Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessing the incremental value of VIIRS nighttime lights for prefecture-level GDP growth in China

Bole Tan
Chinese Academy of Agricultural Sciences
Xiaoyu Jiang
Chinese Academy of Agricultural Sciences
Xin Dai
Chinese Academy of Agricultural Sciences
Zhong Li
lizhong@caas.cn
Chinese Academy of Agricultural Sciences
Yangbo Liu
Harbin Engineering University

Article

Keywords: Earth observation, nighttime lights, VIIRS, socioeconomic measurement, temporal transfer, validation

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11383504/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Assessing the incremental value of VIIRS nighttime lights for prefecture-level GDP growth in China

Article

Bole Tan[1]†, Xiaoyu Jiang[1]†, Xin Dai[1], Zhong Li[1]*, Yangbo Liu[2]*

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China

† Bole Tan and Xiaoyu Jiang contributed equally to this work.

*Corresponding authors: Zhong Li (lizhong@caas.cn) and Yangbo Liu (liuyangbo@hrbeu.edu.cn).

ORCID iDs: Bole Tan, https://orcid.org/0009-0004-5270-2933; Xiaoyu Jiang, https://orcid.org/0009-0006-2192-7886; Zhong Li, https://orcid.org/0000-0002-5475-9560

Abstract

Earth-observation nighttime lights are widely used to map subnational economic activity, yet whether strong cross-sectional fit translates into reliable measurement of annual local growth remains uncertain. Here we develop a forward, loss-based validation framework across 296 Chinese prefectures (2014–2024) using two VIIRS-derived series—EOG Annual VNL (with its continuation) and NASA Black Marble VNP46A4—to separate within-year spatial representation, stand-alone forward temporal transfer, and conditional incremental value beyond four prior-year GDP-history benchmarks. While log light levels strongly track nominal GDP ($R^2 = 0.730 - 0.831$), annual light changes align weakly with certified real GDP growth ($R^2 = 0.0005 - 0.247$). In pre-2020 forward evaluations, stand-alone light calibrations generally underperform GDP-history benchmarks. During 2016–2019, all eight pooled conditional incremental-skill estimates are negative (-0.0362 to -0.0016), with province-block mean-loss confidence intervals spanning zero, indicating no detectable positive gain from adding light to historical GDP. Conversely, conditional point estimates are positive across all benchmarks during the 2020–2021 stress period. In the 2022–2024 extension on common 296-prefecture support, pooled conditional skill is positive for the EOG-derived series (0.0319–0.1097) but negative for Black Marble (-0.2240 to -0.1127), a divergence concentrated in 2022 and robust across verified GDP-source definitions. These findings demonstrate that socioeconomic proxy validation must be tailored to the target estimand, conditioning information set, and temporal domain.

Keywords: Earth observation; nighttime lights; VIIRS; socioeconomic measurement; temporal transfer; validation

Introduction

Earth-observation data are increasingly used to derive socioeconomic information where conventional statistics are sparse, delayed, or spatially coarse. Nighttime lights are a prominent example because artificial illumination records persistent features of settlement, transport, commerce, production, and public infrastructure, and the resulting imagery has become a standard input to economic measurement [1–5]. Recent work uses nighttime lights to construct subnational economic surfaces and higher-frequency output estimates [6, 7]. These applications make validation consequential beyond map production itself. A spatial layer can reproduce the geography of economic activity while still providing little information about annual local change, because persistent infrastructure and settlement patterns can support a stable cross-sectional ranking even when short-run changes in radiance and output diverge.

Existing evidence gives strong reasons not to treat nighttime-light validity as a single property. Performance varies across products and locations [8], small-area relationships are heterogeneous and aggregation-sensitive [9], and processing choices can change substantive conclusions [10]. County-level studies in the United States and China find much stronger performance for cross-sectional differences than

for within-place changes [11, 12]. At national scale, nighttime-light indicators can improve GDP-growth forecasts, but the gains are heterogeneous across countries and specifications [13]. Signal harmonization can strengthen over-time relationships [14, 15], while recent subnational evidence continues to emphasize the contrast between levels and growth [16]. The relevant question is therefore not whether nighttime lights are generally valid or invalid, but whether a mapping validated for one measurement task remains informative when the target, time horizon, or information set changes.

This problem is geographical as well as temporal. Counties, prefectures, provinces, and broader regions are different spatial supports, and aggregation can alter observed relationships even when the underlying process is unchanged [17, 18]. Transfer beyond the empirical domain represented in a calibration sample is related to domain shift and area-of-applicability concerns in spatial prediction [19–21]. Recent work also shows that aggregation can change predictive elasticities [22] and that measurement error in nighttime lights can affect downstream inference [23]. We therefore distinguish spatial representation, temporal transfer, spatial dependence, and cross-series measurement in the validation design.

We examine these distinctions across 296 Chinese prefectures using EOG-derived annual VIIRS lights and NASA Black Marble VNP46A4 Collection 2 through 2024. The products share the VIIRS/Suomi-NPP sensor family and differ in processing. Calibration uses only preceding GDP outcomes and is applied to observed target-year light. Light-only estimates are compared with national, provincial, prefecture and most-recent GDP histories; conditional tests add light to intercept-recalibrated history. Timing is indexed to GDP outcome years, not real-time publication vintages. The primary 2015/2016–2019 domain, 2020–2021 stress period and 2022–2024 extension remain separate. The extension evaluates a fixed 296-prefecture universe, with differences in GDP-source composition and nighttime-light series assessed separately.

The contribution is a forward, loss-based validation framework rather than a new predictive estimator. It separates cross-sectional representation from forward annual-growth validity, stand-alone light performance from incremental value beyond explicit GDP-history benchmarks, and temporal portability from geographic support, GDP outcome-source validity, and cross-series variation. A proxy may fit levels well, transfer poorly across target years, or show different performance across the two VIIRS-derived series even on the same economic targets. Figure 1 summarizes this validation structure.

