The Subnational Urbanization Projections for China, India, and the U.S.

Leiwen Jiang and Hamid Zoraghein

Abstract

The NCAR Community Demographic Model was used to generate global national urbanization projections for all countries under the Shared Socioeconomic Pathways (SSPs) – the new IPCC socioeconomic scenarios. However, the climate change research communities require subnational urbanization projection to account for the large variations in urban growth of big countries in developing extended SSP scenarios. This study takes advantage of the new development of the Projection Model and newly available data to project urbanization trends of subnational regions (provinces of China, states of India and the U.S., as examples). We carried out validation analysis through comparing projected urbanization trends in the past decades against observed urbanization records at both national and subnational levels. It shows that the improved model produces reasonable and unbiased projection outcomes, covering a wide range of plausible urbanization paths for the long- and medium-term for countries and subnational regions.

1. Introduction

Urbanization is one of the most profound socioeconomic and environmental transformations in the human history. Projections of future trends in urban growth are important because virtually all world population growth and most global economic growth are expected to occur in urban areas (Jiang and O'Neill 2015). In the meantime, urban areas in changing climate systems will be more exposed to climate hazards such as heat extremes and sea level rise, making the urban population more susceptible (Jones *et al.* 2015, Jones and O'Neill 2016). Therefore, climate change studies require better understanding of future urbanization trends to support analyses of emissions and mitigation options as well as vulnerability to impacts. Subsequently, the Intergovernmental Panel on Climate Change (IPCC) new socioeconomic scenarios – the Shared Socioeconomic Pathways (SSPs), for the first time, include urbanization projections as one of the key elements that determine the future challenges to climate change mitigation and adaptation (Jiang 2014).

The NCAR CDM Urbanization Projection Model has been developed to meet the requirement of climate change research communities for understanding the long-term and alternative urbanization trends. The model was used to generate global national urbanization projections for all countries under the frameworks of the SSPs. The detailed description of the model and the global alternative national urbanization projections can be found in Jiang and O'Neill (2017).

The climate change research communities have been working to develop the socioeconomic scenarios, extended for understanding the human-climate systems interactions in the socioeconomic, geographic, and institutional contexts of regional and subnational areas. There is increasing demand for urbanization projections for subnational regions of big countries which have large variation in urban growth (Jiang, Zoraghein, and O'Neill 2017).

During the recent years, we improved the previous model and the projection results in which projections for a region are now based on more historical data points of selected reference regions, and the selection of the reference regions is now conducted in a more systematic and holistic way (Zoraghein and Jiang 2018). Using the improved NCAR Urbanization Projection Model, in this paper, we make urbanization projections for the provinces of China, and states of India and the U.S. based on historical urbanization records of national and subnational regions. The model can also be used to make urbanization projections for subnational regions of other countries when subnational historical urbanization records of these countries are available. This capability is especially useful for large countries where urbanization levels and trends vary significantly across subnational regions.

To evaluate the performance of the model, we carry out validation analysis through comparing projected urbanization trends in the past decades by the model against observed urbanization records. Our analysis shows that the improved CDM-Urbanization Projection Model produces reasonable and unbiased projection outcomes, covering a wide range of plausible urbanization paths for the long- or medium-term for the subnational regions under study.

2. Data and Methods

The NCAR Urbanization Projection previously uses the data for making urbanization projections at the national level is from the UN Urbanization Prospects 2014 Revision, which includes historical data on urbanization levels of 220 countries from 1950 to 2015 at a 5-year interval (United Nations 2014), and the UN Population Prospects 2017 Revision, which includes historical population counts of all countries for the same time period (United Nations 2017). At subnational level, we collected historical population counts and urbanization records of U.S. states from the censuses during the period of 1900 to 2010 at 10-year intervals. As the CDM-Urbanization Projection Model runs at a 5-year step and requires the input data arranged in the same manner, we used a linear interpolation to derive values for the middle years between every two censuses. For China and India, the time frames are not as extensive and consistent as for the U.S. While the majority of Chinese provinces have historical population and urbanization records from 1950 to 2010, some lack the values in early years. Historical records of most states of India are available from 1971 to 2011, with some states missing data in the early years. To overcome the problem of inconsistent time frames in the original datasets, the model is designed to adapt itself to employ reference regions with different formats. Since the India census is conducted in the second year of each decade (1971, 1981…) while most other countries carry out their census surveys in the first year (1970, 1980 …), we assumed no big change during one year and similarly used 1970, 1980 … as the collection years for India. On the contrary to the U.S., historical records of subnational population and urbanization in China are reported at 5-year intervals. We could therefore directly include the original records in the datasets. Finally, we merged urbanization levels and population counts of the U.S., China and India, respectively, to generate two csv datasets of subnational units, one containing population and area values of the three countries, and the other including their urbanization records.

Table 1 summarizes the main characteristics of the original datasets used in this paper.

Table 1. Datasets and their main characteristics.

Dataset	Time frame	Interval	Count	Format
National population	1950-2015	5	233	x 1000

National urbanization level	1950-2015	5	220[1]	Percentage
U.S. state population	1900-2010	10	50	x1000
U.S. state urbanization level	1900-2010	10	50	Percentage
China province population	1950-2010 (inconsistent)	5	31	x 1000
China province urbanization level	1950-2010 (inconsistent)	5	31	Percentage
India state population	1970-2010 (inconsistent)	10	35	x 1000
India state urbanization level	1970-2010 (inconsistent)	10	35	Percentage

The previous CDM Urbanization Projection Model was constructed as an extension of and improvement to the U.N. urbanization projection model (Jiang and O'Neill 2017). The U.N. model projects the urbanization level of a target country as a function of the difference between the urban and rural population growth rates. It establishes a linear statistical relationship between these two variables based on historical records. More specifically, it models changes in the urban-rural ratio (URR_t) and therefore urbanization levels as a function of differences between urban and rural population growth rates urr_t (Equation 1) where growth rate differences are a function of urbanization levels (PU_t) (Equation 2):

$$URR_{t+1} = URR_t \times e^{urr_t} \tag{1}$$

$$urr_t = f(PU_t) \tag{2}$$

In Equation 2, f is the global, linear, empirical relationship derived from all historical data. Jiang and O'Neill (2017) modify the U.N. methodology by defining the linear regression separately for each region in 2 stages rather than a single projection to account for uncertainty and distinct socioeconomic pathways represented by the SSPs over the 2010-2100 time frame.

During the most recent years, improves the previous model in Jiang and O'Neill (2017) in several ways. First, the improved model uses all original historical records of urbanization levels and differences in urban-rural population growth rates across all years instead of averaging them for each time lag. Second, the number of years to be included for the second stage is not limited to 35 years unlike the previous model. Third, it conducts the first and second refinement steps for

[1] Thirteen of the 233 countries or regions which have either reached 100% urban or have no data are excluded from the dataset.

identifying reference regions simultaneously. Fourth, it has the ability to use both national and subnational (i.e. U.S., China and India) historical observations for subnational urbanization projections and incorporate additional subnational data of other countries when they become available. Subsequently, the concept region in the improved model encompasses both countries and subnational units as opposed to the previous model, in which it is limited to countries.

The ability of the model to incorporate original historical records across all years without averaging them allows the incorporation of more data points (pairs of urbanization level and difference in urban-rural population growth rates) for the regression analysis. The idea of averaging was introduced in the previous model to make it less prone to outliers. In the improved model, we identified outliers as regions whose difference in annual urban-rural population growth rates is higher than 0.2% despite their urbanization level already being more than 90%. By introducing a threshold to exclude these instances, the new approach is no longer sensitive to outliers, and all data points can contribute to the estimation of regression coefficients.

Reference regions might become too few and diverse as target urbanization levels increase, particularly for the second stage where projected urbanization levels from the first stage might have already reached high values under the central and fast scenarios. Therefore, averaging their urbanization levels and differences in urban-rural population growth rates generates biased unrepresentative values for years after 35. However, using all original values in the analysis, the improved model has no limitation to integrate them from later years. This increases the number of data points for the regression analysis, making the coefficients more robust and representative.

The previous model does the first and second refinement steps for identifying reference regions separately. It identifies the first year in which the historical urbanization record of a potential reference region falls within the target range and then evaluates its difference in urban-rural population growth rates over the past 5 or 10 years. The drawback of this approach is that the model might erroneously discard a potential reference region that does not meet the 75% threshold of the second refinement step in the first year but satisfies the threshold in a later year when its urbanization level is still within the target range. In fact, the region being discarded from the reference might have a rather similar urbanization experience to the target region in a later year. The improved model is now able to assess a potential reference region over all years of its historical records and picks the year in which both the urbanization level is within the target range and the prior difference in urban-rural population growth rates is the most similar to the target region. The new approach has two benefits over the previous one. First, it does not exclude any potential reference region until evaluating its historical records entirely. Second, it applies the most similar historical pattern of each reference region for defining slow, central and fast reference groups, and informing the calculation of regression coefficients in Equation 2 and consequently the urbanization projection for the target region. Figure 1 provides an example when Italy is selected as a potential reference region for projecting urbanization trends of Armenia. In the previous model, Italy would be selected as a fast urbanization reference region because of rapid urbanization growth the country had experienced over the 30 years after reaching the target urbanization level for the first time in 1960. In the improved model, on the other hand, Italy is selected as a central or slow reference region based on its modest urbanization change after 1985, - the year when the urbanization level of the country was still within the target urbanization range and its historical urbanization growth rate was more similar to the present Armenia. Accordingly, the improved model extracts the historical urbanization part of Italy that better resembles the present urbanization regime of Armenia.

