

# GLOBAL ESTIMATION AND SCENARIO-BASED PROJECTIONS OF SEX RATIO AT BIRTH AND MISSING FEMALE BIRTHS USING A BAYESIAN HIERARCHICAL TIME SERIES MIXTURE MODEL

BY FENGQING CHAO<sup>1</sup>, PATRICK GERLAND<sup>2</sup>, ALEX R. COOK<sup>3</sup> AND LEONTINE ALKEMA<sup>4</sup>

<sup>1</sup>*Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, [fengqing.chao@kaust.edu.sa](mailto:fengqing.chao@kaust.edu.sa)*

<sup>2</sup>*United Nations Population Division, DESA, United Nations, [gerland@un.org](mailto:gerland@un.org)*

<sup>3</sup>*Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, [ephcar@nus.edu.sg](mailto:ephcar@nus.edu.sg)*

<sup>4</sup>*School of Public Health and Health Sciences, University of Massachusetts, Amherst, [lalkema@umass.edu](mailto:lalkema@umass.edu)*

The sex ratio at birth (SRB) is defined as the ratio of male to female live births. The SRB imbalance in parts of the world over the past several decades is a direct consequence of sex-selective abortion, driven by the coexistence of son preference, readily available technology of prenatal sex determination and fertility decline. Estimation and projection of the degree of SRB imbalance is complicated because of variability in SRB reference levels and because of the uncertainty associated with SRB observations.

We develop Bayesian hierarchical time series mixture models for SRB estimation and scenario-based projections for all countries from 1950 to 2100. We model the SRB regional and national reference levels and the fluctuation around national reference levels. We identify countries at risk of SRB imbalances and model both: (i) the absence or presence of sex ratio transitions in such countries and, if present, (ii) the transition process. The transition model of SRB imbalance captures three stages (increase, stagnation and convergence back to SRB baselines). The model identifies countries with statistical evidence of SRB inflation in a fully Bayesian approach. The scenario-based SRB projections are based on the sex ratio transition model with varying assumptions regarding the occurrence of a sex ratio transition in at-risk countries. Projections are used to quantify the future burden of missing female births due to sex-selective abortions under different scenarios.

**1. Introduction.** Under normal circumstances the sex ratio at birth (SRB, defined as the ratio of male to female live births) falls within a narrow range of 1.03 to 1.07 and varies slightly by ethnicity [Chahnazarian (1988), Chao et al. (2019a), Dubuc and Coleman (2007), Garenne (2002, 2008), Graffelman and Hoekstra (2000), James (1984, 1985, 1987), Kaba (2008), Marcus et al. (1998), Mathews and Hamilton (2005), Ruder (1985), Visaria (1967)]. For most of human history, SRB remained within that natural range. However, over recent decades SRBs have risen in a number of Asian countries and in Eastern Europe [Attané and Guilmoto (2007), Basten and Verropoulou (2013), Bongaarts (2013), Bongaarts and Guilmoto (2015), Chao and Yadav (2019), Chao et al. (2019a), Chen et al. (2020), Choi and Hwang (2020), Das Gupta et al. (2003), Duthé et al. (2012), Goodkind (2011), Guilmoto, Hoàng and Van (2009), Guilmoto (2009, 2012a, 2012b, 2012c), Guilmoto and Ren (2011), Hudson and Den Boer (2004), Lin (2009), Meslé, Vallin and Badurashvili (2007), Park and Cho (1995), Tafuro and Guilmoto (2020), Vu and Yamada (2020)]. SRB imbalance results from the interaction of three main factors [Guilmoto (2009, 2012a)]: first, prolonged strong

---

Received June 2020; revised November 2020.

*Key words and phrases.* Bayesian hierarchical model, probabilistic scenario-based projection, time series analysis, sex-selective abortion, sex ratio transition, missing female births.

son preference, offering the motivation; second, fertility decreases leading to fewer children per family, inducing the willingness; third, accessible affordable sex-selection technology, providing the means. As a result, couples seek abortion based on the knowledge of the sex of the pregnancy to obtain sons while maintaining a small family size.

Estimation of the degree of SRB imbalance is challenging because of: (i) variation in baseline SRB levels and (ii) uncertainty associated with SRB observations. In prior work, [Chao et al. \(2019a\)](#), we developed a model to estimate SRB and imbalances for 212 countries from 1950 to 2017. The SRB estimation model accounts for the difference in the SRB reference levels across regions and varying uncertainty associated with SRB observations. We identified 29 countries where SRB imbalance may have happened in the past or may happen in the future, which we refer to as countries at risk of SRB inflation. We fitted a model for SRB levels and trends in country-years without risk of sex-selective abortion and obtained estimates for national and regional SRB baseline values. Subsequently, we estimated sex imbalances using a sex ratio transition model to capture periods of increasing, constant and decreasing sex ratio imbalances.

Constructing SRB projections is challenging for countries with ongoing SRB imbalances [[Gupta, Chung and Shuzhuo \(2009\)](#)] and even more so for countries with normal SRB levels and trends but with potential of rising SRB in the future [[Bongaarts and Guilmoto \(2015\)](#)]. While prior studies have assessed the potential future SRB imbalances for selected countries with ongoing transitions [[Bongaarts and Guilmoto \(2015\)](#), [Chao et al. \(2020\)](#)], no work, to date, has quantified the possible additional SRB imbalances for countries where sex ratio transitions may start in the future. Efforts to date to project the SRB globally include those by the United Nations (U.N.) Population Division [[United Nations, DESA \(2019\)](#)]. The U.N. Population Division publishes projections of demographic indicators, including the SRB for all countries in the World Population Prospects (WPP) every two years. While the U.N. WPP projections of fertility, mortality and populations are probabilistic, SRB projections are deterministic and based on expert-based opinions [[United Nations, DESA \(2019\)](#)]. Specifically, WPP methods are based on the assumption that future SRB outcomes either remain at the same level, as most recently observed, or converge to 1.05 within the next 10–40 years, and future SRB imbalance is not assessed. Similarly, the Global Burden of Disease produces SRB estimates and projections for all countries but does not assess SRB imbalances and assumes that the SRB remains constant in projections from 2017 up to 2100 [[Vollset et al. \(2020\)](#)].

In this study we extend the SRB estimation model [[Chao et al. \(2019a\)](#)] to produce scenario-based SRB projections, and associated SRB imbalances, till 2100 for all countries. The update in the SRB inflation model allows for identification of countries with past and ongoing SRB inflation in a fully Bayesian approach. We use the updated model to estimate the sex ratio transition for country-years with risk of SRB inflation. Subsequently, we construct scenario-based SRB projections that are based on the sex ratio transition model with varying assumptions regarding the occurrence of a sex ratio transition in at-risk countries.

This paper is organized as follows: we first summarize the data used for estimating the SRB in Section 2. Section 3 introduces the models for estimating the SRB, followed by the approaches used to produce scenario-based SRB projections. In Section 4 we present results regarding SRB baselines, past SRB imbalances and scenario-based projections as well as validation results. Finally, in Section 5 we end with a discussion of implications of our model and results, limitations and possible future research.

**2. Data.** We produce SRB estimates for 212 countries<sup>1</sup> with total population size greater than 90,000 as of 2017. An overview of observations by data source type is in Table 1. There

---

<sup>1</sup>We use the term “country” to refer to populations that are considered as countries or areas in the U.N. classification.

TABLE 1

*SRB observations by source type. DHS: Demographic and Health Surveys. Other DHS: refers to nonstandard DHS, including special, interim and national DHS, Malaria Indicator Surveys, AIDS Indicator Surveys, World Fertility Surveys, Reproductive Health Survey, Multiple Indicator Cluster Surveys, Pan Arab Project for Child Development and Pan Arab Project for Family Health. CRVS: Civil Registration Vital Statistics. SRS: Sampling Registration System*

Data source type	# Observations	# Country-years
Census	48	48
DHS	2257	5413
Other DHS	1392	3662
Other	142	222
CRVS/SRS	6996	7257
Total	10,835	16,602

are 10,835 data points available from 202 countries in total. The compiled SRB database is publicly available at [Chao et al. \(2019b\)](#).

Data availability is summarized by source type in Table 1. We compile Civil Registration Vital Statistics (CRVS) data from the U.N. Demographic Yearbook and the Human Mortality Database, and Sampling Registration System (SRS) data for India, Pakistan and Bangladesh from annual reports. CRVS and SRS typically provide data on an annual basis, based on government administrative record from birth certificates. International survey data (Demographic and Health Surveys, World Fertility Surveys, Reproductive Health Survey, Multiple Indicator Cluster Surveys, Pan Arab Project for Child Development and Pan Arab Project for Family Health) were compiled from microdata, when possible, and obtained from reports otherwise. Census and national-level survey data were obtained from reports. Censuses usually provide information on SRB for a retrospective period of 12 or 24 months prior to the survey date. Surveys collect data on recent births or full birth histories from women of reproductive ages for longer retrospective periods of five to 20 years before the survey date. Details on data series by country and preprocessing are given in [Chao et al. \(2019a, 2019c\)](#).

### 3. Methods.

*Notation.* We use lowercase Greek letters for unknown parameters and uppercase Greek letters for variables which are functions of unknown parameters. Where relevant for communicating model assumptions, we add hyperparameters associated with parameters in parentheses. For instance, for some model parameter  $\phi$ , the notation  $\phi(\zeta)$  implies that the distribution for  $\phi$  is parametrized using hyperparameter  $\zeta$ . We use super script  $(g)$  to the left of a parameter and a model/scenario name to the right to denote posterior samples of that parameter from a specific model fit/scenario; that is,  $^{(g)}\phi^{[M1]}$  refers to the  $g$ th posterior sample of  $\phi$ , as obtained from model fit M1.  $\hat{\phi}^{[M1]}$  refers to the posterior median of  $\phi$  based on model M1. Roman letters indicate variables that are known or fixed, including data (in lowercase) and estimates provided by other sources or the literature (in uppercase).

$\Theta_{c,t}$  denotes the main outcome of interest to be modeled which is the SRB for country  $c$  in year  $t$ . Observations are combined across countries over time and indexed by  $i \in \{1, \dots, n\}$ ;  $c[i]$  refers to the country to which the  $i$ th observation belongs,  $t[i]$  to the calendar year of the observation and  $s[i]$  to the data source type (Table 1) of the observation.  $r[c]$  refers to the index of the region of which country  $c$  belongs.

3.1. *SRB model summary.* We assume the  $i$ th observed SRB  $y_i$  follows a normal distribution on the log-scale

$$(1) \quad \log(y_i) | \Theta_{c[i],t[i]}, \omega_{s[i]} \sim \mathcal{N}(\log(\Theta_{c[i],t[i]}), \omega_{s[i]}^2 + v_i^2).$$

The variance for the  $i$ th log-scaled SRB observation  $\log(y_i)$  is the sum of known stochastic/sampling variance  $v_i^2$  and unknown nonsampling error variance  $\omega_{s[i]}^2$  for data source type  $s[i]$  (Section 3.2).

In its general form the process model for SRB  $\Theta_{c,t}$  is defined as follows:

$$(2) \quad \Theta_{c,t} = \beta_c \eta_{c,t}(\boldsymbol{\phi}) + \delta_c \Omega_{c,t}(\boldsymbol{\zeta}),$$

where  $\beta_c$  refers to country-specific time-invariant baseline SRB and  $\eta_{c,t}$  the country-year-specific fluctuations around the baseline.  $\delta_c$  indicates the absence or presence of SRB inflations in country  $c$ . For countries with inflations,  $\Omega_{c,t}$  captures the SRB imbalance in year  $t$  and can be interpreted as the additional number of male births for each female birth that is not aborted. The vector of hyperparameters related to  $\eta_{c,t}$  is denoted as  $\boldsymbol{\phi}$ , and we use  $\boldsymbol{\zeta}$  for the vector of hyperparameters related to  $\Omega_{c,t}$ .

We identify a set of countries where SRB inflation may have happened in the past or may happen in the future, that is, countries with  $\Pr(\delta_c > 0) \neq 0$ , which we refer to as “countries at risk of SRB inflation” (see Section 3.3). In our study we focus on sex ratio transitions that are due to sex-selective abortion, not other factors (e.g., natural disaster, economic crises, famine and war) that could result in acute changes in SRB [Catalano et al. (2006), Fukuda et al. (1998), Song (2012), Venero Fernández et al. (2011)]. The estimation of baseline SRB  $\beta_c$  and fluctuations around baseline  $\eta_{c,t}$  is described in Section 3.4. The sex ratio transition model  $\delta_c \Omega_{c,t}$  is given in Section 3.5. We introduce the approach for constructing scenario-based projections in Section 3.6, using the models developed in the preceding sections. The full model specification, priors and details on computation are in Appendix A.

**3.2. Error variances.** We account for differences in error variance across observations from CRVS/SRS, surveys and censuses. Errors—and hence the error variances—associated with non-CRVS/SRS data tend to be larger than errors associated with CRVS/SRS data, and this is reflected in the model fitting, as the weight assigned to a data point increases as its error variance decreases. Resulting model-based estimates are more strongly weighted by observations with smaller errors, and uncertainty ranges are narrower for country-periods with more observations with smaller error variances.

As per equation (1), the variance for the  $i$ th log-scaled observed SRB  $\log(y_i)$  is the sum of known stochastic/sampling variance  $v_i^2$  and unknown nonsampling error variance  $\omega_{s[i]}^2$  for data source type  $s[i]$ , as listed in Table 1. For CRVS/SRS observations,  $v_i$  is the stochastic error and is precalculated, as described elsewhere [Chao et al. (2019a, 2019c)], and we assume that nonsampling error is zero:  $\omega_s = 0$  when  $s = \text{CRVS/SRS}$ . For observations from surveys or censuses,  $v_i$  is the sampling error and is precalculated using a jackknife method, as explained in Chao et al. (2019a, 2019c), to reflect the survey sampling design. Nonsampling error variance term  $\omega_s^2$  captures random errors that may occur during the data collection process. This variance parameter is estimated and assigned a vague prior,

$$(3) \quad \omega_s \sim \mathcal{U}(0, 0.5) \quad \text{for } s \in \{\text{Census, DHS, Other DHS, Other}\}.$$

**3.3. Selection of countries at risk of SRB inflation.** The model for natural fluctuations in the SRB is fitted to the global database [Chao et al. (2019b)] after excluding data from country-years that may have been affected by masculinisation of the SRB. We use inclusive criteria to identify such country-years, based on a combination of qualitative and quantitative approaches. We select countries with at least one of the following manifestations of son preference: (i) a high level of desired sex ratio at birth (DSRB), (ii) a high level of sex ratio at last birth (SRLB) or (iii) strong son preference or inflated SRB suggested by a literature review [Chao et al. (2019a, 2019c)]. DSRB is calculated as the ratio of the reported number of

desired male births to desired female births, as reported by women during survey interviews. The DSRB reflects the desired sex composition at the time of survey interview. The SRLB quantifies the sex ratio among births that are the latest births to women who desire no more children.

A total of 29 countries satisfy at least one of the aforementioned criteria and hence are considered at risk of SRB inflation (see [Chao et al. \(2019a, 2019c\)](#)): Afghanistan, Albania, Armenia, Azerbaijan, Bangladesh, China, Egypt, Gambia, Georgia, Hong Kong-SAR of China, India, Jordan, Republic of Korea, Mali, Mauritania, Montenegro, Morocco, Nepal, Nigeria, Pakistan, Senegal, Singapore, Taiwan-Province of China, Tajikistan, Tanzania, Tunisia, Turkey, Uganda and Vietnam.

3.4. *Model stage 1: Estimating the SRB in country-years without SRB inflation.* In the model for country-years not affected by sex-selective abortion, we assume  $\delta_c = 0$  in equation (2). The SRB is thus given by a product of two components,

$$(4) \quad \Theta_{c,t} = \beta_c \eta_{c,t}(\boldsymbol{\phi}),$$

where  $\beta_c$  is a national baseline for country  $c$ , which is assumed to be constant over time,  $\eta_{c,t}(\boldsymbol{\phi})$  is a country-year-specific multiplier that captures the natural fluctuation of the country-specific SRB around its respective baseline value over time and  $\boldsymbol{\phi}$  is the vector of the hyper-parameters related to  $\eta_{c,t}$ . Above we label equation (4) as ‘‘M1’’ to refer to model stage 1. In this step of modeling, parameters that are not related to prenatal gender discrimination and sex-selective abortion are estimated. To do so, we use a reduced SRB database, referred to as the risk-free database denoted by  $\mathbf{y}^{(\text{risk-free})}$ , by excluding SRB observations that may be affected by prenatal sex discrimination and sex-selective abortion. The risk-free database is obtained by excluding data points with reference years after 1970 from the 29 countries at risk of SRB inflation (listed in Section 3.3) because the sex-selective abortion technology became more accessible and affordable since then [[Allahbadia \(2002\)](#), [George \(2002\)](#), [Goodkind \(1997\)](#), [Oomman and Ganatra \(2002\)](#), [Tandon and Sharma \(2006\)](#)]. Data points with reference years before 1970 from the 29 countries and all data points from remaining countries are included in the risk-free database for fitting the model.