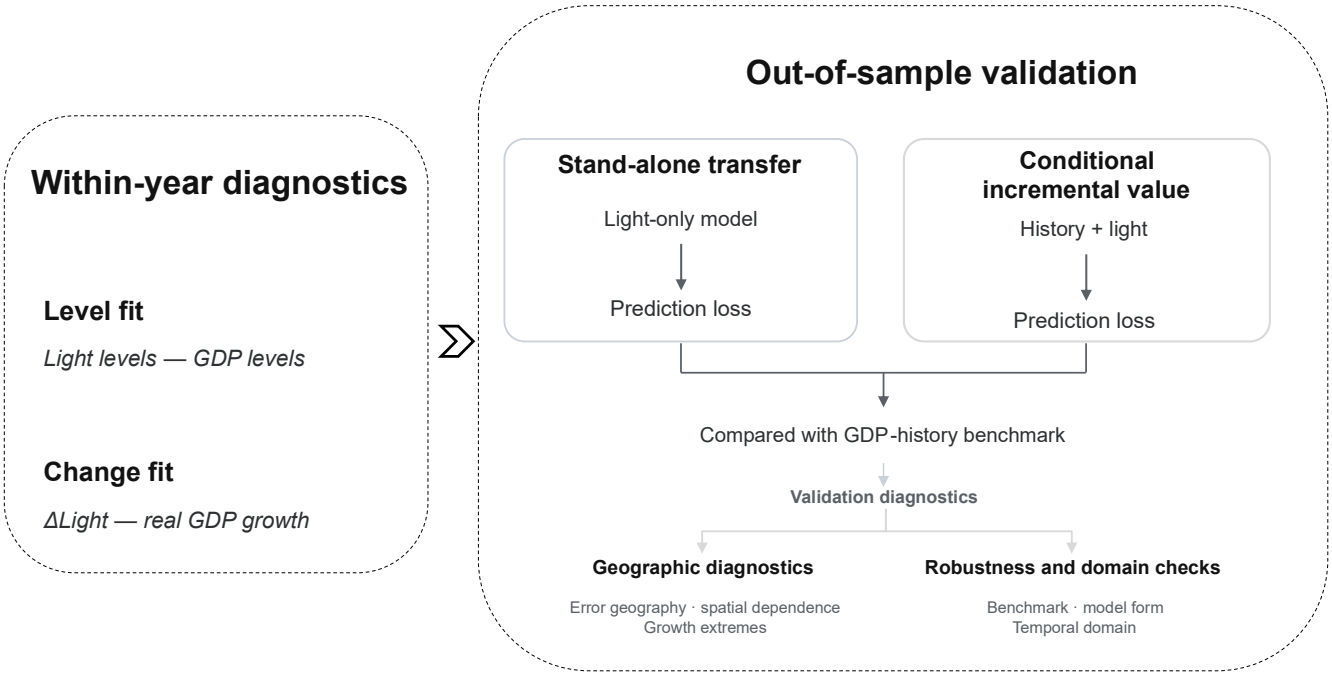


Figure 1. Task-matched validation framework for nighttime lights as a socioeconomic proxy. Within-year diagnostics evaluate spatial representation and annual-change alignment. Forward tests compare light-only mappings with prior-year GDP histories and then compare GDP history augmented with light against GDP history alone. Geographic, processing, sample, and temporal-domain diagnostics define the scope of these comparisons. Target-year light is observed and target-year GDP is excluded from calibration and benchmark construction. The GDP timing restriction is distinct from the full-period fixed raster support.

Results

Study universe and GDP outcome provenance

The study covers a fixed set of 296 prefectures across all 31 mainland provincial-level units. Historical GDP coverage is reported by year in Table 1. In the 2022–2024 extension, the priority-source GDP panel combines 482 independently retrieved official-source outcomes with 406 entries from a compiled workbook, yielding 888 prefecture-year observations. A verified-enhanced sensitivity panel replaces 108 compiled entries with inspected official values, leaving 298 compiled entries. Both panels evaluate the same 296 prefectures in each extension year (Fig. 2; Supplementary Table S50).

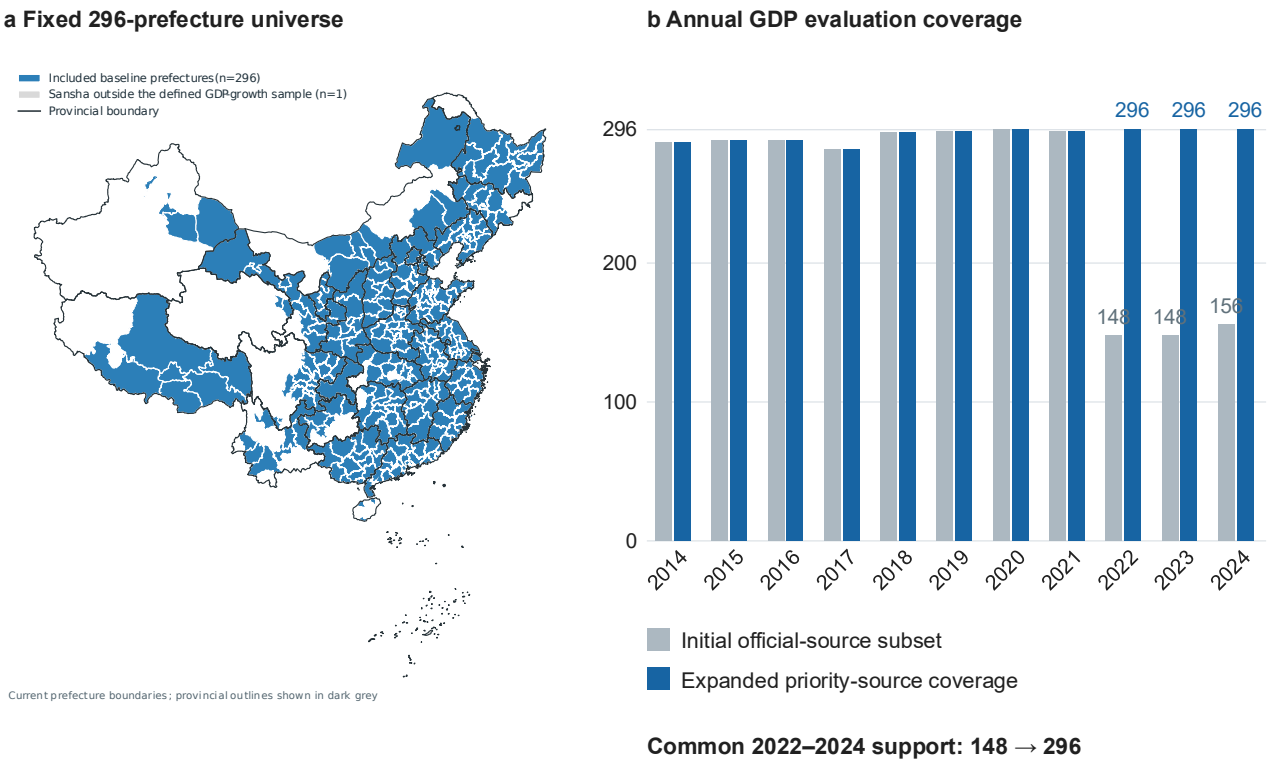


Figure 2. Fixed analysis universe and GDP evaluation coverage. (a) Current-boundary 296-prefecture universe, which is also the common 2022–2024 extension support. Blue identifies retained baseline units; Sansha is outside the defined GDP-growth sample. Provincial boundaries are unchanged. Areas outside the fixed 296-prefecture analysis universe are unfilled. (b) Initial official-source subset (452 observations) versus the expanded priority-source coverage: 2022–2024 coverage increases from 148, 148 and 156 to 296 in each year, and common support rises from 148 to 296. Historical counts are unchanged. Coverage refers to the defined 296-prefecture study sample, not all prefecture-equivalent jurisdictions in China.