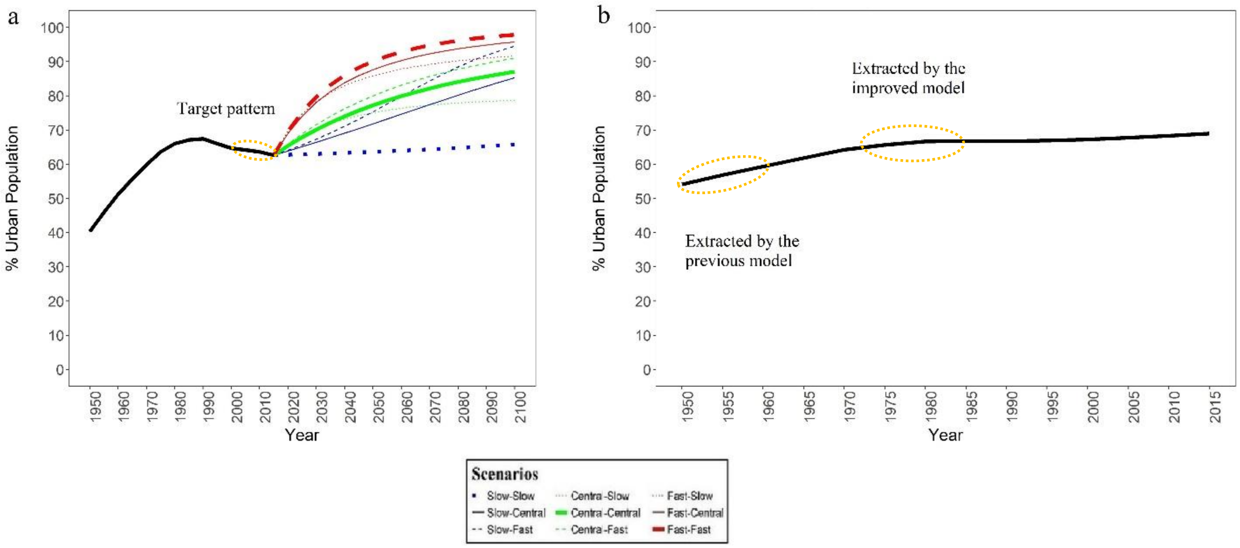

Figure 1. Difference in pattern extraction for (a) the target region, Armenia from (b) the reference region, Italy.

The improved model has also the ability to make subnational urbanization projections through integrating subnational time series of different countries. It currently incorporates historical subnational population and urbanization records of the U.S., India and China, but can also include subnational data from other countries. This new feature enables the model to contribute to developing urbanization projections for the extended SSPs that go beyond national level projections, and pave the way for scholars to possibly modify the SSPs to better meet their requirements of analysis at local or regional scales. Currently, we have only collected historical subnational data from China, India and the U.S. In the subnational urbanization projection for these three countries, we can follow two approaches of projection, one that uses only the subnational datasets for selecting reference regions, and the other that uses the combined global national and subnational level datasets.

Table 2 outlines the main improvements offered by the new CDM Urbanization Projection Model over the previous one. Figure 2 displays the workflow diagram of the new model. In addition to the new capability of projecting urbanization levels at the subnational level, the improved model produces projection results that cover a wider range of uncertainty for scenario building and assessment. This is because the new model can better reflect the paths of declining urbanization levels, and it has fewer instances of projections under the slow-fast scenario surpassing those under the fast-fast scenario.

In this paper, we regard the more thorough and optimum extraction of historical patterns of reference regions as the primary improvement over the previous model introduced in Jiang and O'Neill (2017). Therefore, we modified the previous model by incorporating the other enhancements listed in Table 2 to focus on the differences between the previous and improved models specifically caused by the primary improvement. Consequently, all the results pertaining to the previous model in Section 4 come from applying all additional improvements to the model except the optimum and simultaneous pattern extraction.

Table 2. Main differences between the previous and improved CDM urbanization projection models.

Previous model	Improved model
Averages urbanization records and differences in urban-rural population growth rates of reference regions	Uses original urbanization records and differences in urban-rural population growth rates of reference regions
Limits the number of years to be included for the second stage of projection to 35 years	Uses records collected in as many years as available for the second stage of projection
Implements the first and second refinements separately	Implements the first and second refinements simultaneously and comprehensively
Is limited to national level projection	Is capable of both national and subnational level projections

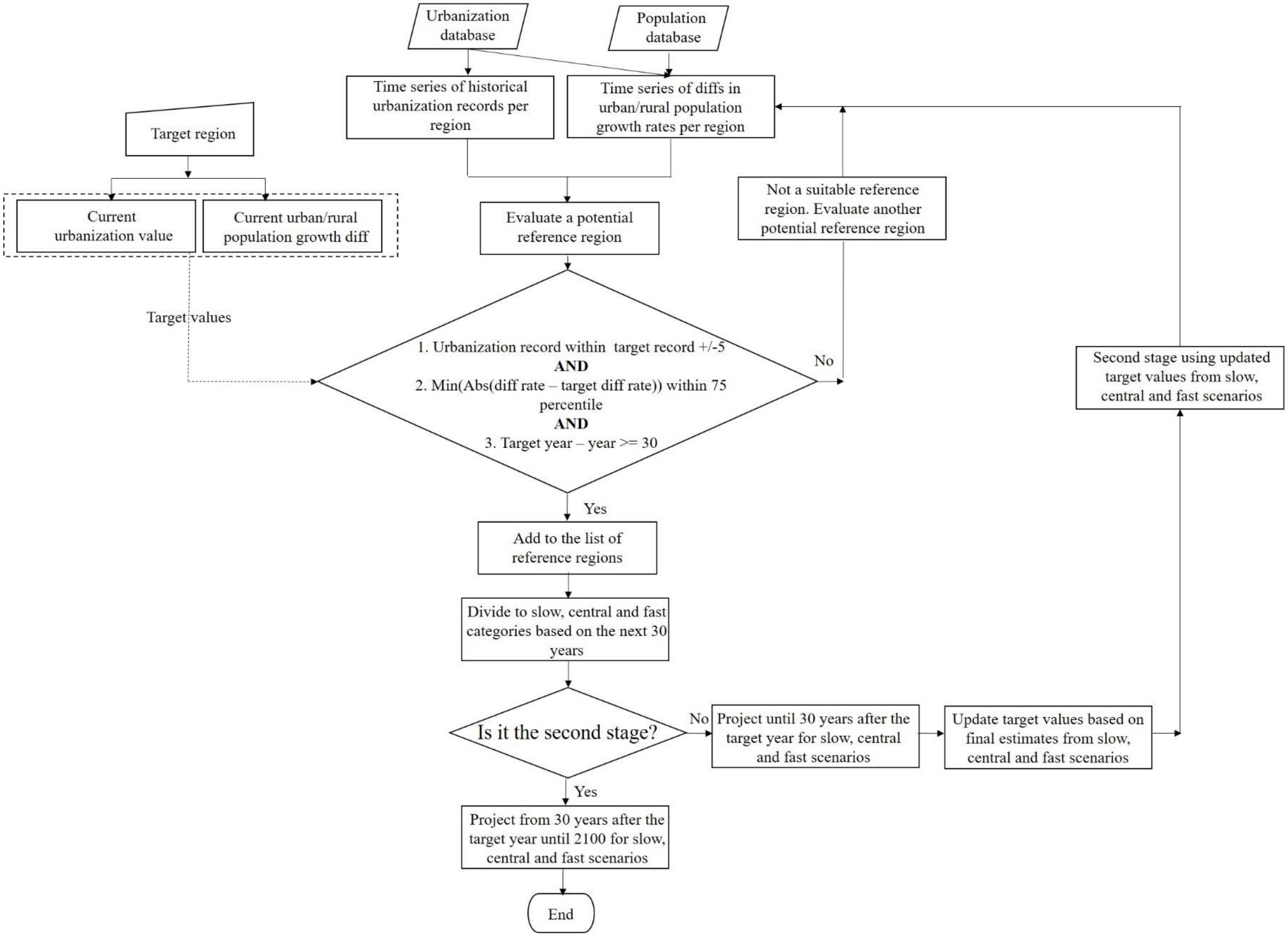
Urbanization database
Population database
Time series of historical urbanization records per region
Time series of diffs in urban/rural population growth rates per region
Target region
Current urbanization value
Current urban/rural population growth diff
Evaluate a potential reference region
Not a suitable reference region. Evaluate another potential reference region
Target values
1. Urbanization record within target record +/-5
AND
2. Min(Abs(diff rate – target diff rate)) within 75 percentile
AND
3. Target year – year >= 30
No
Second stage using updated target values from slow, central and fast scenarios
Yes
Add to the list of reference regions
Divide to slow, central and fast categories based on the next 30 years
Is it the second stage?
No
Project until 30 years after the target year for slow, central and fast scenarios
Update target values based on final estimates from slow, central and fast scenarios
Yes
Project from 30 years after the target year until 2100 for slow, central and fast scenarios
End

Figure 2. Main steps of the new CDM Urbanization Projection Model.

3.3. Validation

We conducted validation analyses to evaluate the performance of the model through comparing projected urbanization results to the actual records over historical periods. The validation analysis was conducted for various historical periods with different starting time points. It is noteworthy that a target region should have at least 10 years of prior observations to ensure the applicability of the second refinement step. For instance, the earliest starting time for validation analysis at the national level is 1960 – we call it the "validation year." The earlier the validation year, the more available historical records to compare to projected values for validation.

We compared actual historical records to projected values of the 9 urbanization pathways for each target region. The analysis help us answer the following first two questions with respect to each target region and the third question with regard to the overall performance of the model:

- To what extent does the range of projected pathways cover the variations of the historical urbanization trend of the target region?
- Which projected urbanization pathway and how well the pathway does reflect its actual historical trend?
- Does the model produce unbiased or non-skewed outcomes compared to historical records?

To answer the first question, we recorded the number of times historical urbanization records of a target region fell within the ranges provided by its projection. Therefore, we extracted the minimum and maximum projected values in each year, corresponding to the slow-slow and fast-fast pathways, and compared them to the historical observation in that year. We summed all instances where the historical observation lay between the minimum and maximum values.

