\Pi^*(t)$ are two descriptions that permit a useful characterization of regime performance over time.

[27] To find an expression for $\langle \Pi^* \rangle$ we use a first-order Taylor approximation to linearize equation (3) around the mean μ_{δ_c} of $\delta_c(t)$, assuming that $\delta_c(t)$ is sufficiently well behaved in the neighborhood of μ_{δ_c}. Hence we obtain

$$\Pi^*(t) \approx 1 + \frac{\delta_a(t)}{\mu_{\delta_c}} + \frac{\delta_a(t)\left(\delta_c(t) - \mu_{\delta_c}\right)}{\mu^2_{\delta_c}} + O\left[\delta_c(t) - \mu_{\delta_c}\right]^2. \quad (6)$$

[28] If we drop the second- and higher-order terms $O[\delta_c(t) - \mu_{\delta_c}]^2$ in equation (6) and after simplification, we get

$$\langle \Pi^* \rangle = 1 - \frac{\mu_{\delta_a}}{\mu_{\delta_c}} + \frac{1}{\mu^2_{\delta_c}} \mathrm{Cov}(\delta_a, \delta_c), \quad (7)$$

where $\mathrm{Cov}(\delta_a, \delta_c)$ denotes the covariance and μ_{δ_a} as well as μ_{δ_c} the mean of the time series $\delta_a(t)$ and $\delta_c(t)$. Note that in equation (7) we dropped the time subscripts for notational convenience.

[29] $\langle \Pi^* \rangle$ is not defined for $\mu_{\delta_c} = 0$. At such level and circumstances, that is, $c(t) = o(t)$, $t = 0, 1, \ldots \tau$, policy makers would probably not initiate a new policy since any deviation from the status quo would affect the performance measure negatively. Again, in the case of optimality, that is, $a(t) = o(t)$, $\langle \Pi^* \rangle = 1$ since $\delta_a(t) = 0$ for all $t = 0, 1, \ldots \tau$ and hence $\mu_{\delta_a} = 0$. Therefore $\mathrm{Cov}(\delta_a, \delta_c) = 0$, which follows from $\mathrm{Cov}(\delta_a, \delta_c) = \langle \delta_a(t) \cdot \delta_c(t) \rangle - \langle \delta_a(t) \rangle \langle \delta_c(t) \rangle = \langle 0 \cdot \delta_c(t) \rangle - 0 \cdot \mu_c = 0$.

[30] The derivation of the variance $\sigma^2_{\Pi^*}$ of $\Pi^*(t)$ is as follows. According to the standard textbook definition

$$\sigma^2_{\Pi^*} = \left\langle \Pi^*(t)^2 \right\rangle - \langle \Pi^*(t) \rangle^2. \quad (8)$$

[31] By using equation (7), the second term on the right-hand side of equation (8) is

$$\langle \Pi^*(t) \rangle^2 = \left(1 - \frac{\mu_{\delta_a}}{\mu_{\delta_c}} + \frac{1}{\mu^2_{\delta_c}} \mathrm{Cov}(\delta_a, \delta_c)\right)^2 = \frac{\left(\mathrm{Cov}(\delta_a, \delta_c) + \mu_{\delta_c\left(\mu_{\delta_c} - \mu_{\delta_a}\right)}\right)^2}{\mu^4_{\delta_c}}. \quad (9)$$

[32] Similarly, by using equation (3), the first term on the right-hand side is

$$\left\langle \Pi^*(t)^2 \right\rangle = 1 + \frac{2 \cdot \mathrm{Cov}(\delta_a, \delta_c) + 4\mu^2_{\delta_a} - 2\mu_{\delta_a}\mu_{\delta_c} + 4\sigma^2_{\delta_a}}{\mu^2_{\delta_c}} + \frac{\left\langle \delta^2_a \delta^2_c \right\rangle}{\mu^4_{\delta_c}} - \frac{4\left\langle \delta^2_a \delta_c \right\rangle}{\mu^3_{\delta_c}}. \quad (10)$$

[33] $\langle \Pi^*(t)^2 \rangle$ cannot be calculated without knowledge of the underlying probability distribution functions of $a(t)$, $c(t)$ and $o(t)$ since third- and fourth-order moments have to be determined (last two terms in equation (10)). However, we can again linearize these terms. By doing so, after a somewhat tedious calculation, we obtain for the individual higher-order terms

$$\frac{\left\langle \delta^2_a \delta^2_c \right\rangle}{\mu^4_{\delta_c}} \approx \frac{\mu_{\delta_a}\left(4\mathrm{Cov}\left(\mu_{\delta_a}, \mu_{\delta_c}\right) + \mu_{\delta_a}\mu_{\delta_c}\right)}{\mu^3_{\delta_c}}, \quad (11)$$

and

$$\frac{4\left\langle \delta^2_a \delta_c \right\rangle}{\mu^3_{\delta_c}} \approx \frac{\mu_{\delta_a}\left(2\mathrm{Cov}\left(\mu_{\delta_a}, \mu_{\delta_c}\right) + \mu_{\delta_a}\mu_{\delta_c}\right)}{\mu^3_{\delta_c}}. \quad (12)$$

[34] Subtracting the right-hand side of equation (9) from the one in equation (10), we finally obtain

$$\sigma^2_{\Pi^*} = \frac{4\sigma^2_{\delta_a}}{\mu^2_{\delta_c}} - \frac{\mu^2_{\delta_a}\sigma^2_{\delta_c}}{\mu^4_{\delta_c}} - \frac{\mathrm{Cov}(\delta_a, \delta_c)^2}{\mu^4_{\delta_c}} - \frac{2\mathrm{Cov}(\delta_a, \delta_c)\mu_{\delta_a}}{\mu^3_{\delta_c}}. \quad (13)$$

In equations (7) and (13), μ_{δ_a}, μ_{δ_c}, ${\sigma_{\delta_a}}^2$, ${\sigma_{\delta_c}}^2$ and $\mathrm{Cov}(\delta_a, \delta_c)$ have to be empirically estimated from available data (see, e.g., *Loucks et al.* [1981] for a detailed explanation of the standard estimation procedure).

[35] Note that if the time series a(t), c(t) and o(t) are available, $\Pi^*(t)$ and the corresponding statistics can be empirically calculated without using equations (7) and (13). The usefulness of these equations becomes apparent in situations of restricted data availability where only first and second moment characteristics of the observed variables are known or in situations where a(t), c(t) and o(t) are of unequal length (see case study below).

[36] *Young* [2001] states that procedures involving counterfactual analysis to assess international regime effectiveness have rarely been applied in a transparent and systematic fashion. According to him, they have relied too much on subjective judgments in scoring individual cases based on simplistic categories. We submit that the upgraded measurement concept presented above addresses the most important shortcomings of the approach proposed by [*Sprinz and Helm*, 2000]. We now demonstrate the empirical relevance of our methodology by examining international water management in the Naryn/Syr Darya basin, a major international river system in Central Asia.