The initial official-source subset comprised 452 prefecture-year observations (148, 148 and 156 in 2022–2024; 148 prefectures common to all three years). Adding 30 official Shaanxi observations increased this subset to 482 observations (158, 158 and 166; 158 common prefectures). The initial 148-prefecture subset differed from the remaining baseline units in several pre-extension economic and nighttime-light characteristics. The complete extension uses all 296 prefectures in the fixed analysis universe, providing common geographic evaluation support across both light series. The sample-selection diagnostics and their definitions, including standardized mean differences, are reported in Supplementary Tables S37 and S41–S42 and Supplementary Fig. S7.

Spatial representation

In the historical five-wave level diagnostics for 2014, 2015, 2016, 2018, and 2019, the bivariate R^2 between log EOG total light and log nominal GDP ranges from 0.730 to 0.779. Annual light change and certified real GDP growth have R^2 values from 0.0005 to 0.247. On the 283-prefecture common positive sample, raw level R^2 is 0.731–0.770 and remains 0.731–0.773 after residualizing both variables on prefecture land area. Black Marble shows the same broad contrast, with level R^2 of 0.775–0.831 and change R^2 of approximately 0.001–0.089 across the corresponding sensitivity grid. These historical representation diagnostics are retained as their original estimands; the extension below concerns forward annual-growth validation. Nominal levels and real growth are different economic concepts, so their contrast motivates rather than itself proves a temporal limitation.

Primary forward validation: stand-alone transfer

Primary stand-alone forward validation transfers each historical light–growth calibration to the next observed target-year signal. Most primary B0 skill estimates are negative despite strong spatial alignment (Fig. 3; Table 1). The displayed extension uses the retained priority-source GDP panel; alternative source definitions are reported in Supplementary Table S52A. Target-year light is observed, while target-year GDP is excluded from calibration.

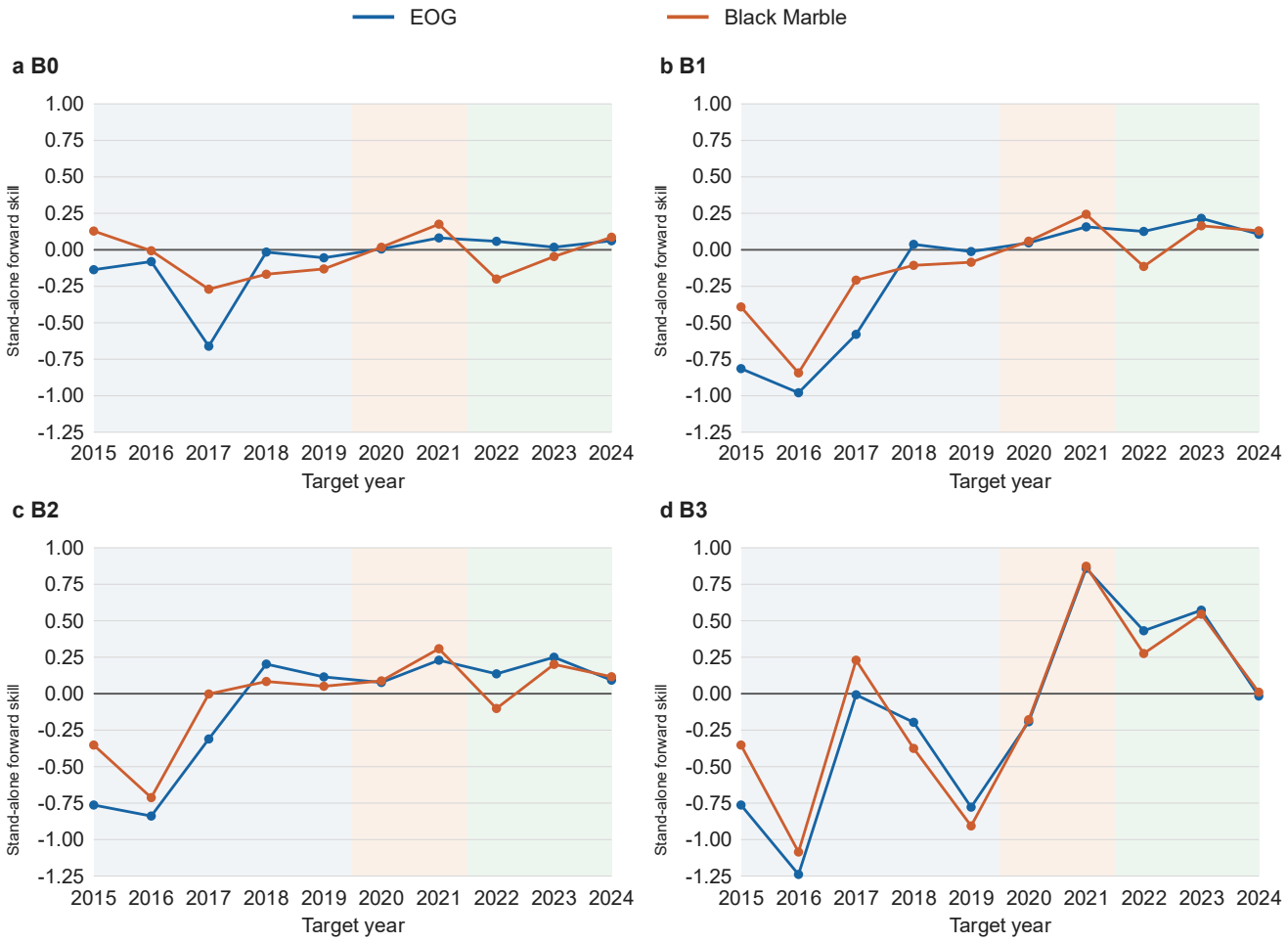


Figure 3. Stand-alone forward skill against GDP-history benchmarks B0–B3. B0 is the national training-period mean; B1, the province training-period mean; B2, the prefecture training-period mean; and B3, the most recent observed prefecture growth. Each annual cell compares transferred light-only predictions with its corresponding history-only benchmark; positive skill favors light. Extension estimates for 2022–2024 use the priority-source GDP panel; earlier estimates for 2015–2021 are unchanged. Blue and orange denote EOG and Black Marble. Shaded backgrounds distinguish primary (2015–2019), stress (2020–2021), and extension (2022–2024) domains. The figure shows point estimates; available historical stand-alone loss intervals and extension source sensitivities are reported in the Supplementary Information.