The national baseline  $\beta_c$  follows a hierarchical normal distribution with mean at its corresponding regional baseline  $\beta_{r[c]}^{(\text{region})}$ ,

$$(5) \quad \log(\beta_c) | \beta_{r[c]}^{(\text{region})}, \sigma_\beta \sim \mathcal{N}(\log(\beta_{r[c]}^{(\text{region})}), \sigma_\beta^2).$$

National baselines are pooled toward regional baseline  $\beta_r^{(\text{region})}$  to capture SRB differences due to ethnic origin [[Chahnazarian \(1988\)](#), [Dubuc and Coleman \(2007\)](#), [Garenne \(2002, 2008\)](#), [Graffelman and Hoekstra \(2000\)](#), [James \(1984, 1985, 1987\)](#), [Kaba \(2008\)](#), [Marcus et al. \(1998\)](#), [Mathews and Hamilton \(2005\)](#), [Ruder \(1985\)](#), [Visaria \(1967\)](#)]. For example, we group countries in Europe, North America, Australia and New Zealand to refer to the regional grouping of countries with a majority of Caucasians. We assume that the national baseline  $\beta_c$  and regional baseline  $\beta_r^{(\text{region})}$  are constant over time. We assign independent uniform priors to each  $\beta_r^{(\text{region})}$  and a vague prior to  $\sigma_\beta$ .

The country-year-specific multiplier  $\eta_{c,t}$  is modeled on the log-scale with an autoregressive AR(1) time series process within a country. For countries without any data or with very limited information,  $\eta_{c,t}$  is shrunk toward 1, such that the estimated SRBs without prenatal sex discrimination are close to their corresponding national baselines  $\beta_c$ . For countries where the data suggest different levels or trends,  $\eta_{c,t}$  captures these deviations from  $\beta_c$ . Let

$\boldsymbol{\phi} = \{\rho, \sigma_\epsilon\}$  be the vector of hyperparameters related to  $\eta_{c,t}$ . We assume

$$(6) \quad \log(\eta_{c,t})|\boldsymbol{\phi} \sim \mathcal{N}(0, \sigma_\epsilon^2/(1 - \rho^2)) \quad \text{for } t = 1950,$$

$$(7) \quad \log(\eta_{c,t}) = \rho \log(\eta_{c,t-1}) + \epsilon_{c,t} \quad \text{for } t \in \{1951, \dots, 2100\},$$

$$(8) \quad \epsilon_{c,t}|\sigma_\epsilon \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2).$$

Vague priors are assigned to  $\sigma_\beta$ ,  $\rho$  and  $\sigma_\epsilon$ .

*3.5. Model stage 2: Estimating the SRB in country-years at risk of SRB inflation.* We model SRB  $\Theta_{c,t}$  for the 29 countries at risk of SRB inflation (listed in Section 3.3) as the sum of two parts: (i) the inflation-free SRB level, given by the model of country-years without SRB inflation, as described in Section 3.4, and (ii) the SRB imbalanced level with probability. We fit the model to all data from the 29 at-risk countries and let  $\mathbf{y}^{(\text{at-risk})}$  denote the database in model stage 2. Specifically,  $\Theta_{c,t}$  for country  $c$ , year  $t$  is modeled as

$$(9) \quad \Theta_{c,t} = \hat{\beta}_c^{[\text{M1}]} \eta_{c,t} (\hat{\boldsymbol{\phi}}^{[\text{M1}]}) + \delta_c \Omega_{c,t}(\boldsymbol{\zeta}),$$

where  $\hat{\beta}_c^{[\text{M1}]}$  is the posterior median estimate for the national baseline and  $\hat{\boldsymbol{\phi}}^{[\text{M1}]} = \{\hat{\rho}^{[\text{M1}]}, \hat{\sigma}_\epsilon^{[\text{M1}]}\}$  the vector of posterior medians of  $\boldsymbol{\phi}$  obtained from the inflation-free model stage 1 (M1) fit described in Section 3.4. Above, we label equation (9) as ‘‘M2’’ to refer to model stage 2. In M2 we impute point estimate  $\hat{\beta}_c^{[\text{M1}]}$  from M1 rather than estimating  $\beta_c$  and inflation  $\Omega_{c,t}$  jointly to avoid upward bias in the estimated start year of SRB inflation in China and India due to overestimation of  $\beta_c$  (see Appendix C).

The product  $\delta_c \Omega_{c,t}(\boldsymbol{\zeta})$  captures SRB imbalance. It is the product of a binary indicator representing the presence or absence of SRB inflation  $\delta_c$  for country  $c$  and a nonnegative SRB inflation  $\Omega_{c,t}(\boldsymbol{\zeta})$  where  $\boldsymbol{\zeta}$  is the vector of hyperparameters related to  $\Omega_{c,t}$ .

The country-specific binary factor  $\delta_c$  detects the existence of SRB inflation, with values either 0 (no inflation) or 1 (with inflation).  $\delta_c$  is modeled with a Bernoulli distribution with country-specific probability of having inflation  $\pi_c$ ,

$$(10) \quad \delta_c|\pi_c \sim \mathcal{B}(\pi_c).$$

Logit-transformed  $\pi_c$  follows a hierarchical normal distribution with a global mean at  $\mu_\pi$  and a global variance at  $\sigma_\pi^2$ ,

$$(11) \quad \text{logit}(\pi_c)|\mu_\pi, \sigma_\pi \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2).$$

Vague priors are assigned to  $\sigma_\pi$  and  $\mu_\pi$ .

$\Omega_{c,t}$  is the upward SRB inflation factor for country  $c$  in year  $t$  to capture higher SRB levels that may be due to sex-selective abortion. We parameterize the sex ratio transition using a trapezoid function to represent consecutive phases of increase, stagnation and decrease back to zero (Figure 1). The inflation factor  $\Omega_{c,t}$  is modeled as

$$(12) \quad \Omega_{c,t} = \begin{cases} \xi_c(t - \gamma_{0,c})/\lambda_{1,c}, & \gamma_{0,c} < t < \gamma_{1,c}, \\ \xi_c, & \gamma_{1,c} < t < \gamma_{2,c}, \\ \xi_c - \xi_c(t - \gamma_{2,c})/\lambda_{3,c}, & \gamma_{2,c} < t < \gamma_{3,c}, \\ 0, & t < \gamma_{0,c} \text{ or } t > \gamma_{3,c}, \end{cases} \quad \text{where}$$

$$\gamma_{1,c} = \gamma_{0,c} + \lambda_{1,c},$$

$$\gamma_{2,c} = \gamma_{1,c} + \lambda_{2,c},$$

$$\gamma_{3,c} = \gamma_{2,c} + \lambda_{3,c},$$

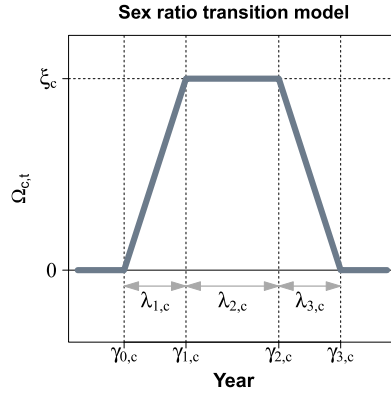


FIG. 1. Illustration of sex ratio transition model. All parameters are for country  $c$ .  $\gamma_{0,c}$ : Start year of SRB inflation.  $\lambda_{1,c}$ ,  $\lambda_{2,c}$  and  $\lambda_{3,c}$ : Period lengths of the SRB inflation phases of increase, stagnation and decrease back to zero.  $\xi_c$ : Maximum value of the inflation.

with sex ratio transition parameters  $\gamma_{0,c}$ , the start year of the inflation,  $\xi_c$ , the maximum inflation and  $\lambda_{1,c}$ ,  $\lambda_{2,c}$  and  $\lambda_{3,c}$ , the lengths of the three phases that constitute the inflation period.

The model for  $\Omega_{c,t}$  makes use of fertility as an external covariate related to SRB inflation to better capture the sex ratio transition process. Specifically, information related to the “fertility squeeze” effect (fertility decreases leading to fewer children per family, inducing the willingness to sex-selective abortion) is incorporated into the model through the parametrization of the start year  $\gamma_{0,c}$  (see Section 3.5.1). Despite the acknowledged role of son preference (offering the motivation of sex-selective abortion) or availability of sex-selection technology (providing the means of sex selection) in driving SRB imbalance [Guilmoto (2009, 2012a)], the model for  $\Omega_{c,t}$  does not make use of covariates related to these two factors because detailed information is generally unavailable for estimation and available candidate predictors do not explain variability across countries [Chao et al. (2019a)]. Instead, sex ratio transition parameters related to the lengths of the transition phases and its maximum are estimated with Bayesian hierarchical models [Gelman et al. (2014), Lindley and Smith (1972)],

$$(13) \quad \xi_c | \mu_\xi, \sigma_\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2) T(0, \infty),$$

$$(14) \quad \lambda_{1,c} | \mu_{\lambda_1}, \sigma_{\lambda_1} \sim \mathcal{N}(\mu_{\lambda_1}, \sigma_{\lambda_1}^2) T(0, \infty),$$

$$(15) \quad \lambda_{2,c} | \mu_{\lambda_2}, \sigma_{\lambda_2} \sim \mathcal{N}(\mu_{\lambda_2}, \sigma_{\lambda_2}^2) T(0, \infty),$$

$$(16) \quad \lambda_{3,c} | \mu_{\lambda_3}, \sigma_{\lambda_3} \sim \mathcal{N}(\mu_{\lambda_3}, \sigma_{\lambda_3}^2) T(0, \infty),$$

where  $\mathcal{N}(\cdot) T(0, \infty)$  refers to a truncated normal distribution with lower truncation at zero. The hierarchical models allow for projecting complete transitions in all countries, including countries that have not yet finished their transitions, based on (partially) observed transitions so far. We assign vague priors to the mean and standard deviation of these truncated distributions (Appendix A).

3.5.1. *Inflation start year.* We model the SRB inflation start year  $\gamma_{0,c}$  as

$$(17) \quad \gamma_{0,c} | \sigma_\gamma \sim t_3(x_c, \sigma_\gamma^2) T(z_c, \infty),$$

referring to a Student- $t_3$  distribution with lower truncation in year  $z_c$ , location  $x_c$  and scale  $\sigma_\gamma$ . The fertility squeeze effect is incorporated into the model for  $\gamma_{0,c}$  using country-year estimates of the total fertility rate  $F_{c,t}$  from the U.N. WPP 2019 [United Nations, DESA

(2019)]. The total fertility rate (TFR) approximates the number of children that would be born to a woman who survives throughout her reproductive ages. The lower truncation  $z_c$  refers to the year that the TFR in country  $c$  decreased to six or the year 1970, whichever occurred later,

$$(18) \quad z_c = \max\{1970, f_{c,6}\},$$

where  $f_{c,6}$  is the first year in which the TFR in country  $c$  declines to six. The location indicator  $x_c$  is the year in which the TFR in country  $c$  declines to the precalculated value of 2.9 (the TFR value 2.9 is determined by the median of fertility levels when the observed sex ratio transition started among countries with high-quality CRVS data [Chao et al. (2019a, 2019c)]) or 1970, whichever occurred later,

$$(19) \quad x_c = \max\{1970, f_{c,2.9}\}.$$

*3.6. Scenario-based projections.* We construct scenario-based SRB projections that are based on the sex ratio transition model in equation (2) with varying assumptions regarding the occurrence of a sex ratio transition in at-risk countries. The three scenarios are:

- S1: SRB inflation continues in countries with strong statistical evidence of past inflations only.
- S2: SRB inflations are projected to occur in all at-risk countries with country-specific inflation probabilities.
- S3: SRB inflations are projected to occur in all at-risk countries with probability 1.

The scenarios are labeled in terms of increasing sex imbalance in future years, with S1 including the sex imbalance due to ongoing transitions in countries with strong statistical evidence of such inflations only and S3 including sex imbalance for at-risk countries without current evidence of inflation. For each country-year  $c, t$ , we let  ${}^{(g)}\Theta_{c,t}^{[S]}$  denote the  $g$ th posterior sample of  $\Theta_{c,t}$  for scenario  $S \in \{S1, S2, S3\}$ .

We identify countries with strong statistical evidence of SRB inflation in a fully Bayesian approach, using  $\psi_c$ , the posterior mean of the inflation indicator  $\delta_c$  for country  $c$  from M2:  $\psi_c = \mathbb{E}(\delta_c | \mathbf{y}^{(\text{at-risk})})$ .  $\psi_c$  represents the relative inclusion of sex ratio inflation in country  $c$ . If  $\psi_c \geq 95\%$ , then we consider the country to have strong statistical evidence of SRB inflation. We use  $\mathcal{C}^{\text{inflation}}$  to denote the set of country indices  $c$  with  $\psi_c \geq 95\%$ .

The projections are based on estimates from different model fits, summarized in Table 2. For countries that are not at risk of SRB inflation, with  $c \in \mathcal{C}^{\text{base}}$ , projections are based on the country-specific baseline estimates and deviations away from that baseline and are the same across scenarios, obtained from M1. Specifically,  ${}^{(g)}\Theta_{c,t}^{[S]}$ , the  $g$ th posterior sample of  $\Theta_{c,t}$  for scenario  $S \in \{S1, S2, S3\}$  is obtained as below:

$$(20) \quad {}^{(g)}\Theta_{c,t}^{[S]} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\eta_{c,t}^{[M1]} \quad \text{for scenario } S \in \{S1, S2, S3\}, c \in \mathcal{C}^{\text{base}}.$$

As indicated in the notation section (see Section 3), we introduce the superscript  $(g)$  on the left-hand side and  $[M]$  on the right-hand side of a parameter to refer to posterior sample  $g$  of that parameter from model  $M \in \{M1, M2, M3, M4\}$ .

For  $c \in \mathcal{C}^{\text{inflation}}$ , the projections under scenarios S1, S2 and S3 are also identical and based on posterior estimates of the model parameters from M2, combined with additional uncertainty related to the country baseline  $\beta_c$ , as obtained from M1,

$$(21) \quad {}^{(g)}\Theta_{c,t}^{[S]} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\eta_{c,t}^{[M2]} + {}^{(g)}\delta_c^{[M2]} {}^{(g)}\Omega_{c,t}^{[M2]}$$

for scenario  $S \in \{S1, S2, S3\}, c \in \mathcal{C}^{\text{inflation}}$ .