For addressing the second question, we did a two stage comparison to identify the most similar pathway to the historical trend. The first stage represented short-term changes and included the first 30 years after the validation year while the second stage reflected long-term changes and encompassed the remaining years. For the first stage, we averaged projected values corresponding to the slow, central and fast scenarios in each year. We then calculated absolute differences between these representative values and historical records across the time frame. Afterwards, we identified the scenario with the lowest mean absolute difference as the most similar one over short time. For the second stage, we calculated the average absolute differences between historical records and projected values corresponding to the three subcategories of the scenario identified from the first stage across the second time frame. We finally labelled the subcategory with the minimum difference at this stage combined with the scenario from the first stage as the most similar pathway to the historical trend of the target region. For example, historical records of Afghanistan start as of 1950. Therefore, the earliest validation year is 1960, to ensure a 10-year urbanization experience prior to the validation year. The short term period spans 1965 to 1990 and the long term period covers 1995 to 2015. There are 9 projected pathways for Afghanistan during 1965-1990, namely slow-slow, slow-central, slow-fast, central-slow, central-central, central-fast, fast-slow, fast-central, and fast-fast. We averaged the three projected values corresponding to each main category, namely slow, central and fast in each year. Thus, we had three sets of averaged values for 1965 to 1990. Then we calculated the absolute differences between the historical urbanization records of Afghanistan over the period and each of the three sets. We identified the category corresponding to the set with the lowest average of absolute differences as the most similar scenario for the short term period. For Afghanistan, this scenario was central. Afterwards, we extracted the three central subcategories over the 1995-2015 period for the second stage. We then calculated the averages of absolute differences between the historical records and the three extracted sets. We considered the scenario associated with the set with the minimum value the closest scenario for the long term period. This scenario was again central for Afghanistan. Finally, based on the outcomes, we determined the “central-central” scenario as the most similar pathway to the historical urbanization records of Afghanistan overall. The characterization of “central-central” urbanization pathway of Afghanistan, to a great extent, is in agreement with other authors. A World Bank report points out that while the whole South Asian region had a slow urbanization trend, Afghanistan was the country with a relatively high urban growth rate that meets the global average during the past decade (Ellis and Roberts 2016).

To respond to the third question, we adopted the rank histogram method, which is a diagnostic tool to evaluate if the distribution of projection results represents an unbiased or non-skewed outcome. The underlying assumption is that projected results have a uniform distribution, in which historical observations equally fall within different bins constructed by projected urbanization values. For a given target region and year, we first ordered its 9 projected values to construct 10 bins. We then identified the bin that contained the observed record. We repeated this process for all years and target regions to construct the final histogram. Notably, the boundaries of each bin were determined repeatedly based on the given target region and year. This rank histograms allow the identification of any systematic under- or over-projections.

Theoretically, the validation analysis is more meaningful for a single region when only its own historical records from earlier periods are used to project changes in later historical periods. In this aspect, there are only sufficient historical data from the U.S. to carry out such a validation analysis. We extracted U.S. historical urbanization and population records for the1900-1980 time frame to project urbanization trends of U.S. states from 1985 to 2010. Then, we compared projected values to actual records over the period 1985-2010 to examine the performance of the model. Even for the U.S., we did not have sufficient data to carry out the projection at two stages and only conducted the projection for 30 years. Therefore, we had 3 projected pathways instead of 9, and could only construct 4 bins for this specific rank histogram analysis.

4. Results

4.1. Projection

Our results indicate that the improved approach outperforms the previous model in several ways. First, it generates urbanization projections at both national and subnational levels. Second, it produces more stable and plausible outcomes in the situations of stagnant or declining urbanization. Third, it generates a wider range of projections than the previous model, due to a more thorough search through historical urbanization records of reference regions),.

The improved model is versatile in generating urbanization projections for both national and subnational levels. In this paper, the presented national projections are based on using only national historical observations while both national and subnational historical records combined underlie the projections at the subnational level. For example, Figure 3 shows the U.S. state Alabama projections under the 9 scenarios based on urbanization and population datasets that contain both national and subnational observations (Figure 3a) alongside those resulting from using only subnational records (Figure 3b). While both suggest a wide range of future urbanization trends for Alabama, the projections based on the combined national and subnational datasets present slightly larger variations and uncertainties than the ones based only on the subnational dataset.

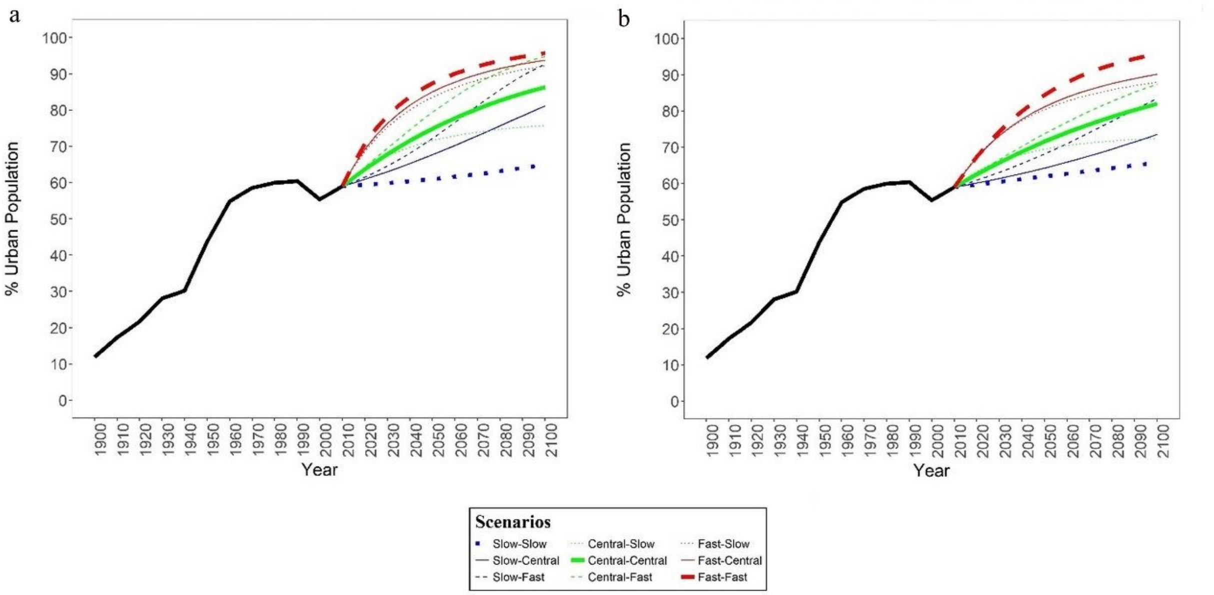

Figure 3. Urbanization projections for Alabama based on (a) combined national and subnational and (b) only subnational historical records.

The improved model is more robust and flexible and can make urbanization projections for regions with various time frames of their historical observations and those of their reference regions. For example, the model generates urbanization projections for the period 2010-2100 for Louisiana, a state in the U.S., Assam State of India, and Chongqing Province of China with their urbanization records being first available in 1900, 1950, 1970 and 2000, respectively, using reference regions with distinct historical time frames (Figure 4).

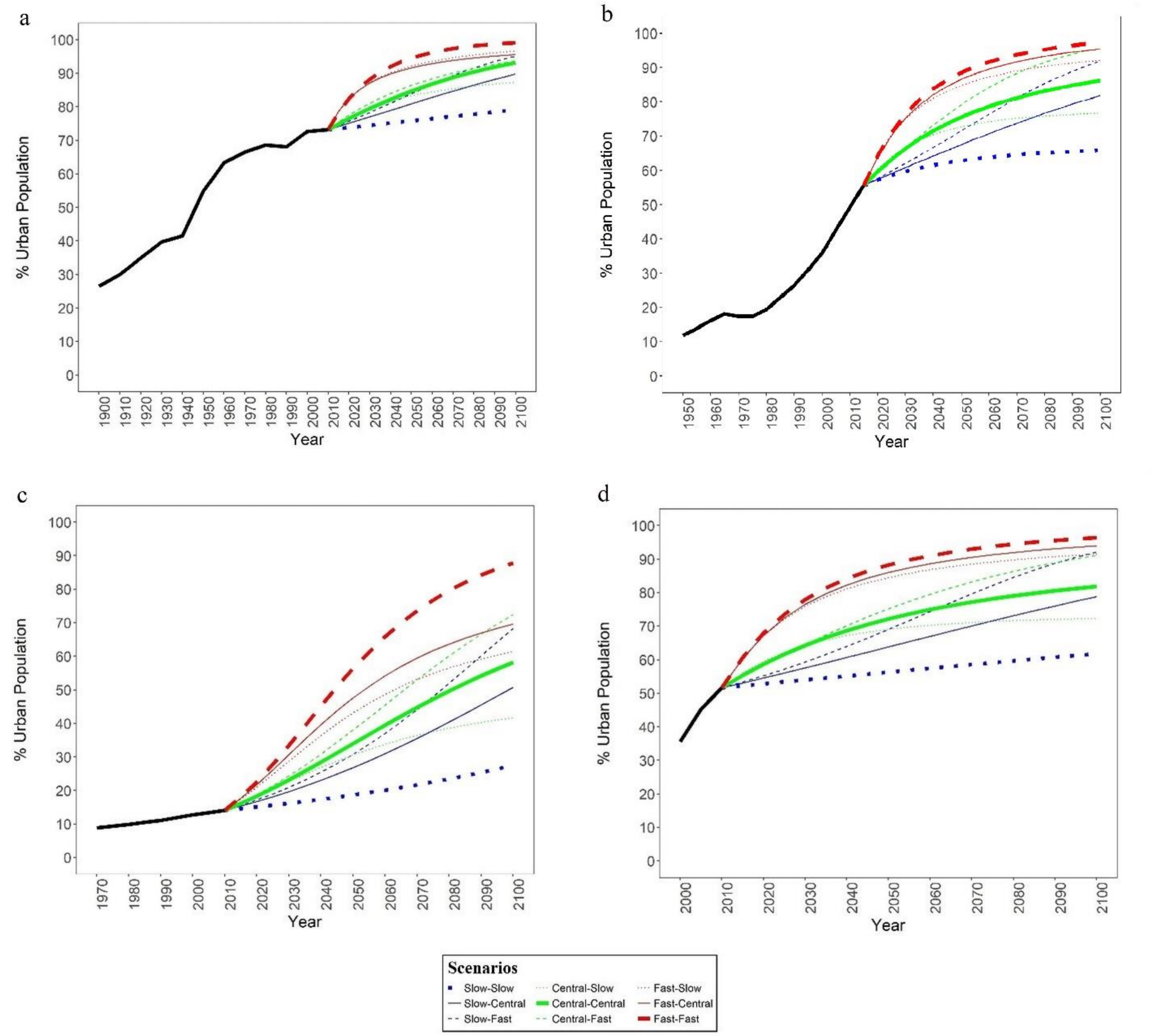

Figure 4. Urbanization Projections for (a) Louisiana, (b) China, (c) Assam, and (d) Chongqing with distinct time series of historical records.