Across the 30 historical stand-alone product–year comparisons with B1–B3, six point estimates favor light and no reported province-block interval lies wholly in that direction. These overlapping comparisons summarize stand-alone transfer rather than 30 independent tests or conditional incremental value.

Primary forward validation: conditional incremental value

In the primary 2016–2019 conditional exercise, all eight pooled incremental-skill point estimates are below zero (−0.0362 to −0.0016), but every corresponding province-block mean-loss interval includes zero. The tested specifications thus show no detectable positive incremental value from adding light to historical GDP information. Table 2 reports the primary results alongside the separately evaluated later domains. The corresponding primary mean-loss intervals are reported in Supplementary Table S14.

2020–2021 temporal stress

The 2020–2021 stress test is evaluated separately from the later extension on 296 and 295 prefectures, respectively, within the defined study sample. Pooled conditional skills B0–B3 are 0.0636, 0.0418, 0.0336, 0.0352 for EOG and 0.1105, 0.0573, 0.0371, 0.0484 for Black Marble. All sixteen product–benchmark–year cells are positive. These point estimates describe the two observed target years and do not identify a causal pandemic mechanism or uncertainty across temporal regimes.

2022–2024 extension

On common 2022–2024 support, priority-source pooled conditional skills B0–B3 are 0.1097, 0.0766, 0.0671 and 0.0319 for EOG and −0.2240, −0.1379, −0.1127 and −0.1759 for Black Marble. Fifteen of 18 stand-alone B1–B3 cells have positive point-estimate skill, but this does not establish conditional gains. EOG B3 has an equal-year mean of −0.0225 despite positive pooled skill; its 2023 and 2024 annual skills are negative. Black Marble is negative in 2022 and 2024 and positive in 2023 for every conditional benchmark. The pooled cross-series divergence is concentrated in 2022: omitting that year from evaluation shifts Black Marble's pooled B0–B3 skills to −0.0441, −0.0072, −0.0007 and +0.0475, respectively; in 2023, paired loss contrasts favor Black Marble for B1–B3 (Supplementary Tables S32B, S54 and S55). These estimates describe observed differences between the two VIIRS-derived series, not the effect of any identified processing step; annual point signs should not be read as uncertainty intervals.

GDP outcome-source validation

The compiled GDP workbook does not retain complete row-level official-source links. Among 482 overlapping city-years with independently retrieved official growth data, exact agreement is 68.880%, agreement within ±0.10 percentage points (pp) is 93.154%, and RMSE is 0.552 pp. The mean signed difference is 0.016 pp, although larger discrepancies occur in parts of Shaanxi and Heilongjiang. Supplementary Appendix I reports the complete concordance and source-assessment diagnostics.

A verification sample selected before source retrieval covers 105 of the 406 workbook-only observations from 36 cities in 12 provinces across four regions and all three extension years. Official-source values were retrieved for 98 sampled observations, including 80 from primary official publication surfaces and 18 from traceable reproductions; seven could not be verified. Among verified rows, exact agreement is 85.714%, agreement within ±0.10 pp is 97.959%, and RMSE is 0.046 pp. The sample is clustered and not nationally probability-weighted, so these concordance rates are not assumed to apply to all remaining compiled entries (Supplementary Appendix I).

Processing-product disagreement under common support

We evaluate four GDP-source definitions: the priority-source panel; an all-workbook panel for all 888 extension rows; an official-only subset of 482 outcomes assessed using priority-source fits; and a verified-enhanced panel substituting 108 additional official values while retaining 298 compiled entries. On the three full-support panels, pooled conditional skill is positive for the EOG-derived series and negative for Black Marble across B0–B3. The official-only subset also changes the set of evaluated prefectures and is therefore not a like-for-like provenance comparison. Full source-defined results are reported in Table 3 and Supplementary Appendix J.

Within the observed 2022–2024 extension, single-province deletions do not change the pooled product-specific signs. EOG B0–B2 remain positive after deleting any one target year; Black Marble B0–B1 remain negative, while the all-workbook B2 comparison reaches 0.0012 in one deletion. Both B3 deletion ranges cross zero. Province-block intervals favor EOG M1 over M0 for B0–B2, but the EOG–B3 interval includes zero; Black Marble mean-loss intervals favor M0 for every benchmark. With identical GDP targets and common M0 predictions, the paired EOG-minus-Black-Marble loss contrast equals the difference in incremental losses. In the verified-enhanced B0 comparison, the mean paired contrast is -1.893×10^{-4} (95% interval $[-3.645, -0.877] \times 10^{-4}$); paired intervals remain negative across B0–B3 under the priority-

source, all-workbook and verified-enhanced definitions. These are cross-processing paired loss contrasts, not independent-sensor validation or general product rankings.

The cross-series comparison uses common prefectures, GDP targets, and history-only predictions, while each light series retains its documented annual inputs and processing support. Constant numerical rescaling cancels in log differences, but this property alone does not establish that annual radiance changes are equivalent across the two product lineages. Supplementary Tables S39–S40 document why other archived light measures are not used as independent corroboration.

Supporting spatial and robustness diagnostics

The historical 272-prefecture primary sensitivity retains negative stand-alone mean B0 skill and eight negative pooled conditional skills. Removing the 2017 evaluation fold changes EOG–B0 pooled skill to 0.0053 but leaves its equal-fold mean negative. These historical checks address the pre-COVID domain (Supplementary Tables S28A–S28D). The expanded extension comparison instead distinguishes test-sample composition from updates to recursive history and residual calibration; it does not refit the historical primary evidence on a new selected subset.

For the 2022–2024 B0 conditional comparison, initial partial-support fits evaluated on the original annual support yield pooled skill 0.1269 for EOG and −0.1955 for Black Marble. Evaluating those same fits on expanded support gives 0.1168 and −0.2448; updating histories and calibration on expanded support gives 0.1097 and −0.2240. Thus expanding the evaluation support changes magnitudes without generating the opposite-product B0 signs. A full two-fit-by-four-support comparison covers original annual support, the earlier common 148 cities, expanded common support and expanded annual support (Supplementary Tables S35A–S35B and S38). The last two coincide at complete coverage. These are descriptive comparisons, not causal composition effects.