4. Application to International Water Management

4.1. Case Study

[37] The Naryn/Syr Darya river system is part of the Aral Sea basin; the other main river of this basin is the Amu Darya. The size of the Aral Sea basin is approx. 1.55 million km^2, its population around 40 million. The economies of the riparian countries (Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan, and Turkmenistan) are heavily dependent on irrigated agriculture (with average shares of 40–50% of GDP in 1960–1990, and around 20–30% thereafter). Farming employs ca. 60% of the rural population and 25–60% of the total labor force [*World Bank*,

Figure 3. Map showing the part of the Naryn and Syr Darya catchment that is of most interest in this paper. The Uch Kurgan gauge station is located in the center of the map. The Toktogul reservoir is located at the top of the Naryn/Syr Darya cascade in Kyrgyzstan.

1996]. While upstream parts of the basin are mostly mountainous and humid, the midstream and downstream areas are arid (low-frequency and irregular, high-intensity precipitation with large daily and seasonal temperature differences). Over the past 40 years, excessive water withdrawals have led to a drastic shrinkage of the Aral Sea; the latter receives the bulk of its water from the two Daryas. The Aral Sea has thus been reduced to around 25% of its original volume and has received worldwide attention as an ecological disaster zone [*Dukhovny and Sokolov*, 2005].

[38] The Syr Darya river originates as the Naryn river in the mountains of Kyrgyzstan (see map in Figure 3). It then flows through Uzbekistan and Tajikistan and ends in the Aral Sea in Kazakhstan (total length around 2800 km). In total, approximately 20 million people inhabit this river catchment, which covers an area of ca. 250,000 km^2. The river is mainly fed by snowmelt and water from glaciers. Approximately 75% of the runoff stems from Kyrgyzstan. The natural runoff pattern, with annual flow ranges of 23.5–51 km^3, is characterized by a spring/summer flood. It usually starts in April and peaks in June. Nowadays, around 93% of the Syr Darya's mean annual flow is regulated by storage reservoirs (in the whole Naryn/Syr Darya basin, total usable reservoir capacity is around 27 km^3) [*Dukhovny and Sokolov*, 2005].

[39] Water abstraction from the Syr Darya basin is mainly for irrigated agriculture. Of the approx. 3.4 million ha of irrigated farm land around 1.7 million ha is irrigated with water taken directly from the river. Figure 4 shows the time series of the Naryn/Syr Darya river flow over the last 72 years as measured at Uch Kurgan gauge station, Uzbekistan (see Figure 3 for the location of the latter).

[40] The runoff of the Naryn/Syr Darya, as measured at the Uch Kurgan gauge station, that is, at the foot of the Naryn/Syr Darya cascade shortly after the river enters Uzbekistan from Kyrgyzstan, varies strongly over time. It is marked by four distinct periods as shown in Figure 4. During the phase of largely natural runoff (1933–1974), mean flow was 388 m^3/s, with a high variability in summer (see Figure 6 for mean monthly flows as well as Table 1 for key hydrologic data at Uch Kurgan gauge). In this period, the pronounced temporal differences in flows are determined by seasonal and climatic variability.

[41] A substantial change in downstream flow patterns occurred with the commissioning of the Toktogul dam in 1974. The Toktogul dam is by far the largest storage facility in the Aral Sea basin. It has 14 km^3 effective capacity, 8.7 km^3 firm yield and a full capacity of ca. 19.5 km^3 (Figure 5 shows storage volumes in 1974–2005). The reservoir area is around 280 km^2, its length approximately 65 km. Hydropower capacity of the Toktogul power plant is 1200 MW, that is, the second biggest in the Aral Sea basin [*Antipova et al.*, 2002]. After the commissioning of the dam, mean flow was reduced to 311 m^3/s, mainly owing to the filling of the Toktogul reservoir. If one assumes an average of 14 km^3 dam storage volume to be filled at a rate of 70 m^3/s (i.e., the difference between pre-impoundment and postimpoundment mean flow at Uch Kurgan), this amounts to an approximate filling time of 6.3 years, which corresponds well with the observations (see Figure 5). After dam closure, a general attenuation of peak downstream flows as well as an overall decline of monthly flow variability was observed (see Figure 4). This decline was most pronounced in the summer months.

[42] The centralized Soviet management system was oriented primarily toward adequate water provision for irrigated agriculture (above all, cotton production) in Uzbekistan and Kazakhstan. The timing of winter and summer flow releases did not change substantially compared to the natural runoff pattern. Management and its infrastructure were fully funded from the federal budget of the USSR. In consultation with the governments of the five republics and on the basis of forecasts by the Central Asia Hydromet Service, the ministry of water resources (Minvodgoz) in Moscow defined annually (on the basis of a multiyear master plan) how much water was to be released for

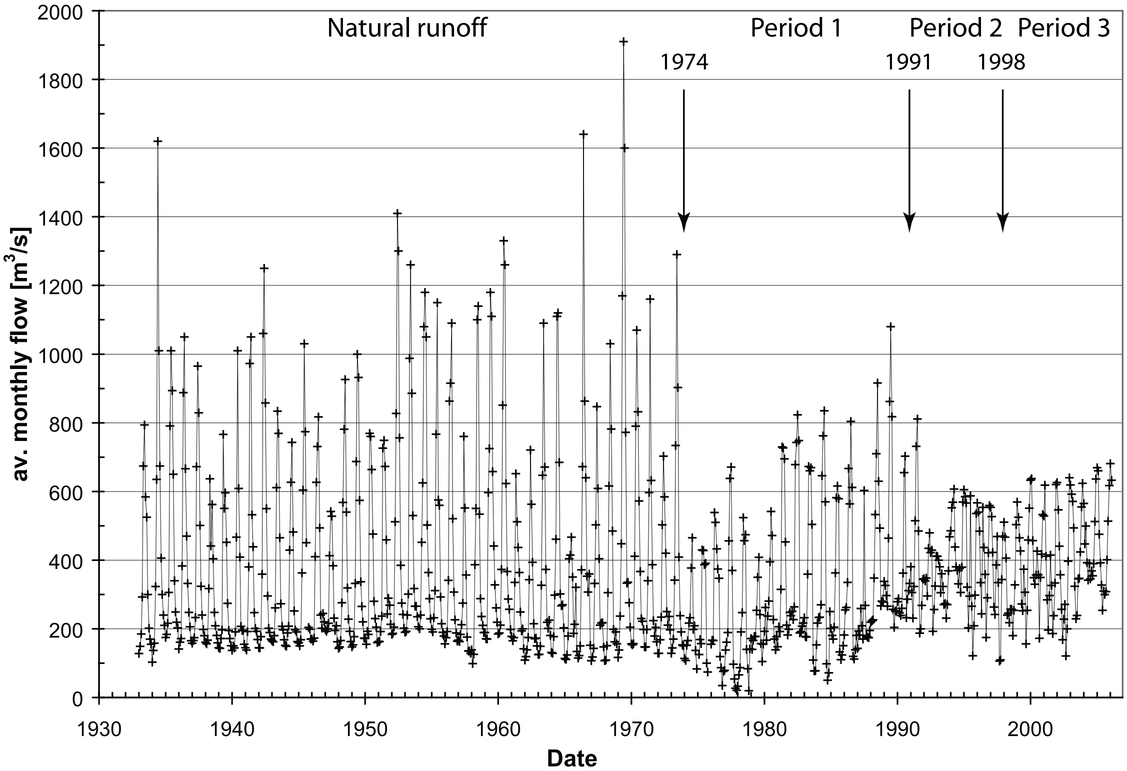

Figure 4. Mean monthly flow of Naryn/Syr Darya river at Uch Kurgan gauge from January 1933 to February 2006. The four flow patterns, including pre-Toktogul (1933–1974), USSR Naryn/Syr Darya management (Period 1: 1974–1990), post-USSR operation (Period 2: 1991–1997), and new Inter-State Commission for Water Coordination (ICWC) regime (Period 3: 1998–today), are clearly visible in the time series. Data are from Global Runoff Data Center (GRDC) and A. Yakovlev (Uzbek Hydrometeorological Service, unpublished data, 2006).

irrigation during the growing season (April to September) to each water management region. The goal was basin-wide, Pareto-optimal allocation with regard to the provision of water for the irrigation agriculture centers that depend on the Syr Darya's water.