TABLE 2  
 Summary of models.  $\mathbf{y}^{(c)}$ : All data points from country  $c$ .  $\hat{\xi}^{[M2]}$  refers to point estimates of the vector of hyperparameters for the inflation factor from M2 ( $\xi = \{\mu_\xi, \sigma_\xi, \mu_{\lambda_1}, \sigma_{\lambda_1}, \mu_{\lambda_2}, \sigma_{\lambda_2}, \mu_{\lambda_3}, \sigma_{\lambda_3}, \sigma_\gamma\}$ )

Model	Model equation	Database
M1	$\Theta_{c,t} = \beta_c \eta_{c,t}(\phi)$	$\mathbf{y}^{(\text{risk-free})}$
M2	$\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]}) + \delta_c \Omega_{c,t}(\xi)$	$\mathbf{y}^{(\text{at-risk})}$
M3	$\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]})$	$\mathbf{y}^{(c)}$
M4	$\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]}) + \Omega_{c,t}(\hat{\xi}^{[M2]})$	$\mathbf{y}^{(c)}$

The construction of the scenario-based projections for the remaining countries, those at risk of inflation but without strong evidence of past or ongoing inflations, denoted by set  $\mathcal{C}^{\text{future-inf}}$ , differs between scenarios. In S2 the country-specific probability of having inflation, as obtained in M2, is used for constructing S2 projected trajectories  ${}^{(g)}\Theta_{c,t}^{[S2]}$ ,

$$(22) \quad {}^{(g)}\Theta_{c,t}^{[S2]} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\eta_{c,t}^{[M2]} + {}^{(g)}\delta_c^{[M2]} {}^{(g)}\Omega_{c,t}^{[M2]}, \quad c \in \mathcal{C}^{\text{future-inf}}.$$

For scenarios S1 and S3 for at-risk countries  $c \in \mathcal{C}^{\text{future-inf}}$ , additional model fits M3 and M4 (see Table 2) are introduced to obtain projections based on estimates without inflation (S1) and with 100% chance of inflation (S3). Projections for S1 are obtained as follows:

$$(23) \quad {}^{(g)}\Theta_{c,t}^{[S1]} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\eta_{c,t}^{[M3]}, \quad c \in \mathcal{C}^{\text{future-inf}},$$

where the model equation for M3 corresponds to that from M1—the model without inflation terms—but M3 is fitted to country-specific data  $\mathbf{y}^{(c)}$ , which includes *all* data points from country  $c$ , as opposed to data up to 1970 only. By including all country-specific data in the model fit, resulting M3 estimates for recent years are informed by the country's data. Projections for S3 are given by

$$(24) \quad {}^{(g)}\Theta_{c,t}^{[S3]} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\eta_{c,t}^{[M4]} + {}^{(g)}\Omega_{c,t}^{[M4]}, \quad c \in \mathcal{C}^{\text{future-inf}},$$

where the model equation for M4 is the model with inflation term fit to country-specific data  $\mathbf{y}^{(c)}$ . The main difference between M2 and M4 is that the inflation  $\delta_c = 1$  for all trajectories in M4, hence forcing an inflation in all trajectories as per S3 assumption.

We incorporate the uncertainty associated with the start year of the inflation, due to uncertainty in TFR projections, into the SRB projections for S2 and S3 in the projection result for the set of countries  $\mathcal{C}^{\text{future-inf}}$ . For trajectories with posterior samples of start year  $\gamma_{0,c}$  beyond the most recent SRB observation, we add in additional uncertainty of  $x_c$ , the year in which TFR in country  $c$  declines to 2.9, based on TFR trajectories. We use 1000 TFR trajectories for associated uncertainty for each country-year based on projections from the `bayesTFR` R-package [Ševčíková, Alkema and Raftery (2011), Ševčíková et al. (2019)].

**3.7. Model validation.** We assess model performance via validation exercises focused on: (i) predicting left-out recent SRB observations (Section 3.7.1) and (ii) predicting SRB inflation transitions (Section 3.7.2).

**3.7.1. Predicting left-out SRB observations.** We assess model performance by leaving out recent data. Specifically, we leave out 20% of the data points that were collected after a certain survey year [Alkema, Wong and Seah (2012), Alkema et al. (2014)] for the out-of-sample validation. After leaving out data, we fit the model to the training dataset and obtain median estimates and prediction intervals for the SRB and SRB inflation that would have been

constructed based on available dataset in the survey year selected. We also assess the model performance by leaving out data at random, that is, leaving out 20% of the data randomly, and repeating this exercise 30 times.

We calculate median errors and median absolute errors for the left-out SRB observations, where errors are defined as:  $e_j = y_j - (\hat{y}_j | \mathbf{y}^{\text{train}})$ , where  $(\hat{y}_j | \mathbf{y}^{\text{train}})$  refers to the posterior median of the predictive distribution based on the training dataset  $\mathbf{y}^{\text{train}}$  for the  $j$ th left-out observation  $y_j$ . Coverage for 95% prediction intervals for left-out observations is given by  $1/J \cdot \sum_{j=1}^J \mathbb{I}_{\mathbf{A}}(y_j, l_j) \cdot \mathbb{I}_{\mathbf{B}}(y_j, u_j)$ , where  $J$  is the total number of left-out observations,  $l_j$  and  $u_j$  correspond to the 2.5th and 97.5th percentiles of the posterior predictive distribution (PPD) for the  $j$ th left-out observation  $y_j$ , and sets  $\mathbf{A} = \{(a, b) \in \mathbb{R}^2 : a > b\}$  and  $\mathbf{B} = \{(a, b) \in \mathbb{R}^2 : a < b\}$ . For the 80% prediction interval coverage,  $l_j$  and  $u_j$  refer to the 10th and 90th percentiles of the PPD, respectively. The validation measures are calculated for 1000 permutations of left-out observations, where each permutation consists one randomly selected left-out observation from each country with data left out. The reported validation results are based on the mean of the outcomes from the 1000 permuted sets of left-out observations.

For the median estimates based on full dataset and training dataset, errors are defined as  $e(\Theta)_{c,t} = (\hat{\Theta}_{c,t} | \mathbf{y}^{\text{full}}) - (\hat{\Theta}_{c,t} | \mathbf{y}^{\text{train}})$ , where  $(\hat{\Theta}_{c,t} | \mathbf{y}^{\text{full}})$  is the posterior median for country  $c$  in year  $t$  based on the full dataset  $\mathbf{y}^{\text{full}}$ , and  $(\hat{\Theta}_{c,t} | \mathbf{y}^{\text{train}})$  is the posterior median for the same country-year based on the training dataset. Similarly, the error for the sex ratio transition process with probability is defined as  $e(\delta\Omega)_{c,t} = (\widehat{\delta_c\Omega_{c,t}} | \mathbf{y}^{\text{full}}) - (\widehat{\delta_c\Omega_{c,t}} | \mathbf{y}^{\text{train}})$ . Coverage is computed in a similar manner as for the left-out observations, based on the lower and upper bounds of the equal-tail 95% and 80% credible intervals for  $\Theta_{c,t}$  based on the training dataset.

**3.7.2. Predicting sex ratio transitions since 1970.** We assess the predictive performance of the inflation model by predicting the SRB inflation for each at-risk country, referring to the set of countries  $\mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}}$ . In this exercise, country data after 1970 is not used to directly inform the transition parameter estimates. Instead, we use median estimates of the baseline and natural fluctuations up to 1970 based on country-specific data prior to 1970 and predict the SRB from 1970 onward as follows:

$${}^{(g)}\tilde{\Theta}_{c,t} = {}^{(g)}\beta_c^{[M1]} {}^{(g)}\tilde{\eta}_{c,t} + {}^{(g)}\tilde{\delta}_c {}^{(g)}\tilde{\Omega}_{c,t} ({}^{(g)}\xi^{[M2]}),$$

where  ${}^{(g)}\tilde{\delta}_c$  and  ${}^{(g)}\tilde{\Omega}_{c,t}$  refer to samples from their respective posterior predictive distributions, where  ${}^{(g)}\tilde{\delta}_c$  is simulated based on  ${}^{(g)}\mu_{\pi}^{[M2]}$  and  ${}^{(g)}\sigma_{\pi}^{[M2]}$  and  ${}^{(g)}\tilde{\Omega}_{c,t}$  is simulated based on  ${}^{(g)}\xi^{[M2]} = \{ {}^{(g)}\mu_{\xi}^{[M2]}, {}^{(g)}\sigma_{\xi}^{[M2]}, {}^{(g)}\mu_{\lambda_1}^{[M2]}, {}^{(g)}\sigma_{\lambda_1}^{[M2]}, {}^{(g)}\mu_{\lambda_2}^{[M2]}, {}^{(g)}\sigma_{\lambda_2}^{[M2]}, {}^{(g)}\mu_{\lambda_3}^{[M2]}, {}^{(g)}\sigma_{\lambda_3}^{[M2]} \}$ . The natural fluctuations  ${}^{(g)}\tilde{\eta}_{c,t}$  are given by M2 for  $t \leq 1970$  and predicted for  $t > 1970$ ,

$${}^{(g)}\tilde{\eta}_{c,t} = {}^{(g)}\eta_{c,t}^{[M2]} \quad \text{for } t \leq 1970,$$

$$\log({}^{(g)}\tilde{\eta}_{c,t}) | {}^{(g)}\tilde{\eta}_{c,t-1} \sim \mathcal{N}({}^{(g)}\rho^{[M1]} \log({}^{(g)}\tilde{\eta}_{c,t-1}), ({}^{(g)}\sigma_{\epsilon}^{[M1]})^2) \quad \text{for } t \in \{1971, \dots, 2100\}.$$

After generating the predicted values, we calculate the same set of results, as described in Section 3.7.1.

**3.8. Estimates of missing female births.** The realization of SRB inflation due to sex-selective abortion is quantified with the number of “missing female births” which refers to the additional number of female births that would have been born if the inflation were absent. It is calculated as the difference between the number of female births under normal circumstance (referred here as the “inflation-free” number of female births) and the actual number under the observed SRB.

The estimated and expected inflation-free number of female live births for a country-year, denoted by  $\Psi_{c,t}$  and  $\Psi_{c,t}^{(\text{inflation-free})}$ , are computed as

$$\begin{aligned}\Psi_{c,t} &= B_{c,t}/(1 + \Theta_{c,t}) \quad \text{and} \\ \Psi_{c,t}^{(\text{inflation-free})} &= (B_{c,t} - \Psi_{c,t})/\Theta_{c,t}^{(\text{inflation-free})},\end{aligned}$$

where  $B_{c,t}$  is the total number of births for a certain country-year, obtained from the U.N. WPP 2019 [United Nations, DESA (2019)]. The number of inflation-free female births,  $\Psi_{c,t}^{(\text{inflation-free})}$ , is obtained from the estimated number of male births ( $B_{c,t} - \Psi_{c,t}$ ) and from the inflation-free SRB,  $\Theta_{c,t}^{(\text{inflation-free})} = \beta_c \eta_{c,t}$ , for the respective country-year. The method of calculating the missing female births follows from what was first introduced by Dréze and Sen (1993) and was recently reviewed and validated by Guilmo, Chao and Kulkarni (2020).

The annual number of missing female births (AMFB) for country  $c$  in year  $t$  is defined as

$$\Psi_{c,t}^{(\text{missing})} = \Psi_{c,t}^{(\text{inflation-free})} - \Psi_{c,t}.$$

The cumulative number of missing female births (CMFB) for period  $t_1$  to  $t_2$  in country  $c$  is defined as the sum of AMFB from the year  $t_1$  up to the year  $t_2$ ,

$$\Lambda_{c,[t_1,t_2]}^{(\text{missing})} = \sum_{t=t_1}^{t_2} \Psi_{c,t}^{(\text{missing})}.$$

**4. Results.** We first summarize findings related to SRB baselines and natural deviations for country-years without risk of inflation (Section 4.1). Second, the sex ratio transition model results are presented (Section 4.2), followed by projections (Section 4.3). Finally, validation results are provided (Section 4.4).

4.1. *SRB baselines and country-year variations in years without risk of inflation.* The median estimates of regional SRB baselines  $\beta_r^{(\text{region})}$  and national baselines  $\beta_c$  are illustrated in Figure 2. The lowest regional baseline is estimated in sub-Saharan Africa at 1.031 (95% credible interval [1.027; 1.036]) and the highest is in Oceania at 1.067 [1.058; 1.077]. Among the 10 regions the probability that baseline is smaller than 1.05 (the widely assumed historical norm) is greater than 97.5% in two regions (sub-Saharan Africa, Latin America and the Caribbean) and the probability that baseline is larger than 1.05 is bigger than 97.5% in four regions (ENAN, southeastern Asia, eastern Asia and Oceania). The national baselines are estimated to range from 1.013 [1.000; 1.026] in Zambia and 1.013 [0.997; 1.028] in Namibia to 1.081 [1.068; 1.093] in Hong Kong.

Australia is an example of a country without risk of SRB inflation. Its estimates and projections are displayed in Figure 3. It typifies countries with high quality annual CRVS data, here available from 1921 to 2015. SRB estimates follow the CRVS data trend, and the uncertainty assessment takes into account the stochastic uncertainty associated with the CRVS data. Its national baseline  $\beta_c$  is estimated at 1.055 [1.049; 1.062]. The national baseline of Australia differs from its regional baseline  $\beta_{r[c]}^{(\text{region})}$  at 1.058 [1.055; 1.061] for the region ENAN (the combination of countries in Europe, North America, Australia and New Zealand) since the national baseline is informed by the CRVS data available in Australia. The estimated SRB for Australia  $\Theta_{c,t}$  (Figure 3, top row) ranges from 1.048 [1.042; 1.055] in 1935 to 1.060 [1.054; 1.067] in 1928. As Australia is not identified to have SRB inflation risk, all the deviations of  $\Theta_{c,t}$  from the national baseline of SRB  $\beta_c$  are considered as natural fluctuation and are captured with  $\eta_{c,t}$  (Figure 3, bottom row). Given that Australia is considered as not having risk of SRB inflation, its projected SRB is based on the model of country-years without SRB inflation. The SRB projection for Australia is approximately constant and given by its national baseline  $\beta_c$ , with the projection for 2100 given by 1.055 [1.037; 1.074].

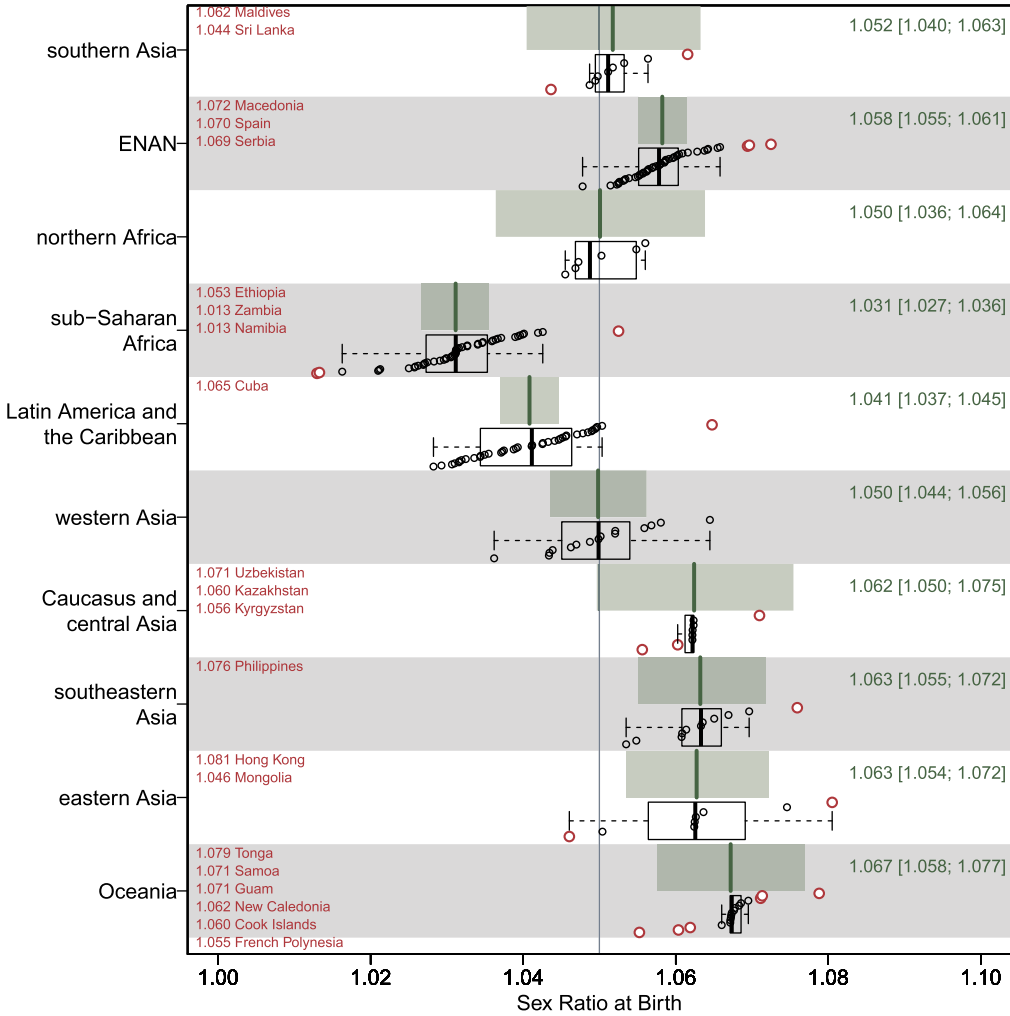


FIG. 2. SRB regional and national baseline. Regional baseline median estimates  $\beta_r^{(\text{region})}$  (dark green line), with 95% credible intervals (shaded area) and printed values (green). Box plots summarize the distributions of national baselines  $\beta_c$  within each region. National median estimates are shown (dots) with those outside the range of [25th percentile  $-1.5IQR$ ; 75th percentile  $+1.5IQR$ ] highlighted in red. The countries with outlying national baselines are listed in the legend with the median estimates reported (ordered by median estimate). ENAN: The combination of countries in Europe, North America, Australia, and New Zealand.