Geography of transfer error and ranking performance

In the 2015–2019 comparison, the two products identify a very similar geography of realized prediction difficulty. Their benchmark-relative B0 loss maps correlate at Pearson $r = 0.835$ and Spearman $\rho = 0.759$ (Fig. 4; Supplementary Table S9). On exactly common prefecture-year rows, raw standardized squared-error maps correlate at Pearson $r = 0.9921$ (province-block 95% CI 0.9731–0.9970) and Spearman $\rho = 0.9349$. After residualizing both benchmark-relative maps on the shared B0 error surface, the partial Pearson correlation remains 0.8516. Raw product errors are also spatially clustered for both products: Moran's I is 0.6285 (Queen) and 0.6447 (4-nearest-neighbor) for EOG and 0.6146 and 0.6386 for Black Marble, with permutation $p = 0.0001$ throughout (Supplementary Table S15). In contrast, B0-relative excess-loss clustering is robust for EOG across Queen and 4NN weights (Moran's I = 0.3077–0.3294; $p = 0.0001$), but weak and weight-sensitive for Black Marble (Moran's I = 0.0507–0.0641; $p = 0.0423$–0.0827; Supplementary Table S9). Because both products share the VIIRS/Suomi-NPP sensor family and the same realized GDP target, this agreement indicates similar realized error geography across processing pipelines, not independent-sensor validation.

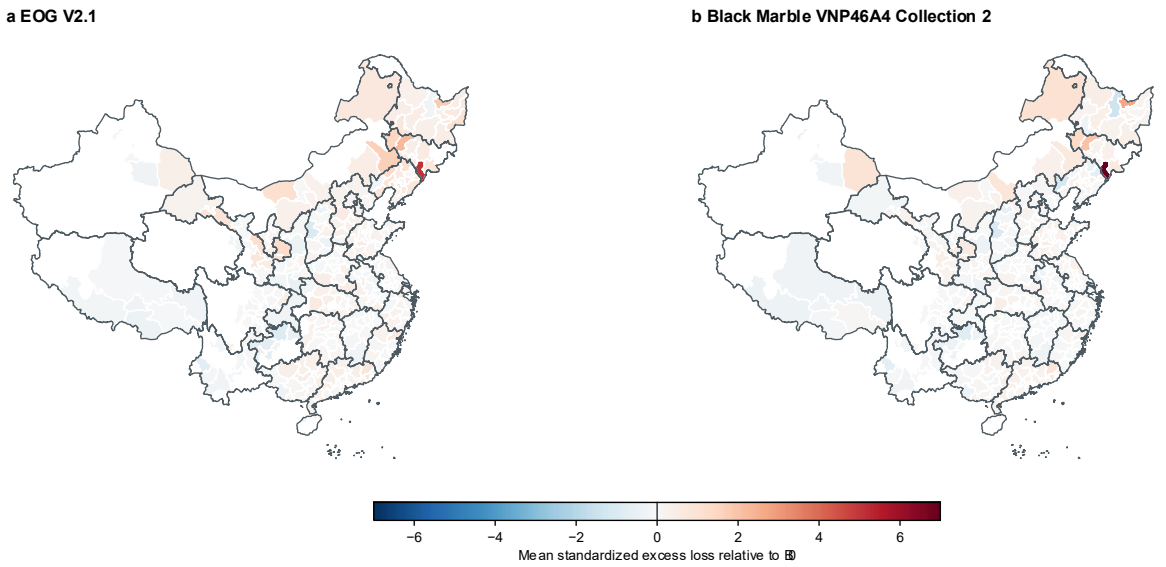


Figure 4. Geography of continuous pre-COVID temporal-transfer loss relative to the national training-period mean. Each panel maps a prefecture's mean standardized excess squared loss averaged across its available transfer folds in 2015–2019; negative values favor the stand-alone light mapping and positive values favor B0. Both panels use the same continuous scale centered at zero and the accepted current-boundary set. Areas outside the fixed 296-prefecture universe are unfilled; near-white shades within included prefectures indicate values close to zero on the diverging scale. The mapped quantity is comparative benchmark-relative performance, not sensor error alone.

Growth-extreme recovery in the pre-COVID ranking exercise is less stable. At the 20% threshold used in the historical analysis, light-based high- and low-growth Jaccard values range from 0.103 to 0.234 for EOG and from 0.054 to 0.261 for Black Marble. Eleven of 20 product-year-tail cells lie above the 95th percentile of the random-ranking reference. By comparison, B2 and B3 Jaccard ranges are 0.295–0.611 and 0.341–0.589, and all 20 historical-ranking cells lie above the same reference. Spatial-HAC inference further limits the robustness claim. EOG stand-alone excess loss relative to B0 and B3 remains positive at 200, 500, and 800 km, whereas Black Marble is more cutoff-sensitive. Every conditional M1–M0 spatial-HAC interval includes zero for both products and the two primary benchmarks (Supplementary Tables S16–S17).

Figure 5 summarizes the annual conditional sequence on the retained priority-source panel, separating the three observed temporal domains. Outcome-source sensitivities and extension uncertainty are reported separately in Table 3 and Supplementary Appendix J.