[43] The electricity produced at Toktogul during that period went into the Central Asian Energy Pool (CAEP) and was thus shared among the riparian republics. In exchange, the neighboring republics supplied coal, oil, and natural gas to Kyrgyzstan in winter to cover the increased Kyrgyz energy demand during the colder months. The fossil fuel was used primarily in thermal power plants in Bishkek and Osh [*Cai et al.*, 2002].

[44] The collapse of the Soviet Union in 1991 led to the breakdown of centralized water resources management and water-energy trade-off arrangements, causing serious dis-

Table 1. Intraperiod Monthly Mean Flows and Corresponding Standard Deviations at Uch Kurgan Gauge in Uzbekistan[a]

	Natural Runoff Regime		Period 1 (1974–1990)		Period 2 (1991–1997)		Period 3 (1998–2005)		
Month	μ	σ	μ	σ	$\mu(c(t))$	σ	μ	σ	Optimal $\mu(o(t))$
1	150.9	27.7	183.9	74.5	478.5	101.1	590.0	55.3	357.7
2	152.2	24.8	196.5	68.9	464.2	113.0	561.8	78.6	426.2
3	178.8	28.6	192.9	49.9	428.9	122.1	465.8	52.9	323.4
4	318.4	92.5	265.7	94.7	350.2	115.6	367.0	79.6	426.2
5	672.2	190.5	443.6	189.7	348.0	120.2	286.8	52.0	452.8
6	987.3	325.8	532.1	205.2	450.1	152.6	270.6	73.8	468.0
7	807.9	258.8	638.5	210.1	481.0	174.5	324.3	78.2	494.7
8	518.1	138.4	518.4	149.6	354.1	79.5	316.6	40.3	490.9
9	289.3	71.2	184.9	96.7	198.5	89.2	228.1	93.0	441.4
10	231.0	49.3	146.5	72.4	234.5	67.7	313.7	86.8	300.6
11	219.0	44.4	143.8	89.0	343.5	51.9	439.4	84.9	304.4
12	176.3	26.8	181.9	80.9	479.7	82.3	590.6	53.0	418.6
Overall	388	307	311	215	384	139	396	141	409

[a]The bottom row displays overall means and standard deviations for the duration of the management periods. Units are m^3/s for μ and σ. The last column shows data from *Cai et al.* [2003]. Mean flows in Period 1 are reduced owing to the effect of reservoir filling. See also Figure 6.

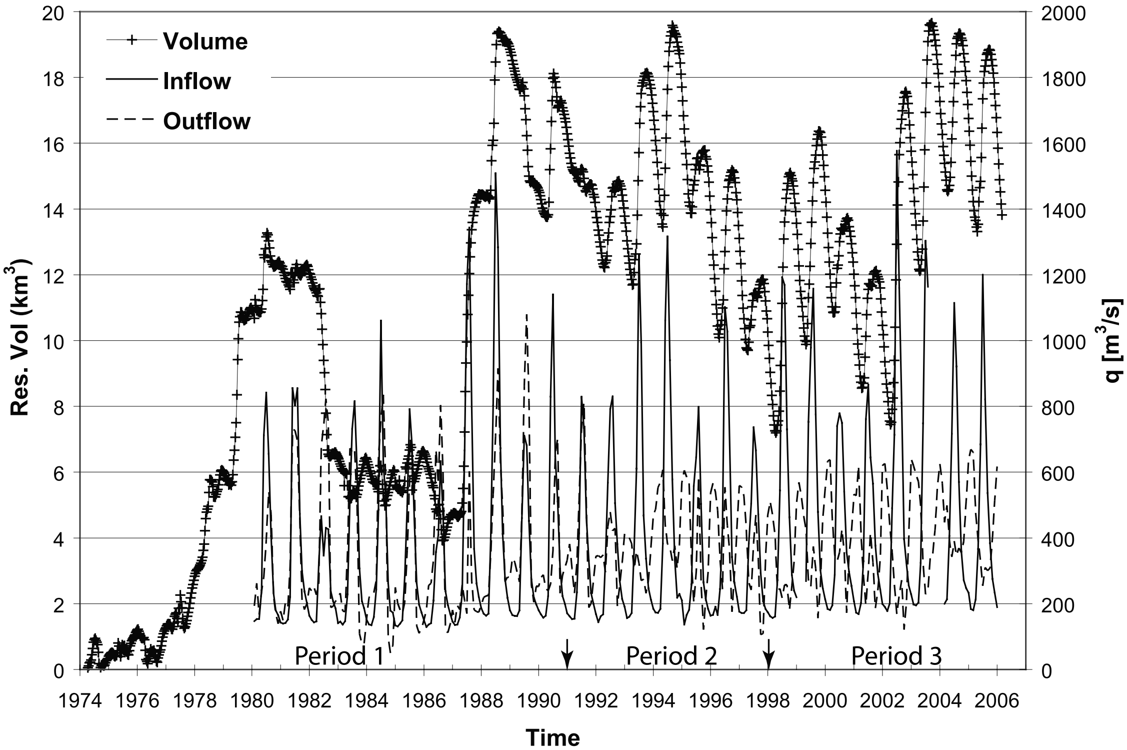

Figure 5. Inflow and outflow of the Toktogul reservoir and reservoir volume since dam closure. Note that the full reservoir capacity was not reached until the beginning of high-inflow years (starting in July 1987). The switch from irrigation to power production mode is clearly visible after 1992 (see also Figure 4). Reservoir inflows and outflows are only shown from 1980. Data are from A. Yakovlev (unpublished data, 2006).

putes between the newly independent states over water allocation issues. Coal, oil, natural gas, and electricity supplies to Kyrgyzstan declined dramatically between 1991 and 1997, and so did the thermal and electric power output of Kyrgyz thermal power plants [*Antipova et al.*, 2002].

[45] Consequently, consumers in Kyrgyzstan turned to electricity, which increased winter demand by more than 100% compared to the Soviet era. Purchases of energy from abroad were (and still are) difficult because the government was (for political and administrative reasons) unable to raise and collect appropriate energy tariffs. Moreover, financial contributions from Moscow and the former republics in the basin for the maintenance of the reservoir ceased. In response to the sharp drop in thermal power output and rising winter demand for electricity, Kyrgyzstan switched the operation of the Toktogul reservoir from irrigation to electric power production mode. Since winter 1993, the flow peaks no longer occur in summer but rather in winter (see Figures 4 and 5).