The improved model is more capable to reflect the possible paths of urbanization stagnation and decline, a phenomenon observed in many parts of the world especially in the more recent decades. A large number of countries in both developed and developing regions experienced rapid urbanization in the 1950s to the 1970s, followed by a considerable reduction in the urbanization growth and even counter-urbanization afterward. The historical urbanization experiences is embedded in the datasets that are used to select references for the urbanization projections. In the previous model and projections, the slower urbanization and even counter-urbanization of the later periods is underrepresented as projection references. Because the previous model searches historical records of regions and selects as the references based on their urbanization experiences prior to the first year when the regions just reached the target urbanization levels. Therefore, the early periods of the reference regions are always more likely selected than the later periods if the regions experienced a rapid urbanization first and an urban decline later even though they had similar urbanization levels at both time periods. As a result, very few cases of reducing urban growth in the later periods are selected for the slow reference group.

The improved model implements a more thorough search through all years of the reference regions and defines the reference groups based on the most relevant and similar urbanization experiences to the region under study. The projection results reveal that the improved model is significantly more capable in capturing the trends of slow urbanization pathways. Figure 5 shows the urbanization projections of Armenia and Vermont, as two national and subnational examples, respectively, by using the previous and improved models. The projection results using the previous model suggest increasing urbanization under the slow-slow scenario for both Armenia (Figure 5a) and Vermont (Figure 5c), even though both recently experienced urbanization decline (Armenia) or stagnation (Vermont). The improved model produces the results (Figure 5b and 5d) that indicate the possible slow-slow pathways more plausibly extend the declining and stagnant historical patterns of these two regions.

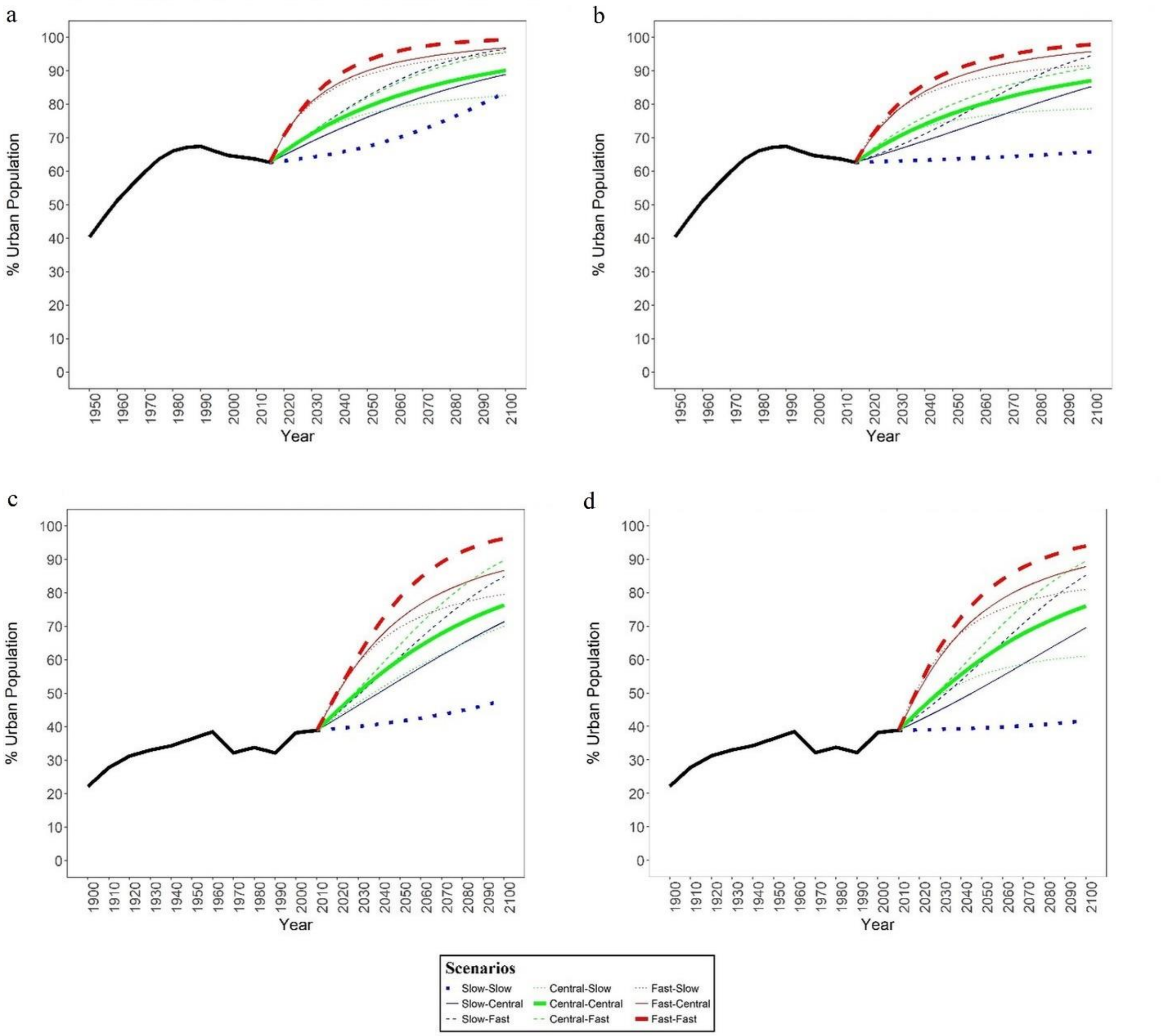

Figure 5. Urbanization projections for (a) Armenia from the previous model, (b) Armenia from the improved model, (c) Vermont from the previous model, and (d) Vermont from the improved model.

The improved capability of the new model in capturing possible trends of urbanization stagnation and decline directly leads to a wider range of urbanization projections. This enhancement has two benefits. First, it better covers the large uncertainty of future urbanization growth and offers a more informed insight into different conceivable urbanization trajectories in the future. Second, it lays out a more distinct set of urbanization pathways that can be associated with different socioeconomic conditions. This will be particularly useful for developing long-term alternative scenarios for assessing socioeconomic and environmental impacts of urbanization across different scales. Figure 6, shows projected urbanization trajectories according to the 9 scenarios for both India (national level) and Vermont (subnational level) based on the previous and improved models, as examples. It is evident that the improved model produces a wider range of projections for both target regions, mainly due to the improved capability of the model in projecting the slow-slow urbanization trajectories.

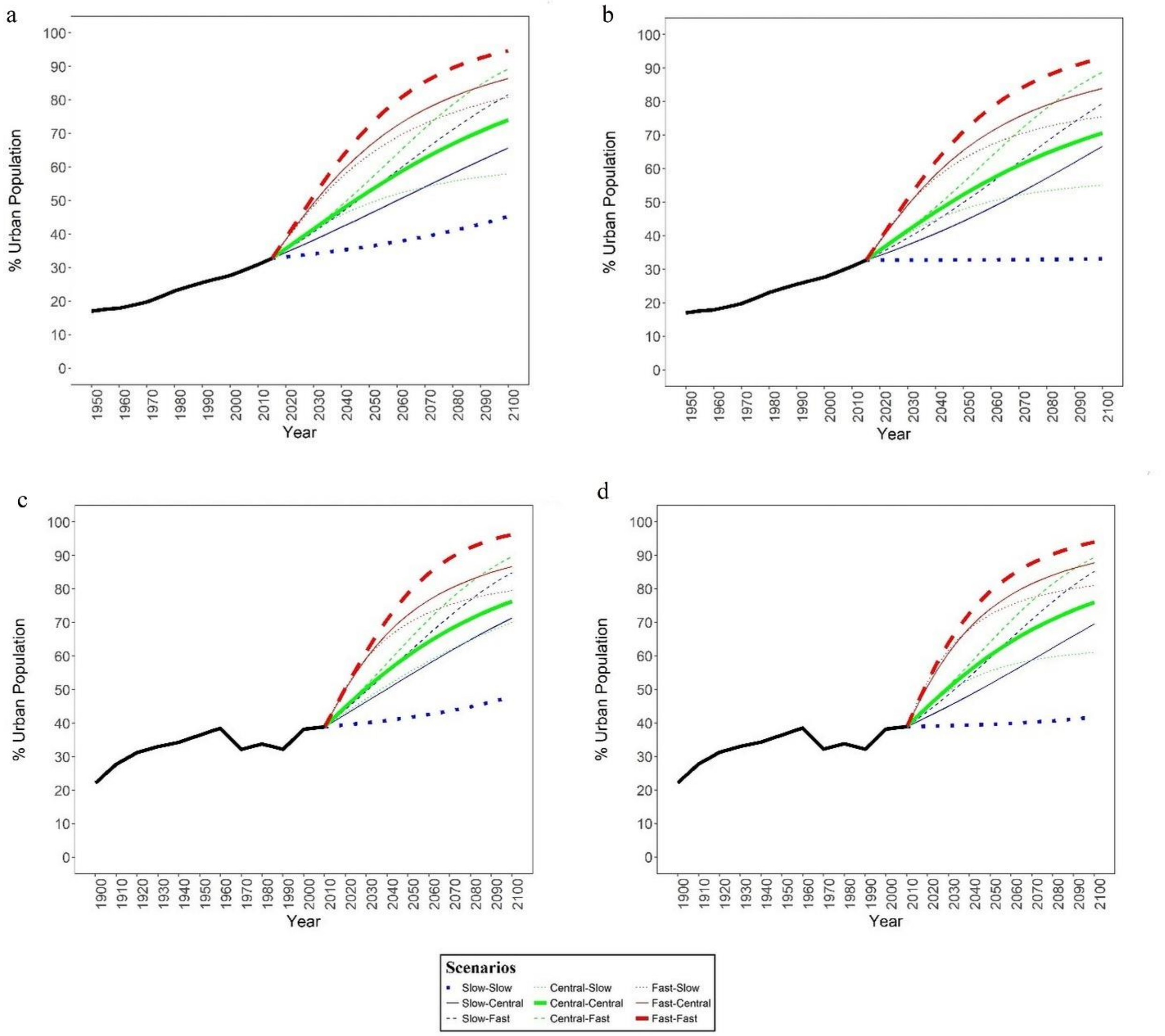

Figure 6. Urbanization projections for (a) India from the previous model, (b) India from the improved model, (c) Vermont from the previous model, and (d) Vermont from the improved model.