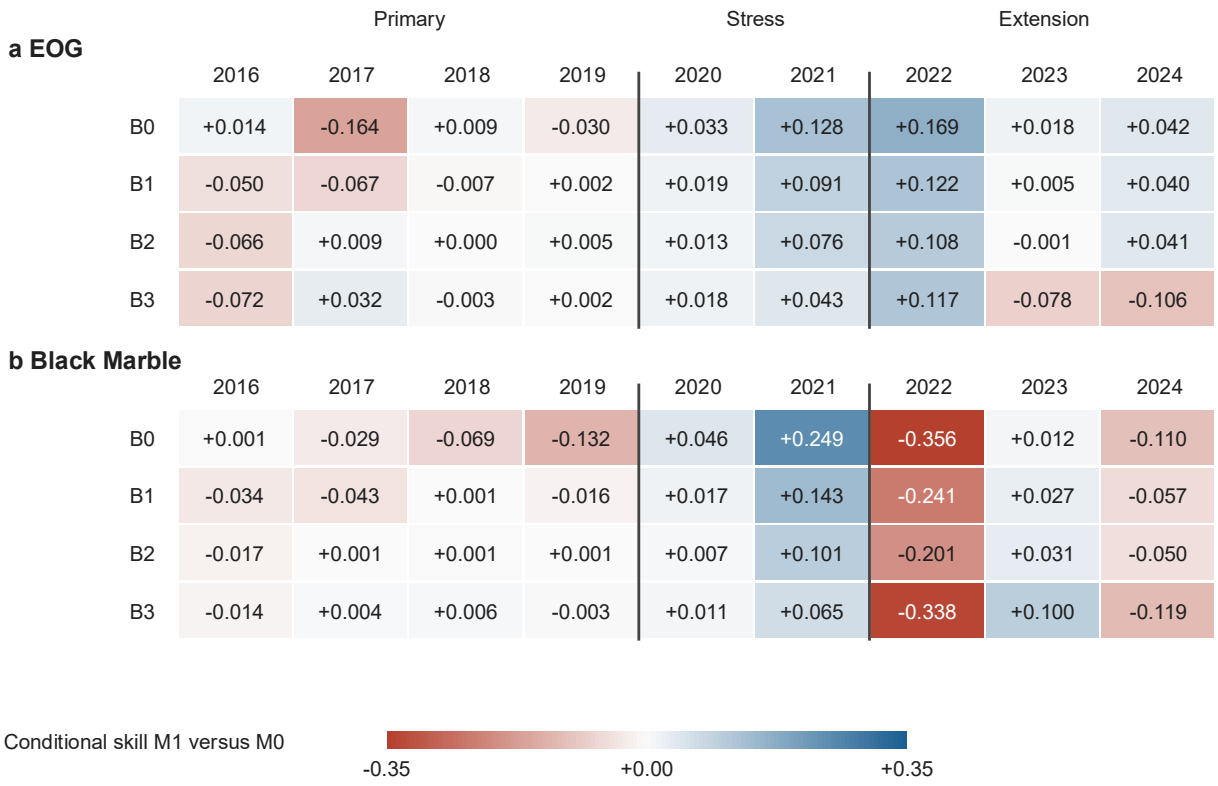


Figure 5. Annual conditional incremental skill on the retained priority-source panel. Each cell is 1 − SSE(M1)/SSE(M0), comparing light-augmented history with intercept-recalibrated history alone. Blue favors M1 and red favors M0. Lines separate primary 2016–2019, stress 2020–2021 and extension 2022–2024. Only preceding GDP outcomes enter calibration and benchmark histories. Extension targets share 296-prefecture support. These are point estimates; outcome-source comparisons and loss intervals are reported in Table 3 and Supplementary Appendix J.

Pre-COVID model-form and benchmark checks do not reverse the pre-COVID result. Five-fold average stand-alone B0 and B3 skill remains negative for linear, quadratic, and fixed four-degree-of-freedom natural-spline forms for both products, and the mean-fold conditional incremental-skill range remains below zero across these fixed forms. A pooled AR(1) historical benchmark is feasible from 2016 onward; adding light improves three of eight product-year cells and worsens five of eight. Exploratory moderator diagnostics based on strictly prior-year light and GDP-history characteristics remain imprecise and do not identify an actionable applicability threshold (Supplementary Tables S18–S25).

Discussion

Strong spatial alignment does not establish forward annual-growth measurement value. In 2016–2019, the conditional tests provide no detectable positive increment from the specified light-augmented models beyond the historical GDP benchmarks. Positive conditional point estimates occur in the separate 2020–2021 stress tests. In 2022–2024, pooled conditional performance differs between the EOG-derived and Black Marble series on common prefecture and GDP outcomes, although the annual comparisons are heterogeneous and the EOG B3 interval includes zero. The findings are therefore conditional on the evaluated period, benchmark, and light series rather than evidence of a persistent processing mechanism.

These results help reconcile a literature containing both strong and weak temporal applications of nighttime lights. Regional and small-area light–GDP relationships vary across settings [24, 25], Chinese provincial associations vary over time [26], and recent cross-country subnational evidence reports much stronger relationships in levels than in growth [16]. National forecasting exercises can nevertheless obtain gains from nighttime lights [13], especially when signal construction or predictive systems are more flexible [6, 7, 14]. Our design evaluates a narrower claim: whether a historical scalar light-to-GDP mapping carried forward to an observed light signal reduces loss relative to explicit prior-year GDP benchmarks, and

whether light adds information after conditioning on those benchmarks. Performance for this task need not determine performance for tuned multivariate or machine-learning systems.

The geographic diagnostics add a separate dimension. EOG and Black Marble identify very similar raw error geographies even though benchmark-relative spatial inference is more product-sensitive. Because both products share the VIIRS/Suomi-NPP sensor family, this is cross-processing agreement rather than independent-sensor replication. The clustering of raw errors and the sensitivity of some benchmark-relative comparisons to spatial-HAC cutoffs indicate that transfer difficulty is geographically structured and cannot be reduced to a conventional standard-error choice.

The comparisons distinguish temporal transfer from geographic coverage, GDP-source concordance, and differences between nighttime-light series. Evaluating both products on the same 296-prefecture extension sample reduces geographic-composition differences in their performance comparison. Source-defined outcome substitutions produce similar pooled estimates on the complete panels, although they do not establish the accuracy of every compiled GDP observation. These dimensions should be considered separately when interpreting the observed cross-series results.

Three limitations define the scope of the extension. First, the EOG-derived series uses a separately maintained continuation after 2021, and Black Marble changes its common-pixel support in the extension. The observed cross-series differences therefore cannot be uniquely attributed to a processing step. Second, part of the 2022–2024 GDP panel uses compiled observations with incomplete row-level provenance; verification and source-substitution exercises assess sensitivity but do not establish equivalence for every city-year. Third, province-block intervals are conditional on the three observed extension years and do not quantify uncertainty over future temporal regimes or GDP-source error. The results concern the specified scalar models, observed target-year light, and current-boundary prefectures; they do not evaluate every possible predictive model.

The broader implication is methodological. Strong spatial fit need not imply that a historically calibrated light signal improves annual local-growth measurement beyond admissible GDP history. Under the tested scalar specifications, the pre-COVID conditional comparisons show no detectable positive incremental gain, while the later comparisons vary across observed periods, benchmarks, and VIIRS-derived series. These results motivate validation that matches the intended outcome, available information, temporal domain, and spatial support; they do not establish a general product ranking or identify the cause of cross-series differences.

Methods

Data sources and study sample

The fixed analysis universe comprises 296 prefectures across all 31 mainland provincial-level units; Sansha remains outside the accepted GDP-growth sample. Annual light levels cover 2013–2024 and GDP-growth observations cover 2014–2024, with all GDP and light values through 2021 unchanged. The 2022–2024 priority-source panel includes 452 previously selected official observations, 30 additional official Shaanxi observations and 406 compiled observations. In the verified-enhanced panel, 108 compiled observations are replaced by inspected official values: 98 from the predeclared verification sample and ten additional Jilin observations. The all-workbook sensitivity uses workbook outcomes throughout the extension; the official-only sensitivity evaluates official outcomes using priority-source fits. All complete extension panels cover 296 prefectures annually, with historical GDP gaps, city codes and boundaries unchanged (Supplementary Tables S5, S29 and S50).

The initial official-source subset of 148 common prefectures was compared with the remaining study units using standardized mean differences for pre-extension economic and nighttime-light characteristics. Because the full extension includes all 296 prefectures in the fixed analysis universe, included-versus-excluded balance statistics are not defined for that panel. The original diagnostics, variable definitions and coverage checks appear in Supplementary Table S37 and Supplementary Fig. S7. No matching, weighting, selection regression or covariate imputation was introduced.