[46] The core of the upstream-downstream problem in the context of the Syr Darya is that upstream interests deriving from temporal water demands are diametrical to downstream water demands and interests. Kyrgyzstan uses very little water consumptively, that is, for irrigation. Since the country has no fossil fuel sources of its own it is interested in producing hydro-electricity at the Toktogul electric power plant, particularly in winter when energy demand is high. This interest has become ever stronger as the downstream countries have cut back on energy supplies to Kyrgyzstan. Kyrgyzstan also views electricity production as a potential export commodity. It is thus eager to store water in spring to autumn and release it in winter to spring for energy production. Conversely, downstream Uzbekistan and Kazakhstan, by far the largest consumers of irrigation water in the river basin, are interested in obtaining much more water during the growing season (April to September) than in the non−growing season (October to March). They are also interested in electricity produced upstream through water release during the growing season for operating irrigation pumps. Moreover, from the perspective of downstream countries, water releases in winter should be rather low, for high flows may cause floods because ice in the river bed reduces water flow capacity. The principal problem to be solved thus pertains to coordinating the management of the Naryn/Syr Darya cascade of reservoirs that are located in Kyrgyzstan, and in particular the handling of trade-offs between consumptive water use for downstream irrigation purposes and nonconsumptive use for upstream energy production in Kyrgyzstan. Ever since 1991, the riparian countries have been struggling to re-establish an effective management scheme [*Savoskul et al.*, 2003].

[47] In February 1992 the five newly independent states set up the Inter-State Commission for Water Coordination (ICWC). In 1993 the International Fund for Saving the Aral Sea was added to the ICWC (available at http://www.icwc-aral.uz/). The five countries agreed to keep the water allocation principles of the former USSR system in place until a new system could be established, albeit without the funding for the infrastructure that had formerly come from

Table 2. Monthly Release Schedule $s(t)$ of Toktogul Reservoir as Established in the 1998 Treaty

Month[a]	q, m^3/s
1	495
2	490
3	300
4	230
5	270
6	500
7	650
8	600
9	190
10	-
11	-
12	-

[a]No values were defined for the months of October to December.

Moscow. Furthermore, the most important hydraulic structures, and in particular the biggest reservoirs in the basin (including the Toktogul), were de facto nationalized by the newly independent countries.

[48] Several declarations by the riparian countries and attempts by European and North American government agencies to help in the problem-solving effort produced only minimal progress. Only in March 1998, under the aegis of the Executive Committee of the Central Asian Economic Community and assisted by USAID, Kazakhstan, Kyrgyzstan, and Uzbekistan signed a formal agreement that marks the beginning of management Period 3. In 1999, Tajikistan joined this agreement (available at http://ocid.nacse.org/cgi-bin/qml/tfdd/treaties.qml). The agreed monthly release schedule $s(t)$ is shown in Table 2 and Figure 6.

[49] The 1998 agreement holds that during the vegetation period, Kyrgyzstan must release more water than it needs for its own hydropower demands, and that the energy surplus is distributed to Kazakhstan and Uzbekistan. In the non–growing period (1 October to 1 April) Uzbekistan and Kazakhstan supply Kyrgyzstan with energy resources in amounts that are approximately equivalent to the electricity they receive from Kyrgyzstan during the growing season. The exact amounts of water and energy are defined annually through negotiations among the countries. Typically, Kyrgyzstan has been scheduled to release around 6.5 km^3 of water during the vegetation period and transfer around 2.2 M kWh of electricity to Uzbekistan and Kazakhstan. We now evaluate the performance of the international regulatory regime.

4.2. Assessment of Performance

[50] In order to assess the 1998 regime performance, we assume in the following that flows are an adequate proxy for the benefits or losses to the individual countries. In doing so, we have to acknowledge that sensitivity with regard to release patterns as well as flows exists to a variable degree. As an example, upstream Kyrgyzstan is highly dependent on timely hydropower production owing to the limited capability to store electricity as well as owing to the lack of alternative supplies. Similarly, Uzbekistan crucially depends on the availability of irrigation water during the growing season in the summer months. Yet, from the downstream perspective, a negative deviation from optimal surface water flows can be partially covered by increased groundwater pumping in certain regions for at least a limited period of time. With that, agricultural yield losses can potentially be averted.

[51] The availability of detailed production data would allow mapping production inputs into benefits. Actual, optimal and counterfactual performance thus would be expressed as potentially nonlinear functions of flows $q(t)$ and other state variables $x(t)$, for example, $a(t) = f(q(t), x(t))$. In the absence of a clear specification of the returns to scale in production as well as information on factor substitutability and associated costs, the quantification of economic effects associated with an inadequate coverage of the

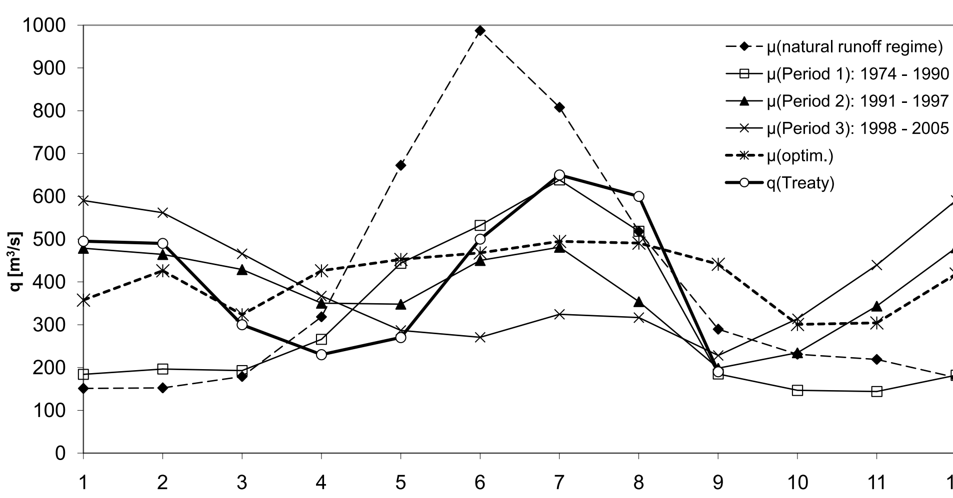

Figure 6. Monthly long-term average flows at the Uch Kurgan gauge (based on data from GRDC and A. Yakovlev, unpublished data, 2006). The data on flow variability for the corresponding months and periods are provided in Table 1. The monthly data μ(optim.) are calculated optimal releases from the Naryn/Syr Darya cascade. Optimization was carried out with a coupled hydrologic-agronomic-economic model on the basin scale by *Cai et al.* [2003]. The 1998 agreement flow volumes are shown for comparison.

electricity and irrigation water demands respectively is, however, difficult. Thus, for the performance assessment, we directly use flow data as surrogates for benefits and losses.

[52] In the following, we assume that the postimpoundment flow at the Uch Kurgan gauge is an outcome of national or international policy respectively. As discussed in section 4.1 and seen in Figures 4–6, three distinct management periods can be identified: Period 1, 1974–1990; Period 2, 1991–1997; and Period 3, 1998–2006. These periods are characterized by differing Syr Darya flow patterns below the Toktogul reservoir (see Figure 6). They are associated with the timeline of political events.