4.2. Validation

To examine whether the model produces robust and unbiased urbanization projections, we conducted validation analysis by comparing projection results over historical period to the observed urbanization records. The projections started from different years of the past decades (validation year) to evaluate the model results across various validation periods. Figure 7 demonstrates a few examples of our validation analysis from Algeria as a country, and 3 subnational regions from the U.S. (Colorado), India (Chhattisgarh) and China (Chongqing).

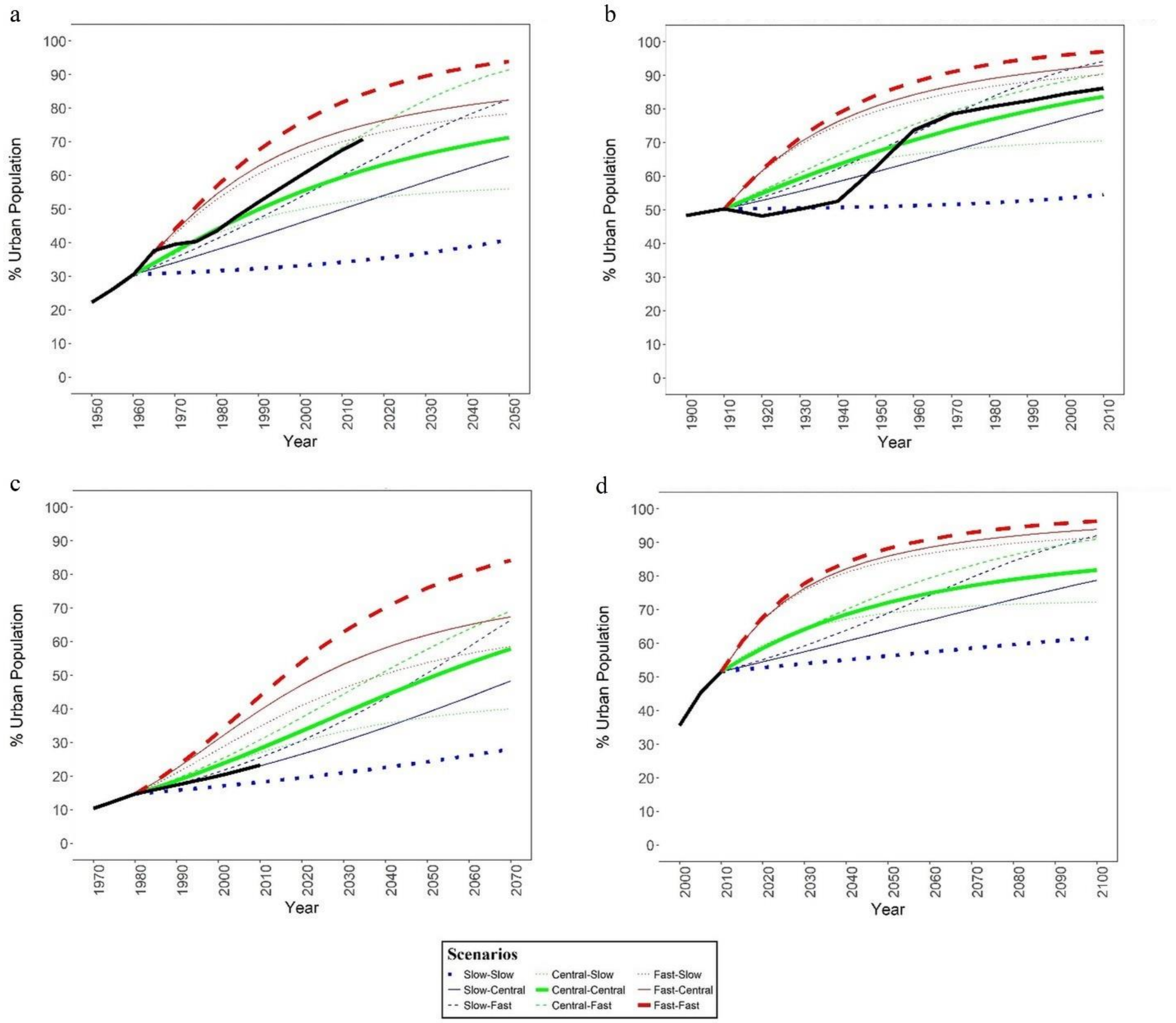

Figure 7. Validation plots of (a) Algeria, (b) Colorado, (c) Chhattisgarh, and (d) Chongqing with different validation years.

The extent of the validation analysis that can be done depends on the temporal extension of historical records. For example, while we can do validation analysis for both short-term (first 30 years) and long-term (beyond 30 years) periods for Algeria and Colorado, we can only do short-term analysis for Chhattisgarh and none for Chongqing because of lacking historical records.

As a step of validation analysis, we collected the number of historical records of a target region falling within the ranges of its projected results. We carried out the analysis for each target region as of their validation years. The number of comparison years is different for each region: 11 years per world country (from1965 to 2015 with a 5-year interval), 20 per U.S. state (from 1915 to 2010 with a 5-year interval), mostly 10 per China province (from 1965 to 2010 with a 5-year interval) and mostly 6 per India state (from 1985 to 2010 with a 5-year interval).

In the projection of historical urbanization trends for the validation analysis of Figure 7, we selected reference regions for national level projections based on the datasets that include all nations and for subnational level projections based on the combined datasets including both nations and subnational regions. The idea behind our validation analysis is that projection results over the validation period to some extent should cover the "normal" plausible range of urbanization trajectories based on the experiences of all relevant countries and subnational regions. One would expect that an effective model will lead to projections that well mimic urbanization histories of target regions. Inevitably, there could be regions with unusual (or "abnormal") urbanization experiences during certain historical periods, particularly some subnational regions, which the model could not address. On the other hand, if its projection results do not represent urbanization histories of majority of target regions, the projection model would be considered ineffective.

We used the quantities in Table 3 as a metric to evaluate if our model led to projection ranges that covered historical records. We identified regions with more than 20% of their historical records falling outside resulting projected ranges as "abnormal" or the failed cases that the model was unable to address. Table 3 summarizes the outcomes of this analysis.

Table 3. Summary of the number of regions whose historical records fall within projected ranges.

		# All Records	#Select Records[1]	#Proportion
National		220	162	0.74
Subnational	USA	50	44	0.88
	China	30	5	0.17
	India	34	19	0.56

[1] Those whose proportion of years falling within ranges to all years is higher than 0.8.

Table 3 shows that 74% of global nations have more than 80% of their historical observations falling inside their ranges of urbanization projections, suggesting that the urbanization projection model sufficiently represents historical urbanization trends at the national level. At the subnational level, the proportion is higher (88%) for U.S. states. The majority (56%) of Indian states also meet the 80% threshold. However, there are only 17% of Chinese provinces with more than 80% of historical observations falling inside ranges of urbanization projections. Therefore, urbanization projections based on experiences of other countries and subnational regions do not represent well the urbanization histories of most of the Chinese provinces for the period of 1960-2010. This is because China had experienced very different urbanization than other countries in the world during the recent decades (Ma 2002; Jiang and Kuijsten 2001). The country rapidly urbanized in the late 1950s during the Great Leap Forward campaign, followed by a dramatic decline in urbanization caused by the 3-year famine in 1959-1961. It sustained very low levels of urban growth for the next two decades under an antiurbanism policy accompanied by centralized planning and city-based industrialization. Urbanization stopped in China until the economic reform in the early 1980s, before taking off with rapid economic growth and industrialization in the early 1990s. The recent history of dramatic changes and large variations in urban growth places China and its provinces in a very different position from other countries and regions. As a result, projected urbanization trajectories of Chinese provinces in the past decades, based on the experiences of other countries and regions, cannot reflect the peculiar historical records in China in the 1960s (Figure 8). It is noteworthy, however, that the urbanization projections by the model does not aim to forecast the exact urbanization levels in the short run, but rather provide the plausible urbanization trends in the long or medium term.

Our projection results suggest that the model meet the requirement quite well. Even in the Chinese case, the urbanization records of most Chinese provinces started to fall within the range of projected values from 1990 - 30-year after validation year (1960).