Nighttime-light products and preprocessing

EOG Annual VNL V2.1 observations cover 2013–2021 [27, 28]. For 2022–2024, we use a separately maintained VNL-derived Cloud-Optimized GeoTIFF continuation (OpenGeoHub Zenodo record 17294744, version v0.4) [29], excluding its extrapolated 2000–2011 values. The continuation is not assumed to be equivalent to the official EOG V2.1 processing chain. We aggregate non-negative radiance within the same current prefecture boundaries (all_touched=True), using the common valid-pixel intersection for 2013–2024. Historical area-weighting and area-residualization diagnostics retain their original grid and support.

NASA Black Marble VNP46A4 Collection 2 [30, 31] is extended through 2024 using processed annual prefecture aggregates. The extension retains finite, non-negative radiance from AllAngle_Composite_Snow_Free, quality code 0, more than three contributing observations, land/water-mask codes 0/1/2/5, and the primary 0.50 nW cm^{-2} sr^{-1} residual-background threshold. Prefecture boundaries and all_touched=True rasterization are unchanged. The continuation uses common admissible-pixel support for 2013–2024, rather than the 2013–2021 support of the historical aggregate. Full-period support holds pixel coverage fixed within the continuation but may use imagery observed after an earlier target year; this choice is distinct from GDP calibration timing. File identities, hashes, screening rules and support definitions are documented in Supplementary Tables S2 and S30.

To align numerical scales across historical and continuation aggregates, the default bridge combines an initial source-based factor with the median positive matched old-to-new ratio over 2,664 city-years in 2013–2021. For EOG, the metadata-based factor of 0.1 and an overlap median ratio of 10 yield a net factor of 1.0; Black Marble has an initial factor of 1 and a net factor of 1.00075615393726. These factors are used to align aggregate numerical scales; they do not establish physical radiance equivalence between source segments. Annual extension log changes use adjacent years from the same continuation: the 2022 change, for example, uses continuation values for both 2021 and 2022, with no cross-segment fallback. Constant multiplicative scaling cancels in log differences. An alternative log-linear bridge is examined only as a sensitivity analysis.

Annual light growth is

$$\Delta \ln L_{it} = \ln(L_{it}) - \ln(L_{i,t-1}) \quad (1)$$

where L_{it} is total light in prefecture i and year t. The analysis uses only positive totals. The first observable level year supplies the lag and does not enter as a growth observation.

GDP sources and provenance

The historical GDP panel combines nominal prefecture GDP levels for within-year level comparisons with source-certified real GDP growth transformed as ln(1 + growth/100); the mixed-source extension is addressed below. GDP values are certified by city and year under a fixed source hierarchy and reference-vintage rule. In the five original level-comparison years, all 1,452 observations have a selected official value: 1,442 agree exactly with the recorded values and eight agree within the published precision. Two 2015 Heilongjiang observations remain outside that precision and are retained without ex post reconciliation. Supplementary Appendix A reports these exceptions and the corresponding substitution sensitivity.

Historical prefecture GDP sources include the China City Statistical Yearbook (中国城市统计年鉴; relevant editions through 2025), compiled by the Urban Socio-Economic Survey Department of the National Bureau of Statistics, and municipal and provincial statistical-bureau publications. For ten Shaanxi cities, we use the comparable-price table in the 2025 Baoji statistical yearbook. Part of the 2022–2024 extension relies on a compiled 296-prefecture workbook supplied by Markdata as an intermediate extraction and harmonization file, not as an original official source; its SHA-256 is retained. We interpret its growth index as g = index − 100 and transform growth as ln(index/100). The available metadata do not identify an official source for every row or establish that all indices share a consistent prior-year-equals-100 real-growth definition. Transformations are not adjusted to improve agreement. Independently retrieved official values take

precedence over differing workbook values in the priority-source and verified-enhanced panels. Missing GDP observations are not interpolated, and no GDP outcomes are derived from light. Supplementary Appendices I and J document source-definition limitations, official-source comparisons and the independently selected verification sample.

Spatial representation and annual-change alignment

Within each certified level year, we compute the squared correlation between log total light and log nominal GDP. We then compute the analogous squared correlation between annual log light change and certified real GDP growth. These comparisons are diagnostic rather than structural fits: the level margin is influenced by city size and built-up extent, and nominal levels and real growth are different economic measurement concepts. Area residualization provides the principal support adjustment for the level margin. The purpose is to test whether strong spatial ranking persists when attention shifts to annual change, not to attribute the full difference in R^2 to differencing alone.

Stand-alone forward temporal transfer

For each product and each target year from 2015 through 2024, pooled OLS is fitted on all admissible observed rows from earlier GDP years. The estimated coefficients are held fixed when applied to observed target-year light. No target-year GDP enters calibration. B0 is the national training mean, B1 the province training mean, B2 the prefecture training mean, and B3 the most recent observed prefecture growth. The benchmark fallback hierarchy follows the documented rules. GDP outcome timing, rather than publication vintage, determines admissible history. Both products use the same available GDP target rows within year.

We report RMSE, MAE, mean error, SSE, test-mean out-of-sample R^2, skill 1 − SSE(light)/SSE(benchmark) and mean ΔLoss. Positive skill and negative ΔLoss favor light. Pooled loss ratios, equal-year means and annual signs are separate. Primary stand-alone targets are 2015–2019, stress targets 2020–2021 and extension targets 2022–2024. The priority-source panel remains the fixed reference; the all-workbook and verified-enhanced panels are refitted on the same geographic support, while the official-only sensitivity evaluates retained official targets using priority-source fits rather than a separately refitted incomplete-history model. No new stand-alone inferential claim is made from point signs.

Conditional incremental validation

The conditional analysis begins in target year 2016 and recursively reconstructs each historical benchmark for every training row using GDP information indexed to years preceding that row year; this design enforces GDP-outcome-year separation rather than a real-time publication-vintage constraint. This creates an outcome-leakage-restricted history-only prediction for each training observation. For benchmark j, the training residual is

$$r_{j,is} = y_{is} - B_{j,is}^{(s-1)} \tag{2}$$

Here, s denotes an admissible training-observation year (s < t); the benchmark for that row uses GDP outcomes strictly preceding s. For light product p, the baseline residual model M0 estimates only a training-period intercept, while M1 adds the contemporaneous light-growth signal from the same admissible training rows:

$$M_0: r_{j,is} = \alpha_{j,t} + \varepsilon_{is}; \quad M_1: r_{j,is} = \alpha_{jp,t} + \gamma_{jp,t}\, \Delta \ln L_{is}^{(p)} + \varepsilon_{is} \tag{3}$$

For target year t, both residual models are added back to the target-year historical benchmark. Conditional incremental skill compares M1 with M0:

$$IncSkill = 1 - SSE(M_1) \,/\, SSE(M_0) \tag{4}$$

Historical conditional uncertainty uses 9,999 province-panel draws (seed 20260916), with its intervals kept separate from extension results. For the 2022–2024 extension, each of 9,999 bootstrap draws (seed

20261002) samples 31 provinces with replacement, retains their complete histories and observed extension targets, and refits recursive fold-specific M0 and M1. Province multiplicities are paired across light products and the priority-source, all-workbook and verified-enhanced panels. Percentile 95% intervals cover mean squared-log-growth M1-minus-M0 loss and paired cross-processing loss, rather than skill ratios. They quantify spatial-cluster uncertainty conditional on the three observed extension years, not uncertainty across temporal regimes.