4.2. *Sex ratio transition model.* Table 3 summarizes model results for the sex ratio transition for the 29 countries with risk of SRB inflation. In total, 12 countries are identified as having strong statistical evidence of SRB inflation: Albania, Armenia, Azerbaijan; China, Georgia, Hong Kong-SAR of China, India, Republic of Korea, Montenegro, Taiwan-Province of China, Tunisia and Vietnam.

Figure 4 presents the estimated SRB inflation  $\Omega_{c,t}$  (rather than SRB itself) together with inflation residuals for the 12 countries with strong evidence of SRB inflation. We obtain the difference between the observed SRBs and the median estimates of country-specific baselines, referred to here as inflation residual  $r_i = y_i - \hat{\beta}_c^{[M1]}$ . The residuals indicate the presence of an inflation, and the parametric form of  $\Omega_{c,t}$  is able to capture its shape. For the 12 countries the median estimate of start year of the sex ratio transition  $\gamma_{0,c}$  is estimated to be the year in which TFR declines to 5.2 in India to 1.0 in Hong Kong-SAR of China. The country-specific maximum value of SRB imbalance  $\xi_c$  is estimated to be higher than 0.100 in

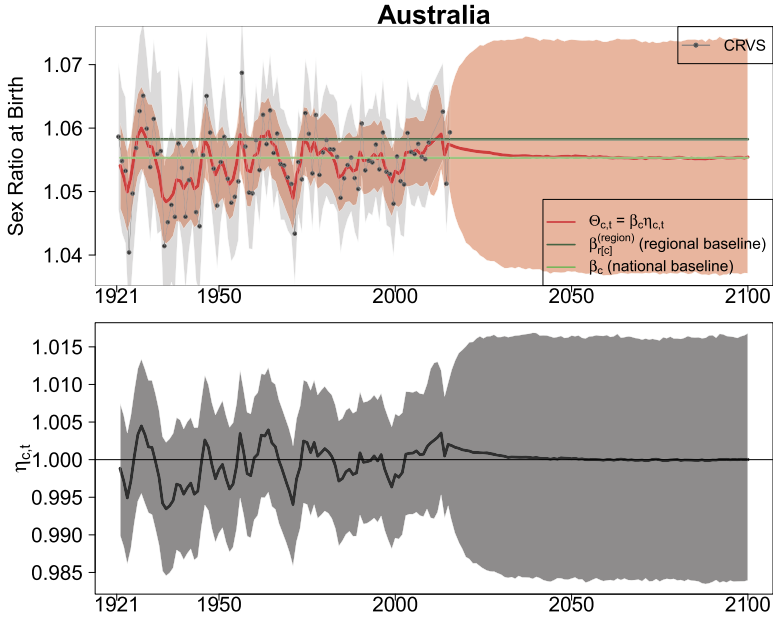


FIG. 3. *SRB estimates and projections for Australia. Top row: SRB median estimates  $\Theta_{c,t}$  (red curve) and 95% credible intervals (red shades), median estimates of the regional baselines  $\beta_{r[c]}^{(region)}$  (dark green horizontal line), median estimates of the national baselines  $\beta_c$  (light green horizontal line). SRB observations are displayed with dots and observations are connected with lines when obtained from the same source. Shaded areas around observation series represent the sampling variability in the series (quantified by two times the stochastic/sampling standard errors). Bottom row: Median estimates of natural deviation  $\eta_{c,t}$  (solid line) and 95% credible intervals (shades).*

China at 0.112 [0.078; 0.151], in Armenia at 0.108 [0.088; 0.129] and in Azerbaijan at 0.104 [0.087; 0.121]. The maximum imbalance is lower than 0.050 in Montenegro at 0.045 [0.022; 0.068], in Tunisia at 0.035 [0.019; 0.050] and in Taiwan-Province of China at 0.030 [0.016; 0.043]. The estimated duration of the sex ratio transition, referring to the sum of the three phases  $\lambda_{1,c} + \lambda_{2,c} + \lambda_{3,c}$ , varies greatly across the 12 countries and ranges from relatively fast transitions that last 11, 25 and 26 years for Hong Kong-SAR of China, Georgia and the Republic of Korea, respectively, to 45, 51 and 58 years in Montenegro, China and India. For the 12 countries the average projection period from the most recent observation year to end of transition is 8.8 years, corresponding to an average of 20.1% of country-specific duration of the sex ratio inflation. All these countries have data available up to the second phase of the sex ratio transition, the stagnation. At the country level, transitions are estimated to have finished prior to the most recent observation year in Hong Kong-SAR of China and in Republic of Korea. Four countries (Georgia, Montenegro, Taiwan-Province of China and Tunisia) have observations available up to transition phase 3 and the remaining six countries (Albania, Armenia, Azerbaijan, China, India and Vietnam) have data coverage up to phase 2.

Figure 5 presents the projected SRB inflation  $\Omega_{c,t}$  for a country prior to observing SRB data. Given the hierarchical structure of the SRB inflation model, the sex ratio transition in Figure 5 represents the average experience of SRB imbalance. The median projected SRB inflation process ( $\lambda_{1,c} + \lambda_{2,c} + \lambda_{3,c}$ ) has a span of 37 [15; 64] years. The maximum SRB inflation  $\xi_c$  for a new country has a median at 0.032 with 95% credible interval [0.000; 0.132]. The inflation maximum is reached around 11 [1; 28] years after the country's TFR declines to 2.9 in the year  $x_c$ .

TABLE 3

*Sex ratio transition model results for the 29 countries at risk of SRB inflation. Numbers above the brackets are posterior median estimates. Numbers inside the brackets are 95% credible intervals.  $\psi_c$ : the relative inclusion of SRB inflation.  $\gamma_{0,c}$ : the start year of SRB inflation.  $\lambda_{1,c}$ ,  $\lambda_{2,c}$ ,  $\lambda_{3,c}$ : period lengths for increase, stagnation and decrease for the sex ratio transition process.  $\gamma_{3,c} = \gamma_{0,c} + \lambda_{1,c} + \lambda_{2,c} + \lambda_{3,c}$ : the end year of SRB inflation.  $\xi_c$ : maximum SRB inflation. 2100+: indicates the year is beyond 2100. Countries with strong statistical evidence of SRB inflation are in boldface. Countries are in descending order of the  $\psi_c$  values*

Country	$\psi_c$ (in %)	TFR in $\gamma_{0,c}$	$\gamma_{0,c}$ (start year)	$\gamma_{3,c}$ (end year)	$\xi_c$	$\delta_c \xi_c$	$\lambda_{1,c}$	$\lambda_{2,c}$	$\lambda_{3,c}$
<b>Albania</b>	100	3.1	1988 [1973; 1997]	2024 [2016; 2043]	0.059 [0.038; 0.079]	0.059 [0.038; 0.079]	15 [3; 31]	3 [0; 26]	15 [6; 33]
<b>Armenia</b>	100	2.5	1992 [1990; 1993]	2029 [2020; 2042]	0.109 [0.088; 0.13]	0.109 [0.088; 0.130]	7 [5; 9]	5 [0; 19]	26 [7; 39]
<b>Azerbaijan</b>	100	3.0	1991 [1988; 1993]	2031 [2019; 2049]	0.104 [0.087; 0.122]	0.104 [0.087; 0.122]	10 [7; 15]	9 [2; 23]	19 [3; 39]
<b>China</b>	100	2.6	1980 [1972; 1988]	2030 [2017; 2051]	0.114 [0.08; 0.156]	0.114 [0.080; 0.156]	20 [8; 32]	11 [1; 28]	18 [3; 39]
<b>Georgia</b>	100	2.1	1992 [1979; 1994]	2016 [2008; 2027]	0.054 [0.039; 0.07]	0.054 [0.039; 0.070]	3 [0; 20]	9 [1; 13]	12 [0; 27]
<b>Hong Kong- SAR of China</b>	100	1.0	2004 [2002; 2005]	2013 [2012; 2014]	0.076 [0.059; 0.096]	0.076 [0.059; 0.096]	6 [4; 9]	2 [0; 4]	2 [0; 3]
<b>India</b>	100	5.2	1975 [1970; 1981]	2033 [2021; 2050]	0.056 [0.042; 0.072]	0.056 [0.042; 0.072]	19 [7; 29]	16 [2; 34]	24 [3; 44]
<b>Republic of Korea</b>	100	2.4	1982 [1978; 1984]	2006 [1997; 2011]	0.072 [0.058; 0.087]	0.072 [0.058; 0.087]	8 [6; 12]	4 [2; 5]	13 [2; 18]
<b>Tunisia</b>	100	4.9	1982 [1976; 1989]	2021 [2012; 2039]	0.036 [0.021; 0.052]	0.036 [0.021; 0.052]	13 [3; 25]	10 [1; 21]	15 [2; 36]
<b>Vietnam</b>	100	2.0	2001 [1991; 2005]	2036 [2017; 2061]	0.066 [0.035; 0.131]	0.066 [0.035; 0.131]	10 [1; 26]	8 [0; 25]	16 [2; 37]
<b>Montenegro</b>	100	2.3	1980 [1971; 1990]	2024 [2014; 2043]	0.048 [0.026; 0.076]	0.048 [0.026; 0.076]	14 [2; 29]	10 [1; 26]	19 [3; 39]
<b>Taiwan- Province of China</b>	99.7	2.4	1982 [1972; 1987]	2023 [2012; 2041]	0.031 [0.018; 0.045]	0.031 [0.017; 0.044]	10 [3; 27]	12 [1; 22]	20 [5; 39]
Mauritania	63.8	2.9	2065 [2038; 2093]	2100+ [2068; 2100+]	0.064 [0.005; 0.163]	0.034 [0.000; 0.149]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Mali	63.4	2.9	2061 [2036; 2089]	2099 [2066; 2100+]	0.064 [0.005; 0.16]	0.033 [0.000; 0.146]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Afghanistan	63.3	2.9	2033 [2013; 2063]	2071 [2041; 2100+]	0.063 [0.005; 0.16]	0.033 [0.000; 0.146]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Nigeria	63.2	2.9	2065 [2038; 2093]	2100+ [2069; 2100+]	0.064 [0.005; 0.163]	0.033 [0.000; 0.148]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Gambia	63.1	2.9	2053 [2027; 2082]	2091 [2057; 2100+]	0.064 [0.005; 0.161]	0.033 [0.000; 0.149]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Pakistan	63.0	2.9	2030 [1995; 2058]	2068 [2027; 2100+]	0.062 [0.005; 0.16]	0.030 [0.000; 0.146]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Senegal	63.0	2.9	2061 [2034; 2089]	2099 [2064; 2100+]	0.064 [0.005; 0.161]	0.033 [0.000; 0.148]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Tanzania	62.7	2.9	2068 [2041; 2096]	2100+ [2071; 2100+]	0.064 [0.005; 0.162]	0.033 [0.000; 0.146]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Uganda	62.6	2.9	2042 [2020; 2070]	2080 [2049; 2100+]	0.064 [0.005; 0.162]	0.032 [0.000; 0.147]	11 [1; 28]	8 [0; 25]	16 [2; 37]
Nepal	62.6	2.6	2009 [1989; 2036]	2047 [2018; 2083]	0.058 [0.005; 0.155]	0.026 [0.000; 0.135]	12 [1; 28]	8 [0; 25]	16 [1; 37]
Tajikistan	62.2	2.9	2038 [2016; 2067]	2076 [2044; 2100+]	0.063 [0.005; 0.162]	0.032 [0.000; 0.147]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Egypt	61.9	2.9	2030 [2006; 2058]	2068 [2034; 2100+]	0.063 [0.005; 0.161]	0.030 [0.000; 0.147]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Jordan	56.3	2.7	2019 [2000; 2048]	2057 [2028; 2095]	0.061 [0.004; 0.158]	0.017 [0.000; 0.140]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Singapore	44.9	2.3	1975 [1970; 2012]	2014 [1990; 2055]	0.029 [0.003; 0.145]	0.000 [0.000; 0.031]	10 [1; 27]	8 [0; 25]	17 [2; 37]
Morocco	40.8	2.6	2003 [1982; 2033]	2041 [2010; 2080]	0.061 [0.003; 0.156]	0.000 [0.000; 0.127]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Bangladesh	37.4	2.4	2009 [1987; 2038]	2046 [2015; 2085]	0.061 [0.003; 0.162]	0.000 [0.000; 0.130]	11 [1; 28]	8 [0; 25]	16 [1; 37]
Turkey	34.9	2.9	1993 [1974; 2028]	2031 [2001; 2074]	0.046 [0.003; 0.151]	0.000 [0.000; 0.078]	11 [1; 28]	7 [0; 25]	16 [2; 37]

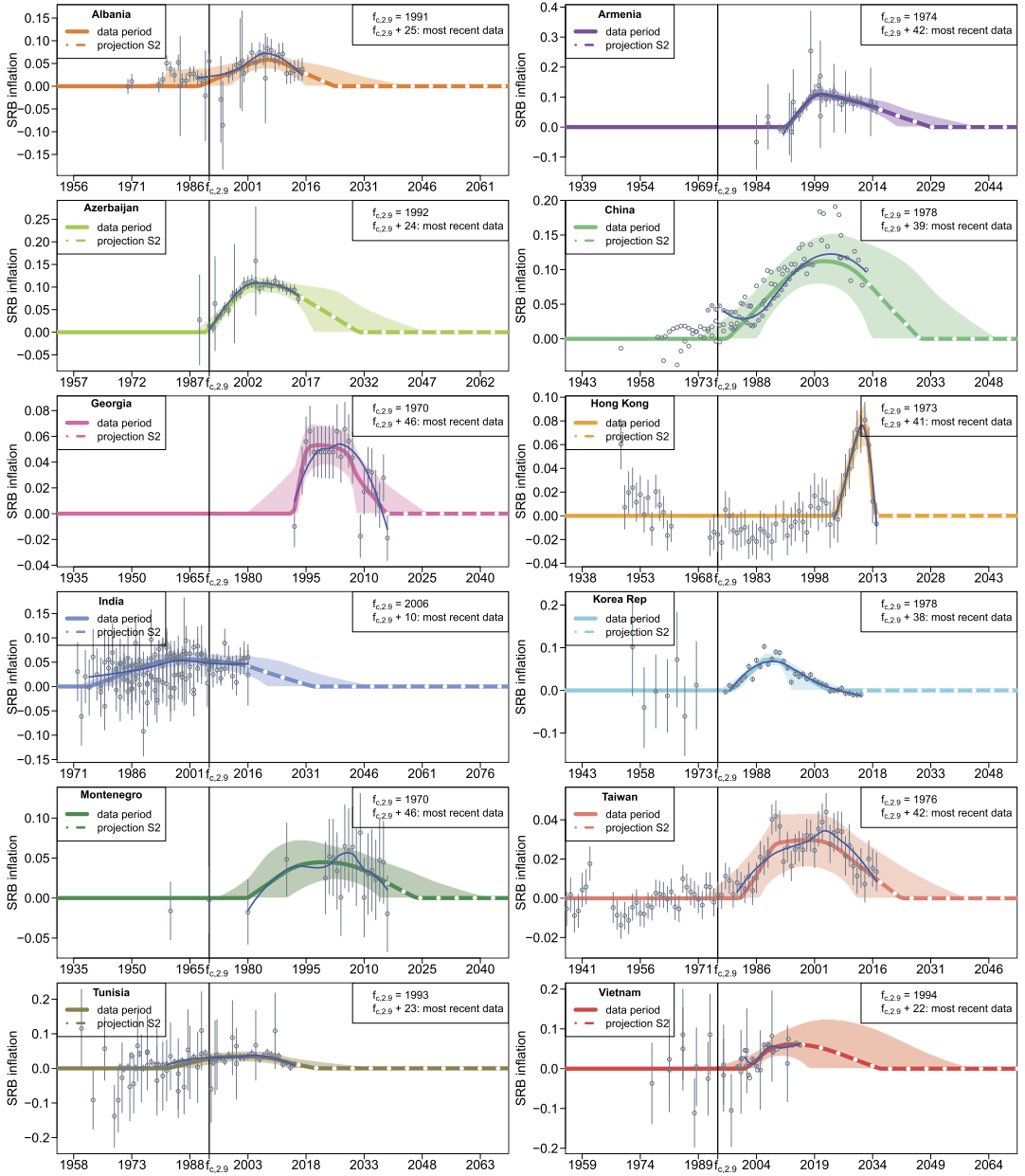


FIG. 4. SRB inflation estimates and projections for countries with strong statistical evidence of SRB inflation with inflation residuals. SRB inflation median estimates during periods with data (solid lines), median projections (dashed lines) and 95% credible intervals (shades). Data shown are the difference between the observed SRB and the median estimates of country-specific baseline ( $y_i - \hat{\beta}_c^{[M1]}$ ). Vertical line segments around data represent the sampling variability in the data (quantified by two times the stochastic/sampling standard errors). Loess curves are in blue. Loess curves are based on data after the inflation start year median estimates and data points are weighted by the inverse of sampling variance. The year in which the TFR declines to 2.9 in each country is shown as well (vertical line labeled as  $f_{c,2.9}$  on the x-axis).