[53] In the period of centralized water resources management under USSR rule (Period 1, 1974–1990), the characteristics of the monthly averages do not differ substantially from the natural flow, with a summer discharge peak and winter low flow. Yet, owing to the filling of the reservoir, the summer peak is less pronounced. This characteristic flow pattern changes after the breakdown of central governance (μ(Period 2) in Figure 6). As discussed above, the increased hydropower demand in upstream Kyrgyzstan led to a pronounced increase of reservoir water releases in the winter months. Furthermore, a generally high variability in monthly flows reflects political turmoil in this phase of transition and unilateral runoff management (see σ of Period 2 in Table 1). Finally, with the implementation of the 1998 agreement in Period 3, flows appear to reflect the trade-offs made in that agreement with a considerable decline in monthly variability compared to the prior period. However, it is also visible that treaty adherence is suboptimal and that reservoir releases are skewed toward the winter period (see Figure 6).

[54] For the performance assessment of the 1998 agreement in Period 3, we define the current flow regime as actual performance $a(t)$. *Cai et al.* [2002, 2003], *McKinney and Cai* [1997], and *McKinney et al.* [1999] developed an optimization modeling framework for integrated river basin management. This optimization model considers risk minimization in water supply, environmental conservation of soil and water resources, spatial and temporal equity in water allocation and economic efficiency in the development of future water infrastructure. The model uses long-term decisions based on quantifiable sustainability criteria to guide short-term water allocation decisions between the upstream and downstream as well between sectors. The authors applied their model to the Syr Darya basin and investigated a full optimization scenario under the assumption of long-term average precipitation in the basin from 1968–1998. This scenario then determines monthly reservoir releases, infrastructure development, irrigated crop patters and area with the objective to maximize the resulting sum of irrigation and ecological benefits and hydropower profits. The resulting optimal monthly flows at Uch Kurgan are shown in Figure 6, that is, μ(optim.), and are utilized in this study to define the optimal performance $o(t)$ (see also Table 1).

[55] As discussed in section 2, measurement of counterfactual performance $c(t)$ can be based on the assumption of unconstrained (by the downstream countries or actors from outside the basin) maximization by Kyrgyzstan of hydropower production to cover domestic energy needs and export excess energy to obtain foreign currency, and to carry out a simulation-optimization (from the Kyrgyz perspective) on that basis. Discussions with experts on the region led us to the conclusion that such a scenario would have been very unlikely given that the downstream countries are economically and militarily more powerful and thus pose a credible threat of future intervention in a scenario of unilateral upstream water allocation. The notion of counterfactual analysis in this context should therefore be based on observed flows during Period 2, that is, from 1991–1997. The period of breakdown of the centralized management system in 1991–1997 (Period 2), that is, the period when there was no international regime, is consequently defined as counterfactual performance, that is, $c(t)$.

[56] The calculation of the regime performance values $\Pi^*(t)$, $\langle\Pi^*\rangle$ and $\sigma^2_{\Pi^*}$ (equations (2), (7), and (13)) based on $a(t)$ and $c(t)$ as defined above may be problematic. While o(t) is based on long-term average supply and demand values respectively, the underlying, implicit assumption for $a(t)$ and $c(t)$ is that boundary conditions have not changed from 1991–2005 so as not to violate causality (see section 2). This assumption is not valid from the perspective of available freshwater supply. Figure 7 shows mean precipitation values in the Syr Darya catchment for the individual periods. The period from 1998–2005 is characterized by increased mean precipitation. Compared to the long-term mean annual precipitation from 1968–1998, the mean values are approx. 30% higher from 1998–2005 in the Syr Darya subcatchments that contribute to runoff at Uch Kurgan gauge. Table 3 summarizes the detailed monthly precipitation statistics for the individual periods based on NCEP/NCAR reanalysis data [*Kalnay et al.*, 1996].

[57] Yet, from the management perspective, the Toktogul reservoir acts as a buffer that protects against short-term fluctuations in interseasonal water availability due to the natural variability in precipitation. Hence, except in cases of exceeded maximum reservoir capacity or insufficient reservoir filling, the demand side is the crucial driver of reservoir releases. With regard to irrigation water demand in downstream Uzbekistan, key characteristics did not change dramatically over the last 15 years (for relevant data, see http://www.fao.org/ag/AGL/aglw/aquastat/countries/uzbekistan/index.stm).

[58] In upstream Kyrgyzstan however, electricity demand from hydropower increased as described above. With that, the fraction of total energy supply derived from hydropower production increased by roughly 25% over this period (compared to 1991 levels) since electric power production from Kyrgyz thermal power plants in 1991–1997 declined from 3.9 to 1.6 M kWh (see, for example, https://www.cia.gov/cia/publications/factbook/geos/kg.html). Hence we have to acknowledge this measurable change in demand boundary conditions from the upstream perspective between periods $c(t)$ and $a(t)$. Note, however, that relative increase in demand form 1998–2005 is comparable to the relative increase in supply over the same period as discussed above. Therefore the comparison of $c(t)$ with $a(t)$ remains valid.

[59] To compute the performance $\Pi^*(t)$ of the international regime installed in 1998 we use averaged within period monthly flow values $\mu(o(t))$ and $\mu(c(t))$ for the optimal and counterfactual performance (see Table 1). Averaging over the respective periods is necessary since comparing individual hydrological years with different

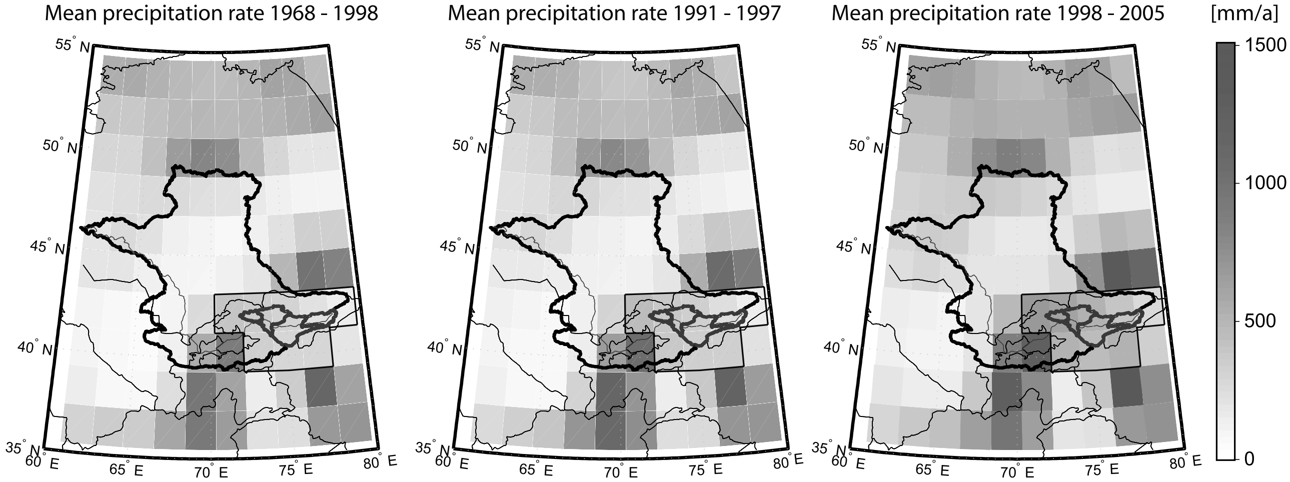

Figure 7. Mean precipitation rates for the periods of optimal (1968–1998), counterfactual (1991–1997), and actual performance (1998–2005). Data are from National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) [see *Kalnay et al.*, 1996]. The black bold line delineates the entire Syr Darya catchment. The subcatchments contributing to runoff at the Uch Kurgan gauge station are shown by bold gray lines. Eight NCEP/NCAR grid cells cover this area as outlined in the figures. Mean per period precipitation rates over these grid cells are 291.3 mm/a, 322.3 mm/a, and 416.0 mm/a, respectively (see also Table 3).

resource endowments (i.e., inflow as well as reservoir levels) and demands (electricity as well as irrigation water) would lead to an arbitrary comparison of reservoir outflows between years that are not necessarily comparable with respect to the above mentioned state variables.