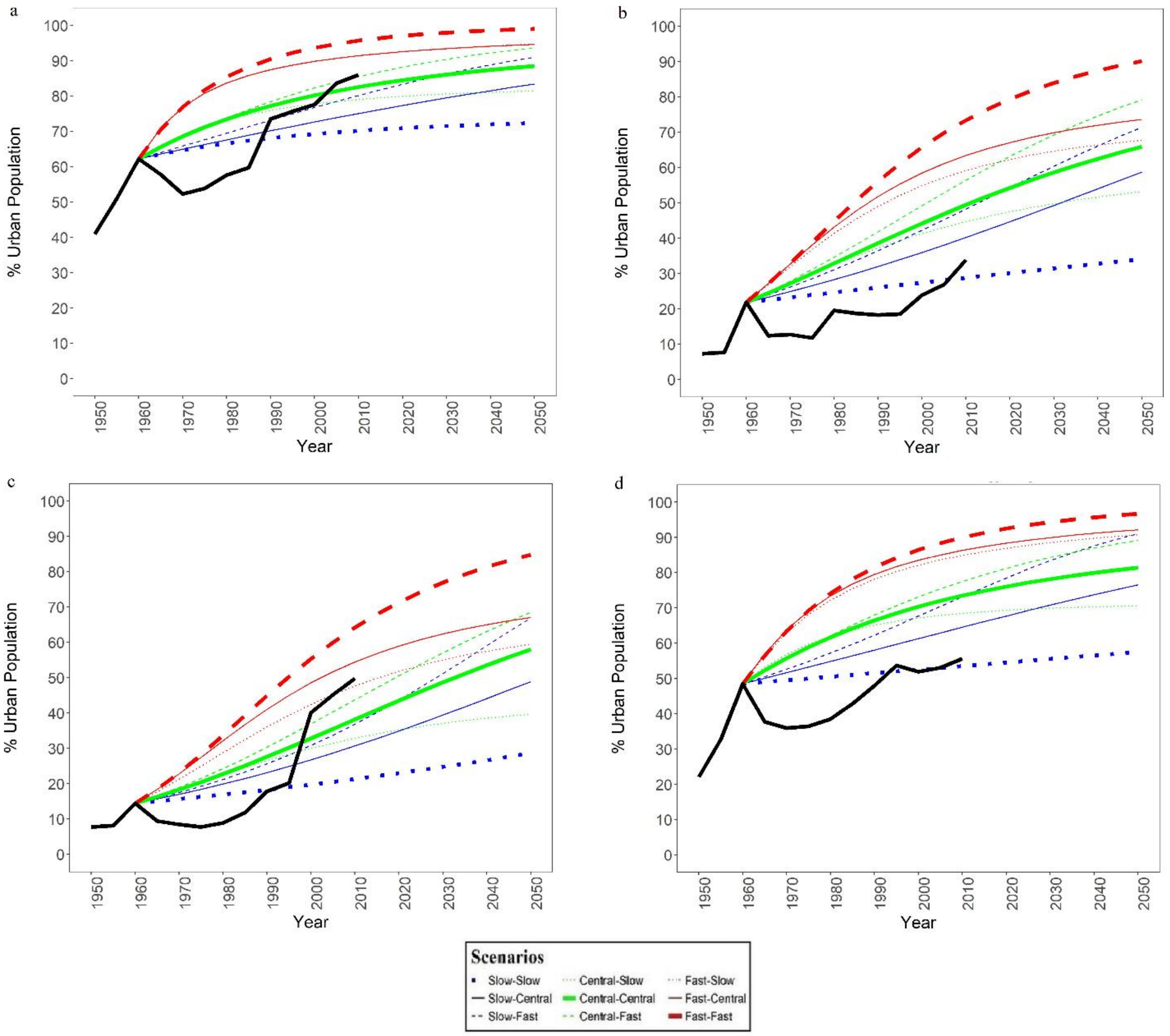

Figure 8. The urbanization validation plots for (a) Beijing, (b) Guizhou, (c) Hainan, and (d) Heilongjiang with the validation year being 1960.

The second validation analysis was to compare historical urbanization trajectories to the 9 projected urbanization pathways by the model and examine how the projected pathways reflect actual urbanization experiences. Table 4 summarizes the number of times that each pathway mostly resembles historical trajectories of regions. For all world countries, U.S. states, and Chinese provinces, the comparisons were made to 9 projected urbanization pathways. For India, there were only 6 comparison years, prohibiting a two-stage evaluation as discussed in Section 3.3. That is why there were only 3 projected pathways for the country to compare. It is noteworthy that for the majority of Chinese provinces, the slow-fast is their most resembling pathway, which reflects the above-mentioned urbanization history of China. The slow pathway is the most dominant representative scenario among India states. This is consistent with the research findings that India had one of the slowest urbanization growth rates during the recent decades among all global regions (Kundu 2011; Bhagat 2018). On the contrary to China and India, in which the slow-fast and slow scenarios dominate, respectively, there is no dominant urbanization pathway for U.S. states and world countries.

Table 4. Counts of representative scenarios for both national and subnational levels.

Scenario	National-Level	Subnational-Level		
		USA	China	India
Slow-Slow	38	6	5	N/A
Slow-Central	22	10	4	N/A
Slow-Fast	15	8	16	N/A
Central-Slow	34	6	0	N/A
Central-Central	27	9	0	N/A
Central-Fast	32	7	2	N/A
Fast-Slow	23	1	0	N/A
Fast-Central	8	0	1	N/A
Fast-Fast	21	3	2	N/A
Slow	-	-	-	24
Central	-	-	-	5
Fast	-	-	-	5
Total	220	50	30	34

Adopting the rank histogram method, we conducted the third validation analysis to test the overall performance of the projection model and examine whether the model generates reasonably unbiased results. The rank histogram analysis generates a set of histogram bars (Figure 9) that show the number of times historical urbanization records of target regions across all comparison years fall within bins constructed by their projected values. We calculated 95% confidence intervals around the representative value of each histogram bar if the distribution was uniform. The boundaries of the confidence interval (hi) were calculated using the equation below:

$$h_i = N \times \frac{1}{i} \pm \sqrt{N \times \frac{1}{i} \times (1 - \frac{1}{i})} \times 1.96$$

Where N is the total number of historical observations and i is the number of bins constructed by projected values.

Figure 9a demonstrates that projection results by the model for world countries compared to historical observations, are generally unbiased and evenly distributed. The analysis provides us with good confidence that the model works well at the national level.

At the subnational level, Figure 9b shows that projections for U.S. states are slightly skewed to the lower bins, meaning that the model over-projected the urbanization growth of U.S. states comparing to their historical records. Figure 9c and 9d display the rank histogram analysis results for China and India, respectively. This validation analysis for China and India shows that their urbanization projections seriously over-project their historical trends, i.e. a high concentration of historical records fall below the slow-slow and slow-central scenarios. This is mainly caused by the dramatic urbanization decline after 1960 in China and exceptionally low urban growth in India, especially in the 1980s and 1990s, as discussed above.

The rank histogram analysis results in Figure 9 are based on using combined national and subnational regions as references. Theoretically, a better evaluation should be based on the projections using own historical data as references. While we cannot carry out such an analysis for China and India due to lack of data, the U.S. historical data is sufficient to be used for making projections for U.S. states. Thus, we derived another set of projections using only the historical data of U.S. states as references. Consequently, the rank histogram analysis suggests improved results, manifested by a more evenly distributed histogram (Figure 10).

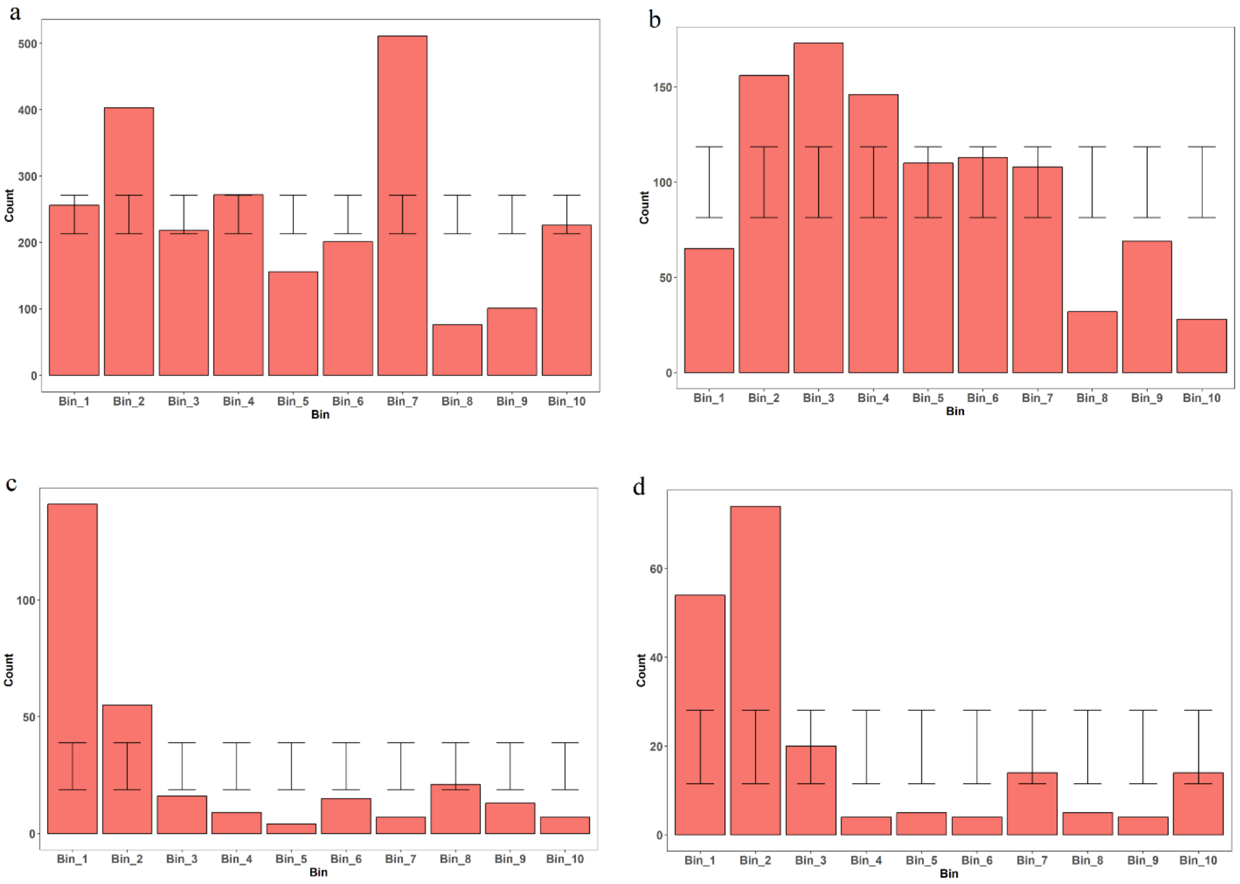

Figure 9. Rank histograms corresponding to the observed and projected urbanization levels of (a) world countries, 1965-2015; (b) U.S. states, 1915-2010; (c) China provinces, 1965-2010; and (d) India states 1985-2010. Projections for world countries (a) are based on the national level data; subnational projections (b, c, and d) are based on the combined national and subnational datasets