4.3. SRB estimates and projections for countries at risk of SRB inflation. SRB and resulting missing female births projections by scenario for all countries at risk of SRB inflation are given in Figure A of the Supplementary Material [Chao et al. (2021a)]. Scenario 1 SRB estimates and projections during 1950–2100 for all countries are given in Figure B of the

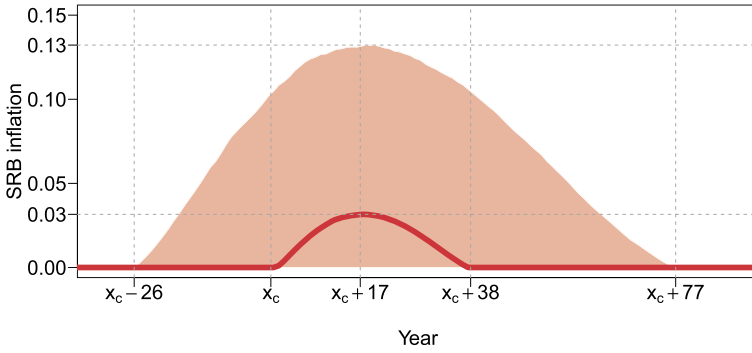


FIG. 5. SRB inflation projection for a country with future SRB inflation/prior to observing data. Median projection (solid line) with 95% credible intervals (shades).  $x_c$  refers to the year in which the total fertility rate (TFR) in a country declines to 2.9 or 1970, whichever occurs later.

Supplementary Material [Chao et al. (2021b)]. Here, we illustrate projections for selected countries.

*Countries with strong statistical evidence of SRB inflation: China.* China is identified with strong statistical evidence of SRB inflation, as listed in Table 3. Its SRB inflation is still ongoing at the start of the projection period as shown in Figure 6. The start year of the inflation is estimated at 1980 [1972; 1988] when the TFR declined to 2.6. We estimate that the SRB in China peaked in 2005 at 1.179 [1.141; 1.220], with associated SRB inflation of 0.112 [0.079; 0.150]. In 2017, the SRB in China is estimated at 1.144 [1.079; 1.206]. We project that the SRB will converge back to the range of natural fluctuations around its national baseline value of 1.063 [1.044; 1.082] in 2030 [2017; 2051].

The annual number of missing female births (AMFB) in China peaked in 2007 at 0.8 [0.6; 1.1] million female births per year. With the inflation projected to decrease to zero by 2030, the resulting AMFB is also projected to decrease to zero at that time. The cumulative number of missing female births (CMFB) for China since 1970 is projected to be 27.9 [18.6; 41.4] million by 2100.

*Countries at risk but without strong evidence of SRB inflation: Afghanistan and Senegal.* Seventeen countries are classified with risk of SRB inflation without strong statistical evidence of SRB inflation, as listed in Table 3. The SRB in these countries are projected under three scenarios based on different assumptions on the occurrence of a sex ratio transition. We use Afghanistan and Senegal as examples to illustrate the scenario-based projections of SRB for such country (Figure 7 and Figure 8).

During the observation period for Afghanistan (Figure 7), data series for Afghanistan do not imply SRB inflation, and, hence, all the fluctuations and deviations of  $\Theta_{c,t}$  away from national baseline  $\beta_c$  are captured with  $\eta_{c,t}$ . For Afghanistan,  $\eta_{c,t}$  is close to one throughout the estimation period.

The three SRB projection scenarios result in substantively different future SRB and associated missing births in Afghanistan. The projection under S1 without future inflation,  $\Theta_{c,t}^{[S1]}$ , remains at its national baseline  $\beta_c$ . The S2 projection  $\Theta_{c,t}^{[S2]}$  includes a sex ratio transition in  $\psi_c = 63\%$  of all future trajectories, while S3 projection  $\Theta_{c,t}^{[S3]}$  includes transitions for all trajectories. For future trajectories with sex imbalances in Afghanistan under scenarios S2 and S3, the transition is projected to start in the 2030s ( $\gamma_{0,c}$  is projected in the year 2033 [2013; 2063]) and ends in 2071 [2041; beyond 2100]. Given that we incorporate the uncertainty in the TFR for S2 and S3 projections  $\Theta_{c,t}^{[S2]}$  and  $\Theta_{c,t}^{[S3]}$ , the SRB inflation process is more flattened than the one shown in Figure 5. Under scenario S2 the average annual number



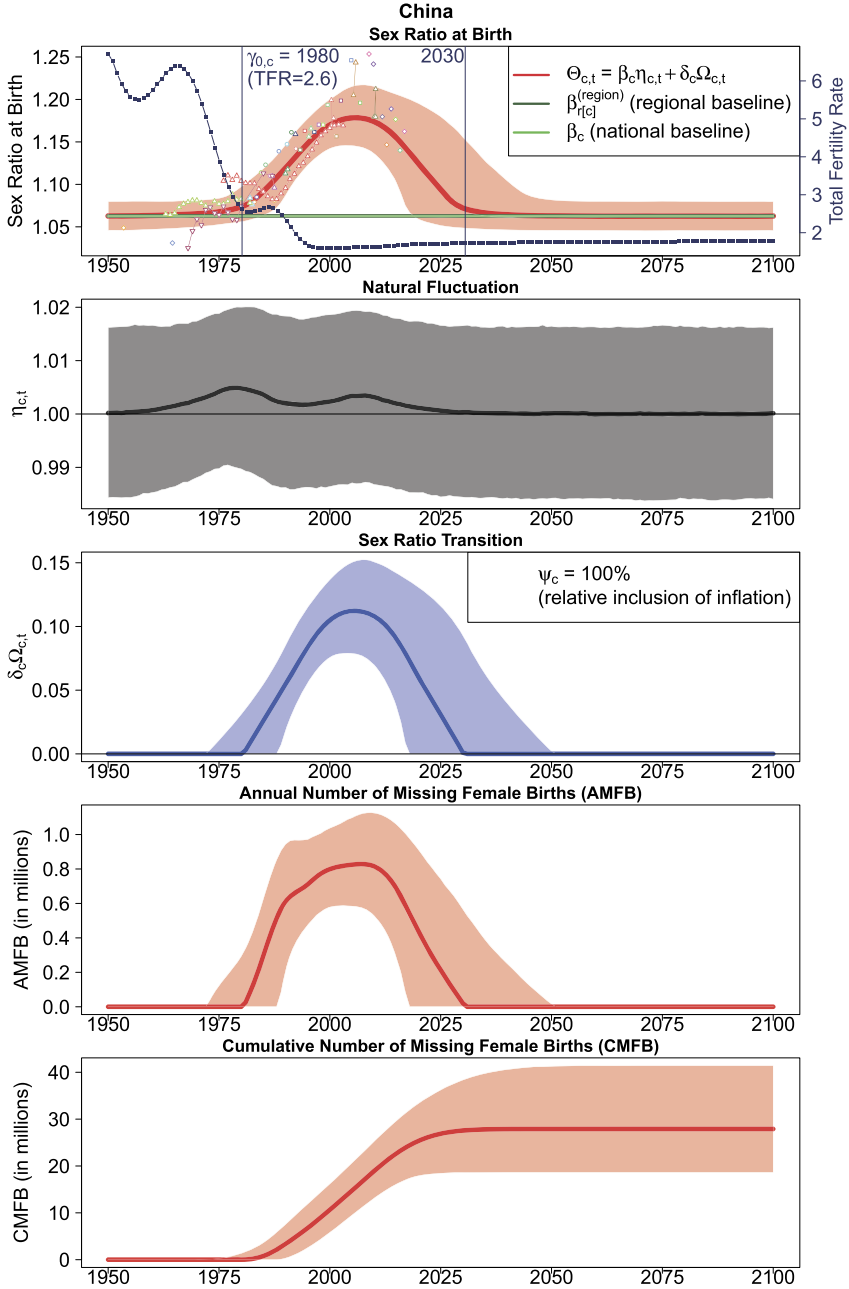


FIG. 6. SRB and missing female births estimates and projections for China. Row 1: SRB median estimates  $\Theta_{c,t}$  (red curve) and 95% credible intervals (red shades), median estimates of the regional baselines  $\beta_{r[c]}^{(region)}$  (dark green horizontal line), median estimates of the national baselines  $\beta_c$  (light green horizontal line). SRB observations are displayed with dots and observations are connected with lines when obtained from the same source. Shaded areas around observation series represent the sampling variability in the series (quantified by two times the stochastic/sampling standard errors). The median estimates and projections of total fertility rate (TFR) from the U.N. WPP 2019 are added (blue squared dots). The median estimates of inflation start year  $\gamma_{0,c}$  and end year  $\gamma_{3,c}$  are the vertical lines. The TFR value in the year  $\gamma_{0,c}$  is shown. Row 2: Median estimates of natural deviation  $\eta_{c,t}$  (solid line) and 95% credible intervals (shades). Row 3: Median estimates of SRB inflation  $\delta_c \Omega_{c,t}$  (curves) and 95% credible intervals (shades). Row 4: Annual number of missing female births (AMFB) estimates and projections. Row 5: Cumulative number of missing female births (CMFB) estimates and projections.

TABLE 4

Validation results for left-out observations. Error is defined as the difference between a left-out observation and the posterior median of its predictive distribution. M1: Model without inflation term (Section 3.4). M2: Model with inflation term (Section 3.5). Prediction from 1970: Predicting sex ratio transition based on prediction setup (Section 3.7.2)

Left-out observations $y_j$	M1		M2	M2
	Recent obs.	Random	Recent obs.	Prediction from 1970
# Country in training dataset	176	184	29	29
# Country in test dataset	143	169	28	29
Median error	0.000	-0.001	-0.003	0.010
Median absolute error	0.015	0.012	0.020	0.027
Below 95% prediction interval (%)	2.7	3.0	4.6	1.4
Above 95% prediction interval (%)	3.6	2.3	1.7	3.6
Expected (%)	2.5	2.5	2.5	2.5
Below 80% prediction interval (%)	9.7	9.8	11.3	7.0
Above 80% prediction interval (%)	10.2	8.2	8.6	14.0
Expected (%)	10	10	10	10

of missing female births (AMFB) during 2018–2100 is projected to be 4 [0; 22] thousand. The corresponding cumulative number of missing female births (CMFB) during 2018–2100 is projected to be 303 [0; 1787] thousand. Under scenario S3 the average AMFB is projected at 8 [0; 22] thousand during 2018–2100 and end up with 624 [36; 1848] thousand missing female births cumulatively.

For Senegal (Figure 8) the model detects no SRB inflation during the data period and projects the SRB imbalance to start in 2061 [2034; 2089] when the TFR declines to 2.9 and to end in 2099 [2064; beyond 2100]. The timing of the inflation is projected to be later than the sex ratio transition in Afghanistan due to differences in TFR projections, with Senegal’s TFR projections suggesting a slower decline than expected in Afghanistan. For Senegal under scenario S2, the average AMFB from 2018 to 2100 is projected to be 1 [0; 16] thousand and corresponding CMFB is projected to be 104 [0; 1298] thousand. The average AMFB and resulting CMFB for scenario S3 are 5 [0; 17] thousand and 399 [0; 1422] thousand.

4.4. *Validation results.* Table 4 summarizes the results related to the left-out SRB observations for: (i) M1 (model without inflation as described in Section 3.4) by leaving out observations obtained from the year 2005 onward, consisting 20.3% of all observations with no risk of SRB inflation  $y^{(\text{risk-free})}$ , (ii) M1 by randomly leaving out 20% observations (repeated 30 times), (iii) M2 (model with inflation term as described in Section 3.5) by leaving out observations obtained from the year 2010 onward, consisting 19.8% of the total observations at risk of inflation  $y^{(\text{at-risk})}$  and (iv) prediction of sex ratio transitions based on the prediction setup as described in Section 3.7.2. Median errors and median absolute errors are close to zero for left-out observations. The coverage of 95% and 80% prediction intervals are as expected and symmetrical.

Table 5 shows results for the comparison between model estimates obtained based on the full dataset and based on the training set for the out-of-sample validation exercises in: (i) M1: for the SRB  $\Theta_{c,t}$  and (ii) M2: for the SRB  $\Theta_{c,t}$  and the inflation  $\delta_c \Omega_{c,t}$ . Median errors and the median absolute errors are close to zero. The proportions of updated estimates that fall above or below their respective credible intervals constructed, based on the training set are reasonable and mostly within the expected range, with at most two countries’ estimates falling outside their respective bounds.

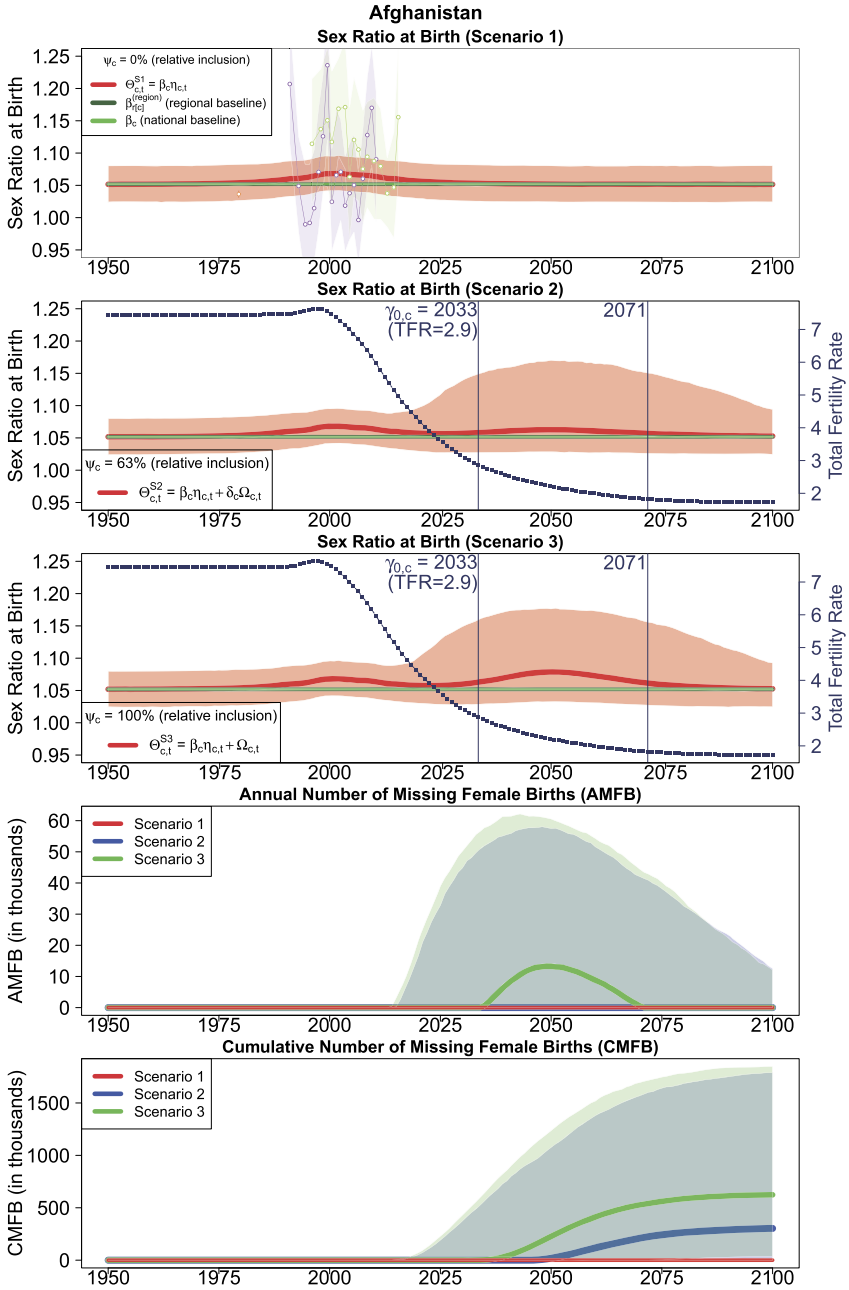


FIG. 7. SRB and missing female births estimates and scenario-based projections for Afghanistan. Row 1: SRB median estimates/projections and 95% credible intervals for Scenario 1  $\Theta_{c,t}^{[S1]}$  (red curve and shades), median estimates of the regional baselines  $\beta_{r[c]}^{(region)}$  (dark green horizontal line), median estimates of the national baselines  $\beta_c$  (light green horizontal line). SRB observations are displayed with dots and observations are connected with lines when obtained from the same source. Shaded areas around observation series represent the sampling variability in the series (quantified by two times the stochastic/sampling standard errors). Row 2: SRB median estimates/projections and 95% credible intervals for Scenario 2  $\Theta_{c,t}^{[S2]}$  (curve and shades). Row 3: SRB median estimates/projections and 95% credible intervals for Scenario 3  $\Theta_{c,t}^{[S3]}$  (curve and shades). The median estimates and projections of total fertility rate (TFR) from the U.N. WPP 2019 are added (blue squared dots). The median estimates of inflation start year  $\gamma_{0,c}$  and end year  $\gamma_{3,c}$  are the vertical lines. The TFR value in the year  $\gamma_{0,c}$  is shown. Row 4: Annual number of missing female births (AMFB) estimates and projections by scenario. Row 5: Cumulative number of missing female births (CMFB) estimates and projections by scenario.