[60] The temporal development of Π^* (t) is shown in Figure 8. As can be seen there, performance of the 1998 regime has been poor with a declining trend over the assessment period. In particular, Figure 8 shows that extremely negative values of Π^* (t) start to occur from 2002 onward, usually in September. This can be explained by the fact that in this month, $|\mu(c(t)) - \mu(o(t))|$; that is, the denominator of Π^* (t) is small, and the difference between actual performance and the monthly averaged performance of Period 1, that is, $|a(t) - \mu(o(t))|$, is large. At the same points in time, the original performance measure Π as defined in equation (1) suggests performance values of $\Pi = \{-13.7, 26.6, 28.0, 19.7\}$. The positive Π values are clearly nonsensical. They thus indicate the inherent problem associated with the utilization of Π as proposed by *Helm and Sprinz* [2000].

[61] To calculate $\langle \Pi^* \rangle$ and $\sigma^2_{\Pi^*}$, we define $\delta_a = |a(t) - \mu(o(t))|$ and $\delta_c = |\mu(c(t)) - \mu(o(t))|$, respectively, for reasons explained above. Note that δ_a is a 7 year monthly time series and δ_c is a 1 year monthly time series. The sample means μ_{δ_a} and μ_{δ_c}, the variances $\sigma^2_{\delta_a}$ and $\sigma^2_{\delta_a}$ as well as the covariance $\mathrm{Cov}(\delta_a, \delta_c)$ were subsequently calculated under the assumption that the monthly values of δ_c do not change over the period of actual performance. The estimates are shown in Table 4. For the estimated covariance, we get $\mathrm{Cov}(\delta_a, \delta_c) = 1250.1$ m^6/s^2. By using equations (7) and (13), we finally obtain for the average regime performance $\langle \Pi^* \rangle = -0.71$ and the variance $\sigma^2_{\Pi^*} = 0.92$. As concluded already from visual inspection of Π^* (t) in Figure 8, overall performance of the 1998 regime is poor indeed. This, together with a highly variable performance is certainly not conducive to regime stability.

[62] It is interesting to compare the above findings with the results from a compliance-oriented assessment approach. To this end, we compute the ratio $r(t)$ of targets for the respective months as defined in the 1998 agreement and the actual water releases from the Toktogul reservoir. More precisely and so as to treat positive and negative deviations from agreed upon releases similarly, compliance $r(t)$ is defined as $r(t) = 1 - |\Delta(t)|/s(t)$ with $\Delta(t) = s(t) - a(t)$ being the difference between treaty and actual monthly

Table 3. Precipitation Statistics for the NCEP/NCAR Grid Cells Depicted in Figure 7[a]

Period	Month	1	2	3	4	5	6	7	8	9	10	11	12
1968–1998	μ	194.5	251.7	363.2	381.7	491.8	511.6	308.3	317.0	210.5	150.1	141.7	173.9
	σ	58.3	94.8	101.3	123.8	222.4	201.8	160.9	212.1	98.8	80.4	83.8	69.2
1991–1997	μ	204.4	308.2	422.5	374.1	614.2	580.3	284.6	373.2	160.8	146.4	181.4	217.5
	σ	35.6	65.9	108.5	137.5	232.5	196.2	124.0	226.6	117.2	60.3	148.7	71.3
1998–2005	μ	258.3	352.7	441.9	514.5	617.8	566.3	558.7	659.7	331.2	175.3	248.1	267.3
	σ	66.6	111.3	115.1	148.5	239.8	221.8	236.3	168.0	87.2	88.9	111.7	85.6

[a]Monthly mean precipitation values and standard deviations are shown in mm/month. For National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) data, see *Kalnay et al.* [1996].

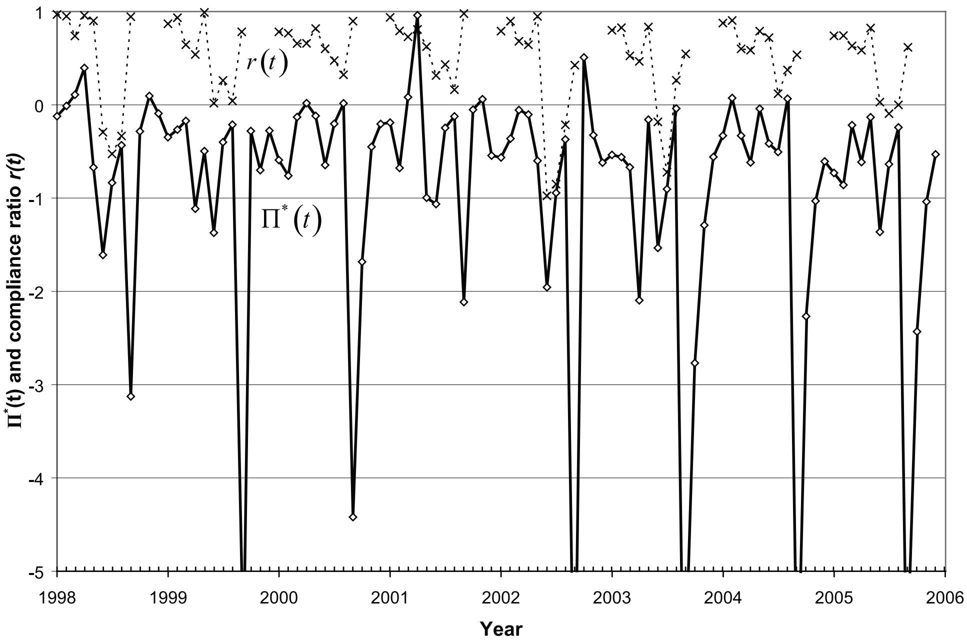

Figure 8. Development of Π* (t) during Period 3. Furthermore, the compliance ratio $r(t)$ is shown (see section 4.2 for the definition of $c(t)$). Consistently, $r(t)$ is lower in the summer months (compare also with Figure 6). For Π* (t), a generally low performance with a declining trend over time is observed. Note that the y scale is truncated at Π* (t) = −5. Values not shown are as follows: Π* (09/1999) = −7.2, Π* (09/2002) = −13.1, Π* (09/2003) = −23.1, Π* (09/2004) = −24.3, and Π* (09/2005) = −16.5.

releases (see also Table 5). Consequently, an average compliance ratio of 0.51 during Period 3 is calculated. This value together with Figure 8 shows that such compliance assessment produces much more positive scores than the more sophisticated estimation of performance.