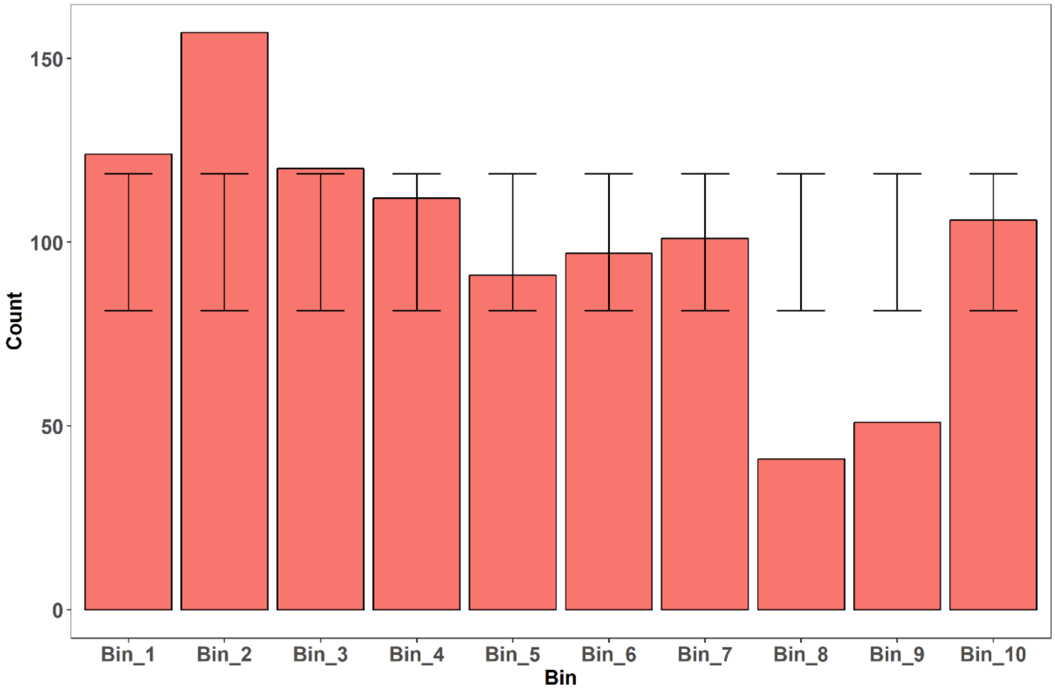

Figure 10. Rank histogram corresponding to the observed and projected historical urbanization levels of U.S. states, 1915-2010. Projections are based on the historical data of U.S. states.

We also tested another approach by using the historical data of U.S. states over the period 1900-1980 to project their urbanization trends for the period 1985-2010. We did this to ensure that there was no overlap between the validation and historical periods. In this case, 4 bins were constructed because only 3 scenarios cover the time frame (Figure 11). The rank histogram analysis leads to a similar impression to the one based on the combined national and subnational datasets (Figure 9b), indicating the model's tendency to over-projecting. This is not unexpected because the urbanization in the U.S. slowed down from the 1980s (Long and DeAre 1983; Boustan et al 2013), similar to the experiences of many other developed and developing countries. The projected results after 1980 based on the experiences of previous decades can easily lead to an over-projection.

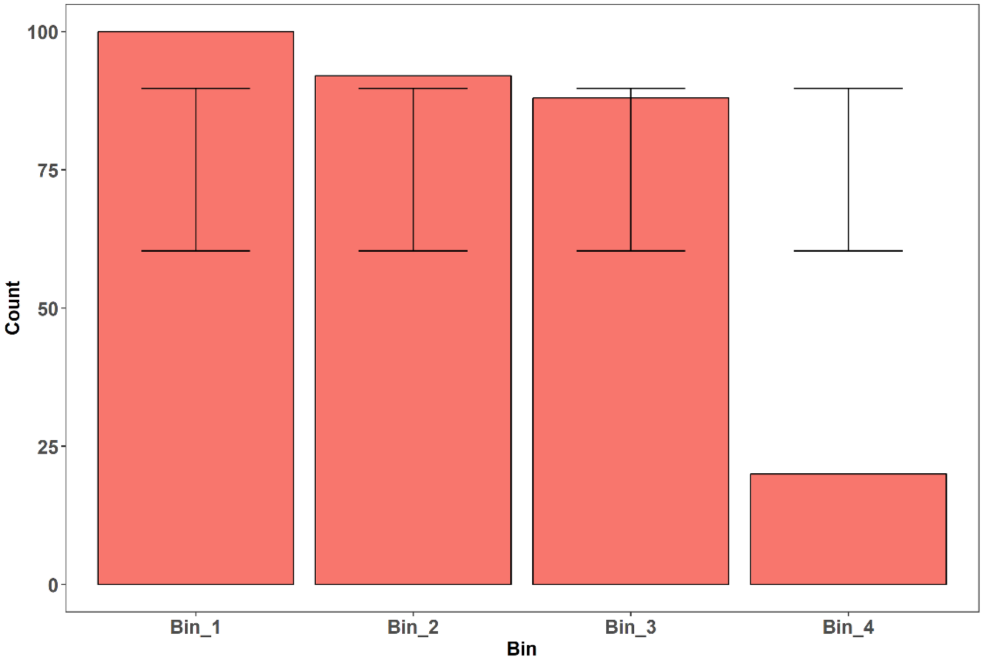

Figure 11. Rank histogram corresponding to the observed and projected historical urbanization levels of U.S. states, 1985-2010. Projections are based on the historical data of U.S. states from 1900 to 1980.

Given the peculiar history of urbanization in China, especially the dramatic changes before and after 1960, we changed the validation year from 1960 to 1970 and made another set of urbanization projections for China provinces over the period 1975-2010, when the spike pattern (rapid urbanization growth in late 1950s and sharp decline in the early 1960s) had already passed and the trend settled. By doing so, the projected trajectories better resemble the historical patterns as shown in Figure 12. Accordingly, the proportion of projection trajectories with more than 80% of observations falling inside their resulting ranges (Table 3) increased from 0.17 to 0.66, indicating the majority of historical urbanization trajectories of China provinces during the period 1975-2010 are well covered by the model. The validation analysis using the rank histogram diagnostic tool also proves that the projection generates more uniform and unbiased results that encompass the historical urbanization trends of China provinces after 1970 (Figure 13).

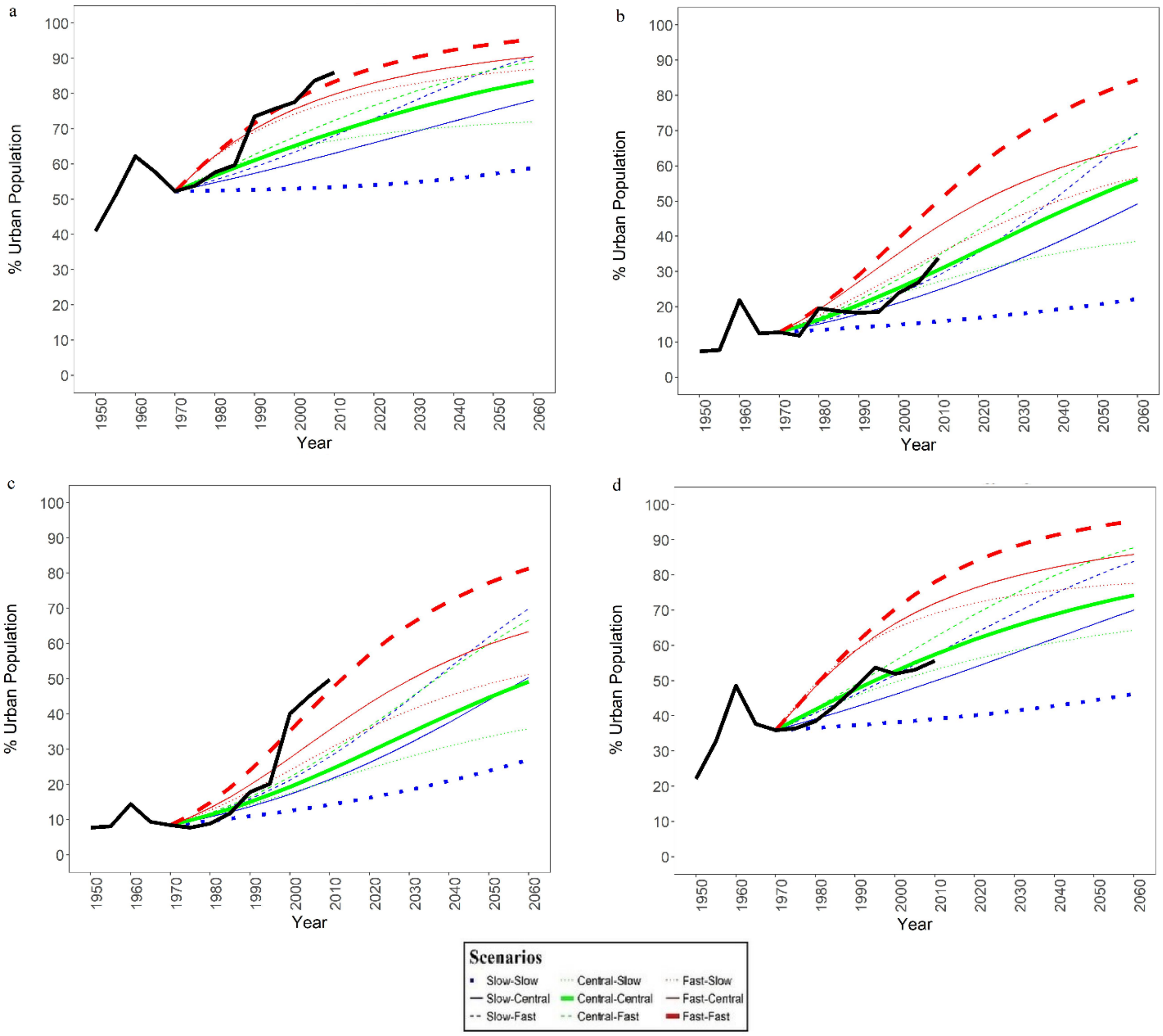

Figure 12. The urbanization validation plots for (a) Beijing, (b) Guizhou, (c) Hainan, and (d) Heilongjiang with the validation year being 1970.