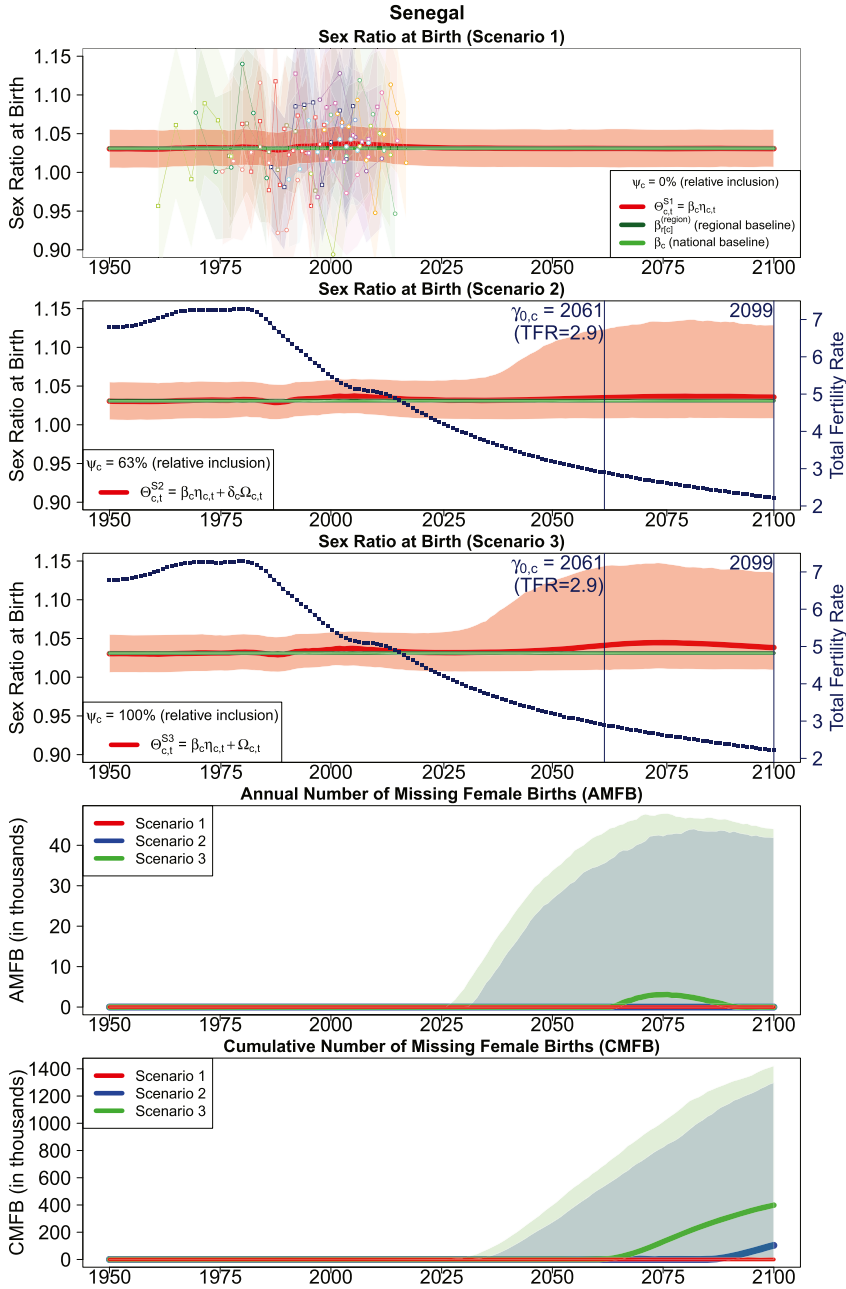


FIG. 8. *SRB and missing female births estimates and scenario-based projections for Senegal. Row 1: SRB median estimates/projections and 95% credible intervals for Scenario 1  $\Theta_{c,t}^{[S1]}$  (red curve and shades), median estimates of the regional baselines  $\beta_{r[c]}^{(\text{region})}$  (dark green horizontal line), median estimates of the national baselines  $\beta_c$  (light green horizontal line). SRB observations are displayed with dots and observations are connected with lines when obtained from the same source. Shaded areas around observation series represent the sampling variability in the series (quantified by two times the stochastic/sampling standard errors). Row 2: SRB median estimates/projections and 95% credible intervals for Scenario 2  $\Theta_{c,t}^{[S2]}$  (curve and shades). Row 3: SRB median estimates/projections and 95% credible intervals for Scenario 3  $\Theta_{c,t}^{[S3]}$  (curve and shades). The median estimates and projections of total fertility rate (TFR) from the UN WPP 2019 are added (blue squared dots). The median estimates of inflation start year  $\gamma_{0,c}$  and end year  $\gamma_{3,c}$  are the vertical lines. The TFR value in the year  $\gamma_{0,c}$  is shown. Row 4: Annual number of missing female births (AMFB) estimates and projections by scenario. Row 5: Cumulative number of missing female births (CMFB) estimates and projections by scenario.*

TABLE 5

Validation results for  $\Theta_{c,t}$  estimates based on training set for M1 and for  $\Theta_{c,t}$  and  $\Omega_{c,t}\delta_c$  estimates based on training set for M2. Error is defined as the difference between an estimate based on full dataset and training set. The percentages (%) of countries in which the SRB median estimates based on the full dataset fall below or above their respective 95% and 80% credible intervals based on the training set are reported. Numbers in the parentheses indicate the number of countries with median estimates based on the full dataset that fall below or above their respective 95% and 80% credible intervals based on the training set. M1: Model without inflation (Section 3.4). M2: Model with inflation (Section 3.5)

Validation (Out-of-Sample)	M1				M2				
# Country in training dataset	176				29				
# Country in full dataset	202				29				
Outcome of interest	$\Theta_{c,t}$				$\delta_c\Omega_{c,t}$				
Reference year	1995	2005	2015	1995	2005	2015	1995	2005	2015
Median error	0.001	0.000	0.001	0.000	0.000	-0.002	0.000	0.000	0.000
Median absolute error	0.003	0.004	0.004	0.001	0.000	0.003	0.000	0.000	0.000
Below 95% credible interval (%)	2.4 (5)	2.4 (5)	0.9 (2)	0.0	0.0	3.4 (1)	0.0	0.0	3.4 (1)
Above 95% credible interval (%)	2.4 (5)	2.8 (6)	2.8 (6)	0.0	0.0	0.0	0.0	0.0	0.0
Expected proportions (%)	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$	$\leq 2.5$
Below 80% credible interval (%)	9.0 (19)	9.0 (19)	4.7 (10)	0.0	0.0	6.9 (2)	0.0	0.0	3.4 (1)
Above 80% credible interval (%)	8.0 (15)	9.4 (20)	8.0 (17)	0.0	3.4 (1)	3.4 (1)	0.0	0.0	3.4 (1)
Expected proportions (%)	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$	$\leq 10$

**5. Discussion.** We described a Bayesian hierarchical time series model for producing SRB estimates and scenario-based projections for all countries from 1950 to 2100. The model produces estimates of SRB baselines by country and region, capturing variations by ethnicity. The model also captures SRB variations under normal circumstance and in settings with sex-selective abortion. The SRB inflation model includes a country-specific indicator to estimate the probability of SRB inflation and identifies countries with strong statistical evidence of SRB inflation in a fully Bayesian approach. Furthermore, the inflation model provides a reproducible way to project SRB inflation based on three scenarios that are data-driven and model-based. Model validation exercises suggest that the SRB models are reasonably well calibrated and have satisfactory predictive performance.

We provide scenario-based probabilistic projections that quantify the SRB inflation if high fertility countries with a son preference would also experience sex ratio transitions, similar to the ones observed so far. In the 17 countries at risk of sex ratio imbalance but without strong statistical evidence of SRB inflation, scenario-based projections S2 and S3 indicate that future sex ratio transitions may happen later this century, when fertility in at-risk countries reaches low levels.

Concerning SRB estimation, our model is based on several model assumptions and is subject to limitations. We summarize here the main ones, as discussed in detail in our prior study [Chao et al. (2019a)]. First, the SRB baselines are modeled according to regional groupings and do not depend on external indicators. Second, aside from the fertility squeeze effect, we are not able to incorporate additional factors in the SRB inflation model that may affect SRB imbalance. Third, out of the 212 countries included in the study, some do not have information to indicate whether they are at risk of SRB imbalance (contributing 3.2% of the global births in 1970–2017). We assumed that these countries have no risk of SRB inflation which is a limitation for monitoring in those specific countries. Lastly, we select countries with risk of SRB inflation prior to model fitting, as opposed to incorporating the selection in the model. This approach was motivated by the need to model the natural fluctuations  $\eta_{c,t}$  and SRB inflation  $\delta_c\Omega_{c,t}$  sequentially to avoid identifiability issues.

The SRB projection scenarios are subject to additional limitations. First, given variability across countries that have started their sex ratio transitions, there is substantial uncertainty associated with the inflation predictions for countries prior to observing country-specific transition data. Second, given that the SRB inflation has been (partially) observed in 12 countries only, the evidence base to extrapolate to other countries is limited. Hence, long-term projections up to 2100 are subject to additional uncertainty that is not captured in our approach. For example, given that no country so far has experienced a reoccurrence of SRB inflation, we do not consider reoccurrence of a sex ratio transition. Lastly, the calculation of missing female births is based on the U.N. WPP projections of the number of total births, as opposed to projections of births that take account of lower or higher SRBs resulting in larger or smaller female cohorts and, subsequently, larger or smaller number of births.

This study focuses on modeling national-level SRB estimation and projection. However, insights from national-level analyses may not be sufficient in national populations with great demographic heterogeneity. Several studies have shown that the SRB can differ greatly across geographic locations or other subpopulations within a country [Chao, KC and Ombao (2020), Chao and Yadav (2019), Guilmoto (2015), Jiang, Ge and Tai (2019)]. For these countries, SRB imbalances may start to decline in some regions while increasing in others, that is, in higher fertility regions. Projections on the national level may mask variability in the sex ratio imbalance at birth across subregions within countries. Constructing subnational scenario-based projections in such settings will provide additional important insights into future missing female births.

While the assumptions made for the scenario-based projections presented in this study may be hypothetical, the associated projections of missing girls are important illustrations of the potential burden of future prenatal sex discrimination and the need to monitor SRBs in countries with son preference. The Sustainable Development Goals (SGDs; see <http://www.un.org/sustainabledevelopment/sustainable-development-goals/>) include the goal to “achieve gender equality and empower all women and girls” by 2030. Monitoring and projecting the sex ratio at birth is an essential part in protecting the gender equality at the prenatal stage. Our scenario-based projections underscore the importance of the monitoring of the sex ratio at birth over time, especially in countries with ongoing inflations and countries where future sex imbalances may occur, to avoid future aborting of girls in favor of male offspring.

## APPENDIX A: MODEL SPECIFICATION AND PRIORS

We fit different variations of the general SRB process model equation,

$$\Theta_{c,t} = \beta_c \eta_{c,t}(\boldsymbol{\phi}) + \delta_c \Omega_{c,t}(\boldsymbol{\zeta}),$$

with details provided in the remainder of this section.

[M1] Model stage 1: The model for the SRB in country-years without SRB inflation,

$$\begin{aligned} \Theta_{c,t} &= \beta_c \eta_{c,t}(\boldsymbol{\phi}), \quad \text{for } c \in \mathcal{C}^{\text{base}}, t \in \{1950, \dots, 2100\}, \\ \log(\beta_c) | \beta_{r[c]}^{(\text{region})}, \sigma_\beta &\sim \mathcal{N}(\log(\beta_{r[c]}^{(\text{region})}), \sigma_\beta^2) \quad \text{for } c \in \mathcal{C}, \\ \beta_r^{(\text{region})} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(1, 1.1) \quad \text{for } r \in \mathcal{R}, \\ \log(\eta_{c,t}) | \boldsymbol{\phi} &\sim \mathcal{N}(0, \sigma_\epsilon^2 / (1 - \rho^2)), \quad \text{for } t = 1950, \\ \log(\eta_{c,t}) &= \rho \log(\eta_{c,t-1}) + \epsilon_{c,t} \quad \text{for } t \in \{1951, \dots, 2100\}, \\ \epsilon_{c,t} | \sigma_\epsilon &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \\ \sigma_\beta &\sim \mathcal{U}(0, 0.05), \end{aligned}$$

$$\rho \sim \mathcal{U}(0, 1),$$

$$\sigma_\epsilon \sim \mathcal{U}(0, 0.05),$$

where  $\mathcal{C}^{\text{base}}$  refers to the set of countries not identified with risk of SRB inflation and  $\mathcal{C}$  is the set of all the 212 countries.  $\mathcal{R}$  is the set of all the 10 regions.  $\phi = \{\rho, \sigma_\epsilon\}$  is the vector of hyperparameters related to  $\eta_{c,t}$ .  $\mathcal{N}(\mu, \sigma^2)$  indicates a normal distribution with mean  $\mu$  and variance  $\sigma^2$ .  $\mathcal{U}(a, b)$  refers to a continuous uniform distribution with lower and upper bounds at  $a$  and  $b$ , respectively.