[63] The latter finding exposes an important analytical problem associated with the compliance-oriented assessment approach. As noted by *Downs et al.* [1996], generally good compliance in international regulatory regimes can be misleading because of endogeneity and selection problems. They note that states often define treaty commitments that require little or no effort beyond what the states concerned would do in the absence of the respective treaty (see also section 2). In our case study, this general statement is confirmed by the presence of consistent positive and negative deviations of monthly flows relative to $s(t)$ (see Table 5). The empirical application of our measurement concept, which uses $o(t)$ and $c(t)$ rather than the 1998 treaty targets $s(t)$ as benchmarks, demonstrates that good news about compliance is in the Syr Darya case is certainly not good news about cooperation.

5. Conclusion

[64] The methodology proposed in this paper addresses several deficiencies in existing concepts for estimating the performance of international regulatory regimes. Notably, it deals in a transparent and tractable way with measuring the substance or depth of cooperation. Assessing substantively the performance of international water management is important from an academic and a practical viewpoint. Developing and testing generalizable explanations for success and failure in international water management must rely on an accurate measurement of the dependent variable, that is, success or failure. Also, helping policy makers understand whether or not a given water management system performs well, that is, developing accurate diagnostic tools, is usually the first step toward improving policies and institutions.

[65] At this point it should be emphasized that our method deals with observable outcomes of international treaties. It does not directly address the issue of whether the goals of these treaties are based on the principles of optimal/efficient allocation or not which is clearly a different question. Ideally, at each allocation scale from the district up to the international level, allocation should be guided by optimality/efficiency considerations as encoded in the level of optimal spatiotemporal allocation $o(t)$. This is even truer in settings such as the Syr Darya where upstream downstream conflicts exist between states as well as between human and ecosystem water uses. The dichotomy is reconciled in cases where a treaty explicitly asks for efficient allocation across as well as within borders and not just specifies transborder flows based solely on needs. In such a case, our measurement of performance would automatically measure the extent to which allocative efficiency is observed. However, again, this depends on the nature of the treaty.

Table 4. Estimated Sample Means $\hat{\mu}$ and Variances $\hat{\sigma}^2$

	δ_a	δ_c	Unit
$\hat{\mu}$	157.7	85.2	m^3/s
$\hat{\sigma}^2$	5900.2	3755.7	m^6/s^2

Table 5. Compliance Analysis[a]

Month	1998	1999	2000	2001	2002	2003	2004	2005
1	−15.7	−74.0	−138.0	−33.3	−131.0	−124.0	−70.0	−174.3
2	23.3	−35.0	−147.7	−128.7	−56.7	−102.0	42.7	−170.0
3	−106.0	−165.3	−156.7	−111.3	−140.7	−271.3	−199.3	−175.3
4	−10.0	−196.7	−119.0	−54.3	−127.3	−264.0	−162.7	−162.3
5	24.3	−2.3	−60.0	73.7	13.3	−53.7	−71.7	−57.7
6	282.0	247.7	143.0	203.3	332.0	271.0	109.7	246.3
7	392.7	276.3	223.7	235.3	422.0	411.3	304.3	339.7
8	343.0	293.7	242.3	273.7	329.0	254.7	231.0	300.0
9	10.0	34.0	17.7	4.0	69.3	−157.0	−164.7	−118.3
10	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-

[a]The monthly differences $\Delta(t)$ between treaty $s(t)$ and actual releases $a(t)$ are shown. Negative values indicate too much flow relative to the treaty, and positive values indicate too little flow relative to the treaty. Consistently, Kyrgyzstan releases too much water from January to April and too little from May to August relative to the treaty standard $s(t)$ (see also Figure 8). Note that $\Delta(t)$ is not defined for September to December, since the treaty does not specify any releases during these months (see Table 2).

[66] Our method assumes that actual performance, optimal performance, and counterfactual performance are time-dependent variables that relate to the presence or absence of particular international regimes. Data from integrated computational water resources modeling can be used to make an informed assessment of unobservable variables. To the extent time series data of reasonable quality for policy outcomes is available, our methodology can be applied to virtually any international (and also national or local) policy or regulatory regime to study its performance. If real-world or model data is more limited, the methodology also lends itself to snapshot assessments of performance at particular points in time, and to assessments based on ordinal-scaled data, for example, data obtained through expert interviews based on the Delphi method or other approaches.

[67] To demonstrate the empirical relevance of the methodology, we carried out a performance assessment of the international regulatory regime for the Naryn/Syr Darya river basin (with a focus on the Toktogul reservoir in Kyrgyzstan). The results show that this regime is characterized by low average performance and high variability.

[68] A comparison of these results with results from a conventional compliance assessment revealed that the more sophisticated method produced much more negative performance estimates. Thus it highlights the analytical problems (selection effects and endogeneity) associated with compliance- or policy-output-oriented performance estimations. Recent work by *Brochmann and Gleditsch* [2006], for example, finds no negative correlation between asymmetric upstream-downstream settings and cooperation in the form of international river management treaties. This result is surprising because it suggests that upstream-downstream asymmetries can be overcome through compensation payments and issue linkages offered by downstream countries in exchange for concessions by upstream countries at reasonably low transactions costs. Yet, the Syr Darya case, a clear upstream-downstream case where we observe a treaty and good compliance with its obligations, shows that conclusions of this kind are probably too optimistic. Our skepticism derives from the argument that international cooperation, as measured by the existence of treaties and compliance with international commitments, may often be more shallow than it looks at first sight.

[69] Our empirical findings also have important policy implications. They suggest that the riparian countries of the Syr Darya (and stakeholders from outside the river basin) should not be misled by good compliance. Our results show that the institutional solution to the problem, as put in place in 1998, is performing very poorly. Conflicts over water allocation among the riparian countries have in the past few years been muted by increased levels of precipitation upstream. As soon as an extended period of low precipitation sets in (owing to climate change or for other reasons) the conflict is likely to heat up. The obvious recommendation is to repair the regime before this happens.

[70] **Acknowledgments.** We are grateful to Andrey Yakovlev, Head of the Department of Operational Hydrology of the Uzbek Hydrometeorological Service, for providing hydrological data that would have been impossible to obtain otherwise. We also thank the Global Runoff Data Center at the Federal Institute of Hydrology (BfG) in Koblenz, Germany, for providing the pre-1991 flow data. Finally, we are grateful to Christoph Schär, Jean Fried, Detlef Sprinz, Gabi Ruoff, Vally Koubi, Thomas Hinkebein, Nils Petter Gleditsch, and Peter Bauer-Gottwein for valuable comments.

References

Antipova, E., et al. (2002), Optimization of Syr Darya water and energy uses, *Water Int.*, *27*(4), 504–516.

Bennear, L. S., and C. Coglianese (2005), Measuring progress: Program evaluation of environmental policies, *Environment*, *47*(2), 22–39.