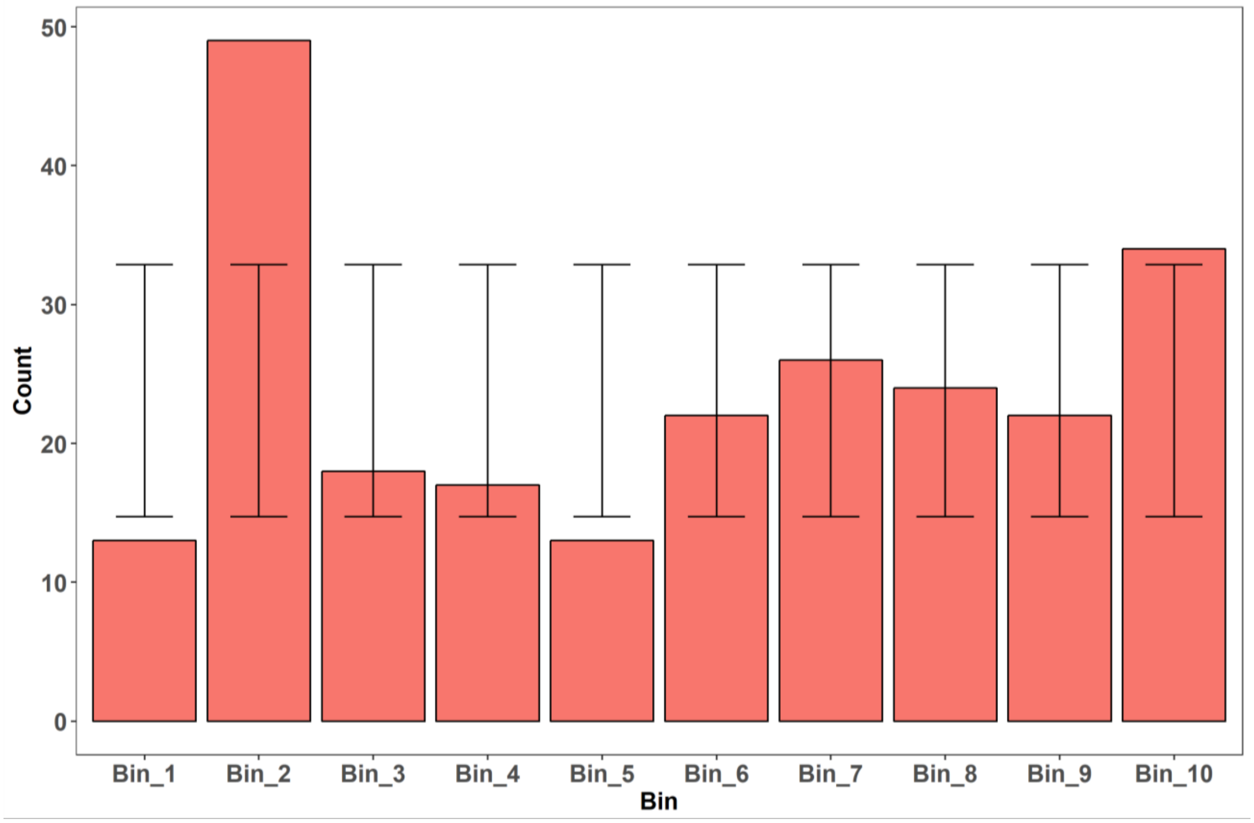

Figure 13. Rank histogram corresponding to the observed and projected urbanization levels of China provinces during 1975-2010. Projections are based on the combined national and subnational datasets.

5. Discussions and Conclusions

Urbanization projection is one of the key components of the CDM demographic tools developed at the Integrated Assessment Modeling Group of NCAR. In this paper, we reported on the new developments in the improved CDM-Urbanization Projection Model and its use for projecting subnational urbanization trends for the provinces of China, and the states of India and the U.S.

The improved model presents versatility in generating urbanization projections at both national and subnational levels and using different combinations of input data. This is important because our model adopts a data-driven approach whose performance for each target region relies on the number and similarity of its selected reference regions. The improved model currently takes the historical population and urbanization observations of all countries as well as historical records of U.S. and India states and China provinces. It employs the national level dataset for projecting urbanization of countries and both national and subnational datasets for projecting urbanization of states or provinces. It also has the capability to take subnational historical records of other countries such as Russia and Brazil provided that they become available and standardized. An inclusive dataset for the model enhances its ability to pick more similar reference regions for each target region and scenario, which in turn leads to the construction of representative regression models and consequently more plausible urbanization projections.

We used the extensive dataset that combines the national and subnational urbanization records of other countries to project subnational urbanization levels, especially for China and India, because there were insufficient historical records for either country to be used separately. For projecting the U.S. subnational level urbanization, on the other hand, we did not encounter this issue and could have used only U.S. historical records. Our model has the capability to use either the U.S. dataset, the U.S. + China + India combinatory dataset or the U.S. + China + India + national level combinatory dataset for projecting urbanization levels of U.S. states. The decision on which dataset should be used is yet to be further contemplated. One way would be to devise quantitative metrics to determine which dataset leads to pathways that both show the highest similarity to historical patterns and model uncertainty inclusively.

Furthermore, the improved model performs a more thorough search through historical urbanization patterns of reference regions and extracts the part from each reference region that mostly resembles the recent urbanization pattern of the target region. This is an important advantage over the previous model, in which it defines reference groups based on urbanization growth rates of reference regions prior to the first year when they reached within the target urbanization range. This more informed pattern extraction for the urbanization projection is one of the contributing factors that the improved model results in urbanization projections with wider ranges and more plausible performance with respect to urbanization decline or stagnation (Figures 5 and 6).

The current model is pure demographic and data-driven and does not explicitly account for any exogenous socioeconomic factors for projection. We opted not to use such factors in order to maintain the generalizability of the model. However, incorporating additional factors specific to a target region might enhance its urbanization projection overall. Moreover, the empirical data used for selecting references for projections comes from regions which may vary in defining urban. For instance, the dataset of UN Urbanization Prospects is based on the national definition of each country or region although numerous efforts are invested to achieve historically consistent statistics on national urbanization levels. The urban definition is a difficult problem that impacts a wide range of demographic and spatial studies. The discrepancy in definitions is a likely source of uncertainty in outcomes given that the method attempts to “learn” from the experience of other regions in projecting possible urban outcomes.

Future work will explore ways to map the 9 resulting pathways to distinct urbanization narratives for developing urbanization projections under the extended SSP scenarios. This will be done to better adopt the socioeconomic and environmental context of specific subnational regions.

References

Bhagat, R.B., 2018: A turnaround in India’s urbanization. *Asia-Pacific Population Journal*, Vol. 27 (2): 23-39.

Boustan, L. P., D. Bunten, and O. Hearey 2013: Urbanization in the United States, 1800-2000. *National Bureau of Economic Research Working Paper 19041*. Cambridge.

Ellis, Peter, and Mark Roberts. 2016. Leveraging Urbanization in South Asia: Managing Spatial Transformation for Prosperity and Livability. *South Asia Development Matters*. Washington, DC: World Bank. doi: 10.1596/978-1-4648-0662-9.

Hunter, L.M. and O’Neill, B.C., 2014. Enhancing engagement between the population, environment, and climate research communities: The shared socio-economic pathway process. *Population and Environment*, 35 (3), 231–242.

Jiang, L., 2014: Internal consistency of demographic assumptions in the shared socioeconomic pathways. *Population and Environment*, 35 (3): 261-285.

Jiang, L. and A. Kuijsten, 2001: Regional disparities of urbanization levels in China. *Chinese Journal of Population Sciences,* No. 1: 45-51.

Jiang, L. and O’Neill, B.C., 2015. Global urbanization projections for the Shared Socioeconomic Pathways. *Global Environmental Change*, 42, 193–199.

Jones, B. and O’Neill, B.C., 2016. Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters,* 11 (8).

Jones, B., O’Neill, B.C., Mcdaniel, L., Mcginnis, S., Mearns, L.O., and Tebaldi, C., 2015. Future population exposure to US heat extremes. *Nature Climate Change*, 5 (7), 652–655.

Kundu, Amitabh, 2011: Trends and processes of urbanisation in India. *Urbanizaton and Emerging Population Issues* - 6. IIED and UNFPA.

Long, L. and D. DeAre, 1983: The slowing of urbanizaiton in the U.S., *Scientific America,* Vol. 249 (1): 33-41.

Ma, Laurence, 2002: Urban Transformation in China, 1949-2000. *Environment and Planning A:*

Economy and Space, vol. 34 (9): 1545-1569.

O'Neill, B.C., Kriegler, E., Ebi, K.L., Kemp-Benedict, E., Riahi, K., Rothman, D.S., van Ruijven, B.J., van Vuuren, D.P., Birkmann, J., Kok, K., Levy, M., and Solecki, W., 2015. The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, 42, 169–180.

United Nations, Department of Economic and Social Affairs, Population Division, 2014: *World Urbanization Prospects: The 2014 Revision*, Methodology Working Paper No. ESA/P/WP.238.

United Nations, Department of Economic and Social Affairs, Population Division, 2017: World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP/248.