[M2] Model stage 2: The SRB model for country-years with potential SRB inflation is

$$\Theta_{c,t} = \hat{\beta}_c^{[\text{M1}]} \eta_{c,t}(\hat{\phi}^{[\text{M1}]}) + \delta_c \Omega_{c,t}(\xi)$$

$$\text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}}, t \in \{1950, \dots, 2100\},$$

$$\Omega_{c,t} = \begin{cases} \xi_c(t - \gamma_{0,c})/\lambda_{1,c}, & \gamma_{0,c} < t < \gamma_{1,c}, \\ \xi_c, & \gamma_{1,c} < t < \gamma_{2,c}, \\ \xi_c - \xi_c(t - \gamma_{2,c})/\lambda_{3,c}, & \gamma_{2,c} < t < \gamma_{3,c}, \\ 0, & t < \gamma_{0,c} \text{ or } t > \gamma_{3,c} \end{cases} \quad \text{where}$$

$$\gamma_{1,c} = \gamma_{0,c} + \lambda_{1,c},$$

$$\gamma_{2,c} = \gamma_{1,c} + \lambda_{2,c},$$

$$\gamma_{3,c} = \gamma_{2,c} + \lambda_{3,c},$$

$$\xi_c | \mu_\xi, \sigma_\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2) T(0, \infty) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\lambda_{1,c} | \mu_{\lambda_1}, \sigma_{\lambda_1} \sim \mathcal{N}(\mu_{\lambda_1}, \sigma_{\lambda_1}^2) T(0, \infty) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\lambda_{2,c} | \mu_{\lambda_2}, \sigma_{\lambda_2} \sim \mathcal{N}(\mu_{\lambda_2}, \sigma_{\lambda_2}^2) T(0, \infty) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\lambda_{3,c} | \mu_{\lambda_3}, \sigma_{\lambda_3} \sim \mathcal{N}(\mu_{\lambda_3}, \sigma_{\lambda_3}^2) T(0, \infty) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\gamma_{0,c} | \sigma_\gamma \sim t_3(x_c, \sigma_\gamma^2) T(z_c, \infty) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\delta_c | \pi_c \sim \mathcal{B}(\pi_c) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\text{logit}(\pi_c) | \mu_\pi, \sigma_\pi \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2) \quad \text{for } c \in \mathcal{C}^{\text{inflation}} \cup \mathcal{C}^{\text{future-inf}},$$

$$\mu_\xi \sim \mathcal{U}(0, 2),$$

$$\mu_{\lambda_1} \sim \mathcal{U}(0, 40),$$

$$\mu_{\lambda_2} \sim \mathcal{U}(0, 40),$$

$$\mu_{\lambda_3} \sim \mathcal{U}(0, 40),$$

$$\text{inverse-logit}(\mu_\pi) \sim \mathcal{U}(0, 1),$$

$$\sigma_{\lambda_1} \sim \mathcal{U}(1, 10),$$

$$\sigma_{\lambda_2} \sim \mathcal{U}(1, 10),$$

$$\sigma_{\lambda_3} \sim \mathcal{U}(1, 10),$$

$$\sigma_\xi \sim \mathcal{U}(0, 2),$$

$$\sigma_\gamma \sim \mathcal{U}(0, 10),$$

$$\sigma_\pi \sim \mathcal{U}(0, 2).$$

$\hat{\beta}_c^{[M1]}$  is the posterior median for the national baseline from M1 and  $\hat{\phi}^{[M1]} = \{\hat{\rho}^{[M1]}, \hat{\sigma}_\epsilon^{[M1]}\}$  the vector of posterior medians of  $\phi$ .  $\xi = \{\mu_\xi, \sigma_\xi, \mu_{\lambda 1}, \sigma_{\lambda 1}, \mu_{\lambda 2}, \sigma_{\lambda 2}, \mu_{\lambda 3}, \sigma_{\lambda 3}, \sigma_\gamma\}$  is the vector of hyperparameters related to  $\Omega_{c,t}$ .  $t_3(\mu, \sigma^2)T(a, \infty)$  refers to a Student- $t_3$  distribution with location  $\mu$ , scale  $\sigma$  and lower truncation at  $a$ .  $\mathcal{N}(\cdot)T(a, \infty)$  refers to a truncated normal distribution with lower truncation at  $a$ .

The data quality model for SRB observations is

$$\begin{aligned} \log(y_i) | \Theta_{c[i], t[i]}, \omega_{s[i]} &\sim \mathcal{N}(\log(\Theta_{c[i], t[i]}), \omega_{s[i]}^2 + v_i^2) \quad \text{for } i \in \{1, \dots, n\}, \\ \omega_s &= 0 \quad \text{for } s = \text{CRVS/SRS}, \\ \omega_s &\sim \mathcal{U}(0, 0.5) \quad \text{for } s \in \{\text{Census, DHS, Other DHS, Other}\}. \end{aligned}$$

*Scenario-based projections.* The  $g$ th scenario-based projection trajectory for country  $c$  and year  $t \in \{2018, \dots, 2100\}$ ,  $^{(g)}\Theta_{c,t}^{[S]}$ , is obtained as follows:

$$^{(g)}\Theta_{c,t}^{[S]} = \begin{cases} ^{(g)}\beta_c^{[M1]} ^{(g)}\eta_{c,t}^{[M1]} & \text{for } c \in \mathcal{C}^{\text{base}}, S \in \{S1, S2, S3\}, \\ ^{(g)}\beta_c^{[M1]} ^{(g)}\eta_{c,t}^{[M2]} + ^{(g)}\delta_c^{[M2]} ^{(g)}\Omega_{c,t}^{[M2]} & \text{for } c \in \mathcal{C}^{\text{inflation}}, S \in \{S1, S2, S3\}, \\ ^{(g)}\beta_c^{[M1]} ^{(g)}\eta_{c,t}^{[M3]} & \text{for } c \in \mathcal{C}^{\text{future-inf}}, S = S1, \\ ^{(g)}\beta_c^{[M1]} ^{(g)}\eta_{c,t}^{[M2]} + ^{(g)}\delta_c^{[M2]} ^{(g)}\Omega_{c,t}^{[M2]} & \text{for } c \in \mathcal{C}^{\text{future-inf}}, S = S2, \\ ^{(g)}\beta_c^{[M1]} ^{(g)}\eta_{c,t}^{[M4]} + ^{(g)}\Omega_{c,t}^{[M4]} & \text{for } c \in \mathcal{C}^{\text{future-inf}}, S = S3, \end{cases}$$

where  $\mathcal{C}^{\text{base}}$  refers to countries without risk of SRB inflation,  $\mathcal{C}^{\text{inflation}}$  to countries with strong statistical evidence of SRB inflation and  $\mathcal{C}^{\text{future-inf}}$  to countries at risk of SRB inflation but without strong statistical evidence of SRB inflation.

The models used are defined as follows:

- M1: Model without inflation term,  $\Theta_{c,t} = \beta_c \eta_{c,t}(\phi)$ , fit to dataset  $\mathbf{y}^{(\text{risk-free})}$ , as described in Section 3.4;
- M2: Model with inflation term,  $\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]}) + \delta_c \Omega_{c,t}(\xi)$ , fit to all data from at-risk countries  $\mathbf{y}^{(\text{at-risk})}$ , as described in Section 3.5;
- M3: Model without inflation term,  $\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]})$ , fit to country-specific dataset  $\mathbf{y}^{(c)}$ , used for scenario S1 for  $c \in \mathcal{C}^{\text{future-inf}}$ ;
- M4: Model with inflation term,  $\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]}) + \Omega_{c,t}(\hat{\xi}^{[M2]})$ , fit to country-specific dataset  $\mathbf{y}^{(c)}$ , used for scenario S3 for  $c \in \mathcal{C}^{\text{future-inf}}$ .

## APPENDIX B: COMPUTATION

We obtain posterior samples using a Markov chain Monte Carlo (MCMC) algorithm, implemented in the open source software R 3.6.1 [R Core Team (2019)] and JAGS 4.0.1 [Plummer (2003)], using R-packages rjags [Plummer (2011)], R2jags [Su and Yajima (2011)] and MCMCpack [Martin, Quinn and Park (2011)]. Convergence of the MCMC algorithm and the sufficiency of the number of samples obtained are checked through visual inspection of trace plots and convergence diagnostics of Gelman and Rubin [Gelman and Rubin (1992)], implemented in the coda R-package [Plummer et al. (2006)]. Table 6 summarizes the MCMC specifications for model runs.

Due to the multimodal nature of the posterior distributions for start year parameters in Albania and Republic of Korea, we apply a stacking approach, as discussed in Yao, Vehtari and Gelman (2020), to obtain representative samples from the posterior distribution. Figure 9 shows the posterior density and trace plots of start years pre- and poststacking for Albania and Republic of Korea.



TABLE 6  
MCMC specifications

MCMC Specifications	Normal Model			Inflation Model	
	Full	Validation Out-of-Sample	Validation In-Sample	Full	Validation Out-of-Sample
# Chains	8	8	8	14	8
# Burn-in	8000	8000	8000	7600	1000
# Thinning	20	20	20	10	2
# Posterior samples per parameter	4000	4000	4000	28,000	11,000

In the stacking approach a weight  $w_k$  is assigned to each parallel MCMC chain indexed by  $k \in \{1, 2, \dots, K\}$ , where  $K$  is the total number of parallel chains. The chain-specific weights  $w_k$  are optimized to maximize the leave-one-out cross validation performance of the distribution formed from the weighted average of the sets of simulation draws [Yao, Vehtari and Gelman (2020)]. The weight for each chain is calculated by solving the following:

$$\hat{w} = \arg \max_{w \in \mathbb{S}(N)} \sum_{i=1}^n \log \sum_{k=1}^K w_k f_k(y_i | y_{-i}),$$

where  $\mathbb{S}(N) = \{w : 0 \leq w_k \leq 1, \forall 1 \leq k \leq N; \sum_{k=1}^K w_k = 1\}$  and  $f_k(y_i | y_{-i})$  refers to the pointwise leave-one-out (loo) log predictive density for  $y_i$  based on the posterior samples of model parameters in chain  $k$ . We use the R-package `loo` [Vehtari et al. (2018)] to approximate  $f_k(y_i | y_{-i})$ . After obtaining the weights we resample the parallel chains based on their respective stacking weight. The greater the stacking weight for a chain, the higher the probability that the posterior samples from the chain are resampled. The resulting post-stacking samples are used to produce estimates and credible bounds.

#### APPENDIX C: MODEL COMPARISON FOR M2 RELATED TO ESTIMATING NATIONAL BASELINES SEQUENTIALLY OR JOINTLY

The model equation for M2 (Section 3.5) is as follows:

$$\Theta_{c,t} = \hat{\beta}_c^{[M1]} \eta_{c,t}(\hat{\phi}^{[M1]}) + \delta_c \Omega_{c,t}(\xi),$$

where  $\hat{\beta}_c^{[M1]}$  refers to the point estimate for the national baseline from model M1. We compare the result from M2 to an alternative model, referred to as M2Joint, in which  $\beta_c$  and  $\Omega_{c,t}$  are estimated jointly. Specifically, in model M2Joint, we use country-specific priors on  $\beta_c$ ,

$$\begin{aligned} \Theta_{c,t} &= \beta_c \eta_{c,t}(\hat{\phi}^{[M1]}) + \delta_c \Omega_{c,t}(\xi), \\ \beta_c &\sim \mathcal{N}(\hat{\beta}_{r[c]}^{(\text{region})[M1]}, (\hat{\sigma}_{\beta}^{[M1]})^2), \end{aligned}$$

where  $\hat{\beta}_{r[c]}^{(\text{region})[M1]}$  and  $\hat{\sigma}_{\beta}^{[M1]}$  are posterior medians obtained from M1.

M2Joint model fittings result in estimates of start years  $\gamma_{0,c}$  in China and India that are deemed too late based on external information regarding the timing of first occurrences of sex-selective abortion. Figure 10 compares results between M2 and M2Joint for both countries. There are noticeable difference in the  $\beta_c$  between the two models. The estimated  $\beta_c$ 's from model M2Joint are higher than those from model M2. Consequently, the estimated start years based on model M2Joint are later than the estimated start years based on model M2. For China the M2-based start year is 1980, compared to 1985 for M2Joint. For India, the

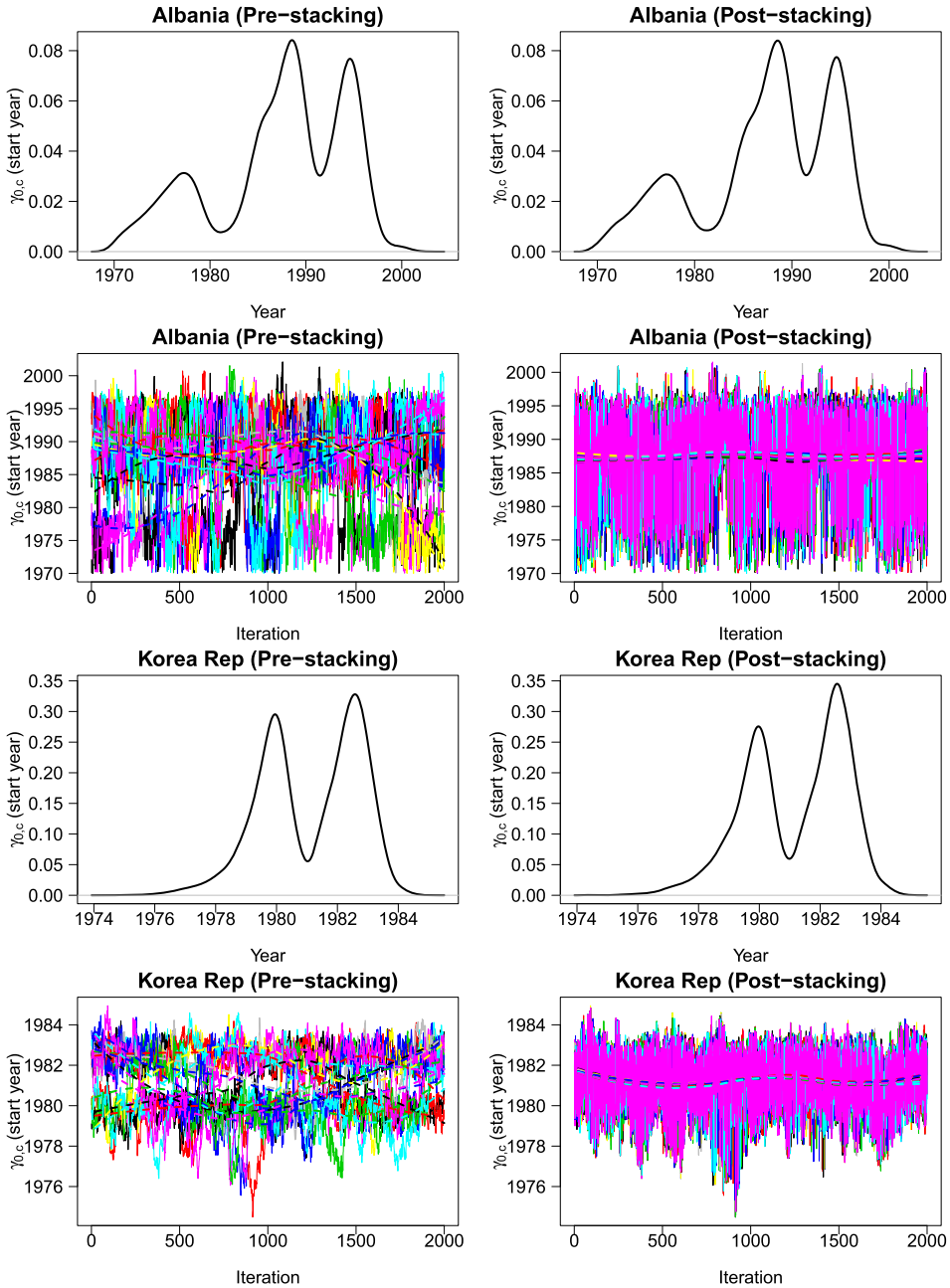


FIG. 9. Posterior density and trace plots of start years pre- and poststacking, Albania and Republic of Korea.

M2-based start year is 1975, compared to 1984 for M2Joint. In China the “one-child policy” was implemented in 1980 and marked the start of a nationwide abortion campaign. Given the timing of this campaign, an estimated start year of 1985 for SRB inflation based on model M2Joint is deemed to be biased upward for China. In India prenatal diagnosis (PD) became available soon after abortion was legalized in 1971. PD was introduced in India as a method for detecting fetal abnormalities but was soon used for prenatal sex selection [Allahbadia (2002), Tandon and Sharma (2006)]. Since then, the combination of prenatal sex determination and selective abortion has been widely used for the systematic elimination of female

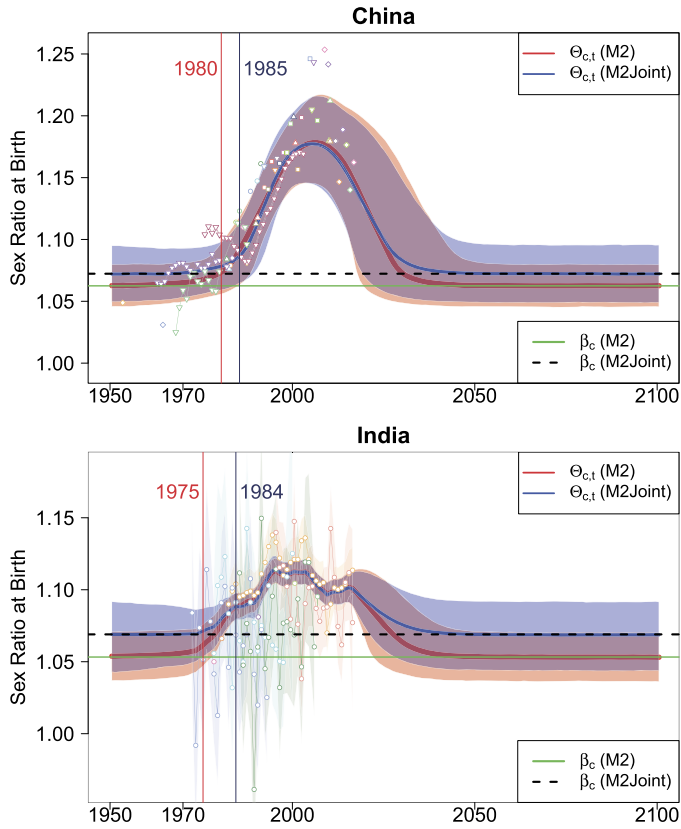


FIG. 10. SRB results comparison between M2 and M2Joint for China and India. M2: estimating  $\beta_c$  and  $\Omega_{c,t}$  sequentially (model presented in Section 3.5; in red). M2Joint: estimating  $\beta_c$  and  $\Omega_{c,t}$  jointly (in blue). SRB median estimates/projections and 95% credible intervals (curves and shades) from M2 (in red) and from M2Joint (in blue). Median estimates of  $\beta_c$  from M2 (light green line).  $\beta_c$  model results for M2Joint (black dashed line). Median estimates of the start year of inflation from M2 (red vertical lines) and from M2Joint (blue vertical lines), with years stated. SRB observations are displayed with dots.

fetuses [Madan and Breuning (2013)]. Given the timing of the introduction of PD in 1971, start year 1984 based on model M2Joint is deemed biased upward.