Spatial diagnostics, ranking, and dependence

For the primary error geography, B0-relative excess squared loss is standardized by target-year GDP-growth variance and averaged within prefecture across the available pre-COVID folds in 2015–2019. Raw product-specific squared-error maps are also constructed on exactly common prefecture-year rows. Pearson and Spearman correlations quantify cross-product geographic agreement; a partial Pearson correlation residualizes both B0-relative loss maps on the shared B0 error surface. Spatial dependence is evaluated with a fixed Queen contiguity graph and a four-nearest-neighbor great-circle sensitivity, each using 9,999 permutation draws.

The retained pre-COVID applied ranking exercise defines actual and proxy-estimated high- and low-growth prefectures separately within each target year at a 20% threshold specified in the historical analysis, with 10% and 25% sensitivities. Precision, recall, F1, Jaccard overlap, and Spearman rank correlation are reported for EOG, Black Marble, B2, and B3. A 9,999-draw random-ranking reference preserves target-year sample size and set cardinality. Cross-border spatial dependence in the primary loss differential is assessed with Bartlett-kernel spatial-HAC/Conley inference [32] at 200, 500, and 800 km cutoffs specified in the historical analysis. Pooled-panel inference allows arbitrary within-prefecture serial covariance and same-year cross-prefecture covariance within the cutoff.

Sensitivity analyses and temporal domains

Historical model-form, AR(1), ranking, spatial-HAC and moderator diagnostics retain their original scope. The sample-expansion analysis crosses the initial and updated fits/histories with four evaluation supports: original annual support, the initial 148-city common support, expanded common support and expanded annual support. Initial coefficients are applied to additional targets without using their GDP in fitting. Comparing supports at fixed fits describes sample-composition differences, whereas updating histories and calibration at fixed support describes changes associated with refitting; neither contrast identifies a causal composition effect. Leave-one-year-out sensitivity recomputes pooled evaluation losses without refitting later models and is applied to the primary and extension domains, which have at least three target years, but not the two-year stress period. DMSP is not concatenated with VIIRS, and supplementary light archives are screened for empirical eligibility.

GDP source sensitivity and independent verification

To assess GDP-source sensitivity, compiled growth outcomes are compared with independently retrieved official growth on 482 overlapping city-years. A 105-row workbook-only verification sample was selected before independent source retrieval using city clusters stratified by region, economy size, and workbook-reported growth; 98 selected rows could be verified and seven remain unresolved. The verified-enhanced panel substitutes all 108 additional inspected official observations, whereas verification agreement rates use only the 98 verified sampled rows. The source-defined model refits, target substitutions, and paired-loss comparisons follow the preceding validation procedures. Detailed provenance, sample-selection rules, and audit results appear in Supplementary Appendices I–J.

Computational environment and reproducibility

The public deposit contains statistical code, aggregate outputs, Table 3, input fingerprints and environment documentation. Supplementary Data 1 provides additional machine-readable aggregate results. Source-assessment records, verification evidence and study-specific processing records are retained by the authors; access to restricted third-party inputs follows the conditions in Data availability.

Use of generative AI

During the preparation of this manuscript, the authors used ChatGPT 5.6 (OpenAI) to improve readability and language. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the final text.

Code availability

Data and aggregate results are available in the GitHub release (URL: https://github.com/tanbole0214/Temporal-Validation-of-VIIRS-Nighttime-Lights/), with additional aggregate outputs in Supplementary Data. The public archive does not include prefecture-year GDP observations due to licensing restrictions.

Data availability

Aggregate results and replication materials are available in the GitHub repository cited above and Supplementary Data 1. The underlying GDP data originate from the*China City Statistical Yearbook*and municipal and provincial statistical publications, with a third-party compiled extract used for part of the 2022–2024 panel. Restricted city-year inputs are not publicly redistributed and remain subject to the relevant data providers' access conditions. Nighttime-light and boundary-data sources are described in Methods.

References

1. Donaldson, D. & Storeygard, A. The view from above: Applications of satellite data in economics. *J. Econ. Perspect.* **30,** 171–198 (2016). https://doi.org/10.1257/jep.30.4.171
2. Levin, N. et al. Remote sensing of night lights: A review and an outlook for the future. *Remote Sens. Environ.* **237,** 111443 (2020). https://doi.org/10.1016/j.rse.2019.111443
3. Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *Am. Econ. Rev.* **102,** 994–1028 (2012). https://doi.org/10.1257/aer.102.2.994
4. Hu, Y. & Yao, J. Illuminating economic growth. *J. Econom.* **228,** 359–378 (2022). https://doi.org/10.1016/j.jeconom.2021.05.007
5. Martínez, L. R. How much should we trust the dictator's GDP growth estimates? *J. Polit. Econ.* **130,** 2731–2769 (2022). https://doi.org/10.1086/720458
6. Zhang, H. et al. Developing an annual global sub-national scale economic data from 1992 to 2021 using nighttime lights and deep learning. *Int. J. Appl. Earth Obs. Geoinf.* **133,** 104086 (2024). https://doi.org/10.1016/j.jag.2024.104086
7. Gong, W., Chen, Z., Wang, C., Zhang, L. & Yu, B. Estimating monthly gross value of industrial outputs at the pixel level with nighttime lights image products and a spatial-analysis-based CNN model. *Int. J. Appl. Earth Obs. Geoinf.* **143,** 104781 (2025). https://doi.org/10.1016/j.jag.2025.104781
8. Gibson, J., Olivia, S., Boe-Gibson, G. & Li, C. Which night lights data should we use in economics, and where? *J. Dev. Econ.* **149,** 102602 (2021). https://doi.org/10.1016/j.jdeveco.2020.102602
9. Bluhm, R. & McCord, G. C. What can we learn from nighttime lights for small geographies? Measurement errors and heterogeneous elasticities. *Remote Sens.* **14,** 1190 (2022). https://doi.org/10.3390/rs14051190
10. Bluhm, R. & Krause, M. Top lights: Bright cities and their contribution to economic development. *J. Dev. Econ.* **157,** 102880 (2022). https://doi.org