Bernauer, T. (1996), Reducing pollution of the River Rhine: The influence of international cooperation, *J. Environ. Dev.*, *5*(4), 391–417.

Bernauer, T. (2002), Explaining success and failure in international river management, *Aquat. Sci.*, *64*(1), 1–19.

Binmore, K. (1994), *Game Theory and the Social Contract*, vol. 1, MIT Press, Cambridge, Mass.

Brochmann, M., and N. P. Gleditsch (2006), Conflict, cooperation, and good governance in international river basins, paper presented at Governance and the Global Water System: Institutions, Actors, Scales of Water Governance. Facing the Challenges of Global Change, Global Water Syst. Proj., Bonn, Germany, 20–23 June.

Cai, X. M., D. C. McKinney, and L. S. Lasdon (2002), A framework for sustainability analysis in water resources management and application to the Syr Darya Basin, *Water Resour. Res.*, *38*(6), 1085, doi:10.1029/2001WR000214.

Cai, X. M., et al. (2003), Integrated hydrologic-agronomic-economic model for river basin management, *J. Water Resour. Plann. Manage.*, *129*(1), 4–17.

Chayes, A., and A. Handler Chayes (1993), On compliance, *Int. Organ*, *47*(2), 175–205.
Dinar, A. (2004), Exploring transboundary water conflict and cooperation, *Water Resour. Res.*, *40*, W05S01, doi:10.1029/2003WR002598.
Downs, G. W., et al. (1996), Is the good news about compliance good news for cooperation?, *Int. Organ.*, *50*(3), 379–406.
Dukhovny, V., and V. Sokolov (2005), Integrated water resources management, experience and lessons learned form Central Asia: Towards the fourth world water forum, report, 27 pp., Inter-State Comm. for Water Coord. in the Aral Sea Basin, Tashkent, Uzbekistan.
Greenstone, M. (2004), Did the Clean Air Act cause the remarkable decline in sulfur dioxide concentrations?, *J. Environ. Econ. Manage.*, *47*, 585–611.
Harsanyi, J. C., and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*, vol. 13, 378 pp., MIT Press, Cambridge, Mass.
Hashimoto, T., et al. (1982), Reliability, resiliency, and vulnerability criteria for water-resource system performance evaluation, *Water Resour. Res.*, *18*(1), 14–20.
Helm, C., and D. Sprinz (2000), Measuring the effectiveness of international environmental regimes, *J. Conflict Resolut.*, *45*(5), 630–652.
Hovi, J., et al. (2003a), Regime effectiveness and the Oslo-Potsdam solution: A rejoinder to Oran Young, *Global Environ. Polit.*, *3*(3), 105–107.
Hovi, J., et al. (2003b), The Oslo-Potsdam solution to measuring regime effectiveness: Critique, response, and the road ahead, *Global Environ. Polit.*, *3*(3), 74–96.
Jenkins, G. M., et al. (1994), *Time Series Analysis Forecasting and Control*, vol. 16, 3rd ed., 598 pp., Prentice-Hall, Upper Saddle River, N. J.
Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437–471.
Kjeldsen, T. R., and D. Rosbjerg (2004), Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems, *Hydrol. Sci. J.*, *49*(5), 755–767.
Loucks, D. P., et al. (1981), *Water Resource Systems Planning and Analysis*, Prentice-Hall, Upper Saddle River, N. J.
Marty, F. (2001), *Managing International Rivers Problems, Politics and Institutions*, 409 pp., P. Lang, Bern.
McKinney, D., and X. M. Cai (1997), Multiobjective water resource allocation model for Toktogul Reservoir, report, 31 pp., Cent. for Res. in Water Resour., Univ. of Tex., Austin.
McKinney, D., et al. (1999), Integrated water resources management model for the Syr Darya Basin, report, U.S. Agency for Int. Dev., Almaty, Kazhakhstan.
Nurmi, P. (2003), Recommendations on the verification of local weather forecasts, consultancy report, Finn. Meteorol. Inst., Helsinki.
Pearl, J. (2001), *Causality Models, Reasoning and Inference*, 384 pp., Cambridge Univ. Press, New York.
Przeworski, A. (2004), Institutions matter?, *Gov. Opposition*, *39*(4), 527–540.
Rieckermann, J., et al. (2007), Assessing the performance of international water management at Lake Titicaca, *Aquat. Sci.*, *68*(4), 502–516.
Saleth, R. M. (2004), Introduction to special section on river basin management: Economics, management, and policy, *Water Resour. Res.*, *40*, W08S01, doi:10.1029/2004WR003368.
Savoskul, O., et al. (2003), Water, climate, food, and environment in the Syr Darya Basin, in *Adaptation Strategies to Changing Environments*, report, 62 pp., Inst. for Environ. Stud., Vrije, Univ., Amsterdam.
Siegfried, T., and W. Kinzelbach (2006), A multiobjective discrete stochastic optimization approach to shared aquifer management: Methodology and application, *Water Resour. Res.*, *42*, W02402, doi:10.1029/2005WR004321.
Sprinz, D. (2005), Regime effectiveness: The next wave of research, paper presented at Conference on the Human Dimensions of Global Environmental Change, Volkswagen Found., Berlin.
Sprinz, D., and C. Helm (2000), Measuring the effectiveness of international environmental regimes, *J. Conflict Resolut.*, *44*(5), 630–652.
Underdal, A. (1992), The concept of regime "effectiveness", *Cooperation Conflict*, *27*(3), 227–240.
Van Damme, E., and J. W. Weibull (1995), Equilibrium in strategic interaction: The contributions of John C. Harsanyi, John F. Nash and Reinhard Selten, *Scand. J. Econ.*, *97*(1), 15–40.
Wolf, A. T. (1999), The transboundary freshwater dispute database project, *Water Int.*, *24*(2), 160–163.
Wolf, A. T., et al. (2005), Managing conflict and cooperation, in *State of the World 2005: Redefining Global Security*, chap. 5, World Watch Inst., Washington, D. C.
World Bank (1996), Developing a regional water management strategy: Issues and work plan, *Aral Sea Basin Program Tech. Pap. Ser. 1*, Washington, D. C.
Wu, X., and D. Whittington (2006), Incentive compatibility and conflict resolution in international river basins: A case study of the Nile Basin, *Water Resour. Res.*, *42*, W02417, doi:10.1029/2005WR004238.
Yoffe, S., G. Fiske, M. Giordano, M. Giordano, K. Larson, K. Stahl, and A. T. Wolf (2004), Geography of international water conflict and cooperation: Data sets and applications, *Water Resour. Res.*, *40*, W05S04, doi:10.1029/2003WR002530.
Young, O. (2001), Inferences and indices: Evaluating the effectiveness of international environmental regimes, *Global Environ. Polit.*, *1*(1), 99–121.
Young, O. (2003), Determining regime effectiveness: A commentary on the Oslo-Potsdam solution, *Global Environ. Polit.*, *3*(3), 97–104.

T. Bernauer, Center for Comparative and International Studies, ETH Zurich, Seilergraben 49, SEI–G.9, 8092 Zurich, Switzerland. (bernauer@ir.gess.ethz.ch)

T. Siegfried, Earth Institute, Columbia University, 500 W. 120th St., Room 842A, New York, NY 10027, USA. (ts2392@columbia.edu)