**Acknowledgments.** The authors would like to thank the anonymous referees, the Associate Editor and the Editor for their insightful and constructive comments that improved the quality of this paper. The study described is solely the responsibility of the authors and does not necessarily represent the official views of the United Nations.

**Funding.** This work was supported by a research grant from the National University of Singapore.

SUPPLEMENTARY MATERIAL

**Supplement figure A: Scenario-based SRB projection during 1950–2100, by country** (DOI: 10.1214/20-AOAS1436SUPPA; .pdf). SRB median estimates/projections (curve) and 95% credible intervals (shades) for Scenario 1  $\Theta_{c,t}^{[S1]}$  (in red), Scenario 2  $\Theta_{c,t}^{[S2]}$  (in blue) and Scenario 3  $\Theta_{c,t}^{[S3]}$  (in green), and median estimates of the national baselines  $\beta_c$  (black dashed horizontal line). The median estimates and projections of total fertility rate (TFR) from the UN WPP 2019 are added (blue squared dots). The median estimates of inflation start year  $\gamma_{0,c}$  and end year  $\gamma_{3,c}$  are the vertical lines. The TFR value in the year  $\gamma_{0,c}$  is shown.

**Supplement figure B: Scenario 1 SRB projection during 1950–2100 for all countries** (DOI: [10.1214/20-AOAS1436SUPPB](https://doi.org/10.1214/20-AOAS1436SUPPB); .pdf). SRB median estimates/projections and 95% credible intervals for Scenario 1  $\Theta_{c,t}^{[S1]}$  (red curve and shades), median estimates of the regional baselines  $\beta_{r[c]}^{(\text{region})}$  (dark green horizontal line), median estimates of the national baselines  $\beta_c$  (light green horizontal line). SRB observations are displayed with dots and observations are connected with lines when obtained from the same source. Shaded areas around observation series represent the sampling variability in the series (quantified by two times the stochastic/sampling standard errors). Model results are shown before 1950 if observations are available prior 1950.

## REFERENCES

- ALKEMA, L., WONG, M. B. and SEAH, P. R. (2012). Monitoring progress towards Millennium Development Goal 4: A call for improved validation of under-five mortality rate estimates. *Statistics, Politics, and Policy* **3**.
- ALKEMA, L., CHAO, F., YOU, D., PEDERSEN, J. and SAWYER, C. C. (2014). National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: A systematic assessment. *Lancet Glob Health* **2** e521–e530. [https://doi.org/10.1016/S2214-109X\(14\)70280-3](https://doi.org/10.1016/S2214-109X(14)70280-3)
- ALLAHBADIA, G. N. (2002). The 50 million missing women. *Journal of Assisted Reproduction and Genetics* **19** 411–416.
- ATTANÉ, I. and GUILMOTO, C. Z. (2007). *Watering the Neighbour's Garden: The Growing Demographic Female Deficit in Asia*. Committee for International Cooperation in National Research in Demography, Paris.
- BASTEN, S. and VERROPOULOU, G. (2013). Maternity migration and the increased sex ratio at birth in Hong Kong SAR. *Popul. Stud.* **67** 323–334.
- BONGAARTS, J. (2013). The implementation of preferences for male offspring. *Popul. Dev. Rev.* **39** 185–208.
- BONGAARTS, J. and GUILMOTO, C. Z. (2015). How many more missing women? Excess female mortality and prenatal sex selection, 1970–2050. *Popul. Dev. Rev.* **41** 241–269.
- CATALANO, R., BRUCKNER, T., MARKS, A. R. and ESKENAZI, B. (2006). Exogenous shocks to the human sex ratio: The case of September 11, 2001 in New York city. *Human Reproduction* **21** 3127–3131.
- CHAHNAZARIAN, A. (1988). Determinants of the sex ratio at birth: Review of recent literature. *Social Biology* **35** 214–235.
- CHAO, F., KC, S. and Ombao, H. (2020). Levels and trends in the sex ratio at birth in seven provinces of Nepal between 1980 and 2016 with probabilistic projections to 2050: A Bayesian modeling approach. Preprint. Available at [arXiv:2007.00437](https://arxiv.org/abs/2007.00437).
- CHAO, F. and YADAV, A. K. (2019). Levels and trends in the sex ratio at birth and missing female births for 29 states and union territories in India 1990–2016: A Bayesian modeling study. *Foundations of Data Science* **1** 177–196.
- CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019a). Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. *Proc. Natl. Acad. Sci. USA* **116** 9303–9311.
- CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019b). SRB database (pnas.1812593116.sd01). Available at [https://www.pnas.org/highwire/filestream/859048/field\\_highwire\\_adjunct\\_files/1/pnas.1812593116.sd01.xlsx](https://www.pnas.org/highwire/filestream/859048/field_highwire_adjunct_files/1/pnas.1812593116.sd01.xlsx).
- CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019c). Web appendix systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. Available at <https://www.pnas.org/content/pnas/suppl/2019/04/10/1812593116.DCSupplemental/pnas.1812593116.sapp.pdf>. <https://doi.org/10.6084/m9.figshare.12442373>
- CHAO, F., GUILMOTO, C. Z., KC, S. and Ombao, H. (2020). Probabilistic projection of the sex ratio at birth and missing female births by state and union territory in India. *PLoS ONE* **15** e0236673.
- CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2021a). Supplement to “Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model.” <https://doi.org/10.1214/20-AOAS1436SUPPA>
- CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2021b). Supplement to “Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model.” <https://doi.org/10.1214/20-AOAS1436SUPPB>
- CHEN, C., CHOU, S.-Y., GIMENEZ, L. and LIU, J.-T. (2020). The quantity of education and preference for sons: Evidence from Taiwan's compulsory education reform. *China Econ. Rev.* **59** 101369.
- CHOI, E. J. and HWANG, J. (2020). Transition of son preference: Evidence from South Korea. *Demography* **1**–26.

- DAS GUPTA, M., ZHENGHUA, J., BOHUA, L., ZHENMING, X., CHUNG, W. and HWA-OK, B. (2003). Why is son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea. *The Journal of Development Studies* **40** 153–187.
- DRÉZE, J. and SEN, A. (1990). *Hunger and Public Action*. Clarendon Press, Oxford.
- DUBUC, S. and COLEMAN, D. (2007). An increase in the sex ratio of births to India-born mothers in England and Wales: Evidence for sex-selective abortion. *Popul. Dev. Rev.* **33** 383–400.
- DUTHÉ, G., MESLÉ, F., VALLIN, J., BADURASHVILI, I. and KUYUMJYAN, K. (2012). High sex ratios at birth in the Caucasus: Modern technology to satisfy old desires. *Popul. Dev. Rev.* **38** 487–501.
- FUKUDA, M., FUKUDA, K., SHIMIZU, T. and MØLLER, H. (1998). Decline in sex ratio at birth after Kobe earthquake. *Human Reproduction* **13** 2321–2322.
- GARENNE, M. (2002). Sex ratios at birth in African populations: A review of survey data. *Hum. Biol.* **74** 889–900. <https://doi.org/10.1353/hub.2003.0003>
- GARENNE, M. (2008). Poisson variations of the sex ratio at birth in African demographic surveys. *Hum. Biol.* **80** 473–482.
- GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](https://doi.org/10.1201/b16006)
- GEORGE, S. M. (2002). Sex selection/determination in India: Contemporary developments. *Reproductive Health Matters* **10** 190–192.
- GOODKIND, D. (1997). *Sex-Selective Abortion, Reproductive Rights, and the Greater Locus of Gender Discrimination in Family Formation: Cairo's Unresolved Questions*. Univ. Michigan, Population Studies Center.
- GOODKIND, D. (2011). Child underreporting, fertility, and sex ratio imbalance in China. *Demography* **48** 291–316. <https://doi.org/10.1007/s13524-010-0007-y>
- GRAFFELMAN, J. and HOEKSTRA, R. F. (2000). A statistical analysis of the effect of warfare on the human secondary sex ratio. *Hum. Biol.* 433–445.
- GUILMOTO, C. Z. (2009). The sex ratio transition in Asia. *Popul. Dev. Rev.* **35** 519–549.
- GUILMOTO, C. Z. (2012a). Sex imbalances at birth: Trends, consequences and policy implications. UNFPA, United Nations Population Fund of Asia and the Pacific Regional Office, Thailand.
- GUILMOTO, C. Z. (2012b). Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100. *Demography* **49** 77–100. <https://doi.org/10.1007/s13524-011-0083-7>
- GUILMOTO, C. Z. (2012c). Son preference, sex selection, and kinship in Vietnam. *Popul. Dev. Rev.* **38** 31–54. <https://doi.org/10.1111/j.1728-4457.2012.00471.x>
- GUILMOTO, C. Z. (2015). Mapping the diversity of gender preferences and sex imbalances in Indonesia in 2010. *Popul. Stud. (Camb.)* **69** 299–315. <https://doi.org/10.1080/00324728.2015.1091603>
- GUILMOTO, C. Z., CHAO, F. and KULKARNI, P. M. (2020). On the estimation of female births missing due to prenatal sex selection. *Popul. Stud. (Camb.)* **74** 283–289. <https://doi.org/10.1080/00324728.2020.1762912>
- GUILMOTO, C. Z., HOÀNG, X. and VAN, T. N. (2009). Recent increase in sex ratio at birth in viet nam. *PLoS ONE* **4** e4624.
- GUILMOTO, C. Z. and REN, Q. (2011). Socio-economic differentials in birth masculinity in China. *Development and Change* **42** 1269–1296.
- GUPTA, M. D., CHUNG, W. and SHUZHUO, L. (2009). Evidence for an incipient decline in numbers of missing girls in China and India. *Popul. Dev. Rev.* **35** 401–416.
- HUDSON, V. M. and DEN BOER, A. (2004). *Bare Branches: The Security Implications of Asia's Surplus Male Population*. MIT Press, Cambridge.
- JAMES, W. H. (1984). The sex ratios of black births. *Ann. Hum. Biol.* **11** 39–44. <https://doi.org/10.1080/03014468400006871>
- JAMES, W. H. (1985). The sex ratio of oriental births. *Ann. Hum. Biol.* **12** 485–487.
- JAMES, W. H. (1987). The human sex ratio. Part 1: A review of the literature. *Hum. Biol.* 721–752.
- JIANG, Q., GE, T. and TAI, X. (2019). Change in China's sex ratio at birth since 2000: A decomposition at the provincial level. *Appl. Spat. Anal. Policy* 1–28.
- KABA, A. J. (2008). Sex ratio at birth and racial differences: Why do black women give birth to more females than non-black women? *Afr. J. Reprod. Health* **12** 139–150.
- LIN, T.-C. (2009). The decline of son preference and rise of gender indifference in Taiwan since 1990. *Demogr. Res.* **20** 377–402. <https://doi.org/10.4054/DemRes.2009.20.16>
- LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. [MR0415861](https://doi.org/10.2307/2344111)
- MADAN, K. and BREUNING, M. H. (2013). Impact of prenatal technologies on the sex ratio in India: An overview. *Genetics in Medicine* **16** 425–432.

- MARCUS, M., KIELY, J., XU, F., MCGEEHIN, M., JACKSON, R. and SINKS, T. (1998). Changing sex ratio in the United States, 1969–1995. *Fertility and Sterility* **70** 270–273.
- MARTIN, A. D., QUINN, K. M. and PARK, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *J. Stat. Softw.* **42** 22.
- MATHEWS, T. J. and HAMILTON, B. E. (2005). Trend analysis of the sex ratio at birth in the United States. *Natl. Vital Stat. Rep.* **53** 1–17.
- MESLÉ, F., VALLIN, J. and BADURASHVILI, I. (2007). A sharp increase in sex ratio at birth in the Caucasus. Why? How? *Watering the Neighbour's Garden: The Growing Demographic Female Deficit in Asia, Paris: Committee for International Cooperation in National Research in Demography* 73–88.
- OOMMAN, N. and GANATRA, B. R. (2002). Sex selection: The systematic elimination of girls. *Reproductive Health Matters* **10** 184–188.
- PARK, C. B. and CHO, N.-H. (1995). Consequences of son preference in a low-fertility society: Imbalance of the sex ratio at birth in Korea. *Popul. Dev. Rev.* 59–84.
- PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria.
- PLUMMER, M. (2011). rjags: Bayesian graphical models using MCMC. R package version 3-5.
- PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.
- R CORE TEAM (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- RUDER, A. (1985). Paternal-age and birth-order effect on the human secondary sex ratio. *Am. J. Hum. Genet.* **37** 362–372.
- ŠEVČÍKOVÁ, H., ALKEMA, L. and RAFTERY, A. E. (2011). bayesTFR: An R package for probabilistic projections of the total fertility rate. *J. Stat. Softw.* **43** 1–29.
- ŠEVČÍKOVÁ, H., ALKEMA, L., RAFTERY, A. E., FOSDICK, B. and GERLAND, P. (2019). bayesTFR: Bayesian fertility projection. R package and documentation version 6.4-0. R-CRAN. Available at <https://cran.r-project.org/web/packages/bayesTFR/>. Accessed 16 Feb 2020.
- SONG, S. (2012). Does famine influence sex ratio at birth? Evidence from the 1959–1961 great leap forward famine in China. *Proc. Roy. Soc., Biol. Sci.* **279** 2883–2890.
- SU, Y. S. and YAJIMA, M. (2011). R2jags: A package for running jags from R. R package version 0.02-17.
- TAFURO, S. and GUILMOTO, C. Z. (2020). Skewed sex ratios at birth: A review of global trends. *Early Hum. Dev.* **141** 104868. <https://doi.org/10.1016/j.earlhumdev.2019.104868>
- TANDON, S. L. and SHARMA, R. (2006). Female foeticide and infanticide in India: An analysis of crimes against girl children. *International Journal of Criminal Justice Sciences* **1**.
- UNITED NATIONS, DESA (2019). World population prospects: The 2019 revision. Available at <http://esa.un.org/unpd/wpp/Download/Standard/Population/>. Accessed 16 Feb 2020.
- VEHTARI, A., GELMAN, A., YAO, Y. and GABRY, J. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.1.0 1003.
- VENERO FERNÁNDEZ, S. J., MEDINA, R. S., BRITTON, J. and FOGARTY, A. W. (2011). The association between living through a prolonged economic depression and the male: Female birth ratio—a longitudinal study from Cuba, 1960–2008. *Am. J. Epidemiol.* **174** 1327–1331.
- VISARIA, P. M. (1967). Sex ratio at birth in territories with a relatively complete registration. *Eugenics Quarterly* **14** 132–142.
- VOLLSET, S. E., GOREN, E., YUAN, C.-W., CAO, J., SMITH, A. E., HSIAO, T., BISIGNANO, C., AZHAR, G. S., CASTRO, E. et al. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: A forecasting analysis for the global burden of disease study. *Lancet*.
- VU, T. M. and YAMADA, H. (2020). Sex ratio and religion in Vietnam. *Munich Personal RePEc Archive*.
- YAO, Y., VEHTARI, A. and GELMAN, A. (2020). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. Preprint. Available at [arXiv:2006.12335](https://arxiv.org/abs/2006.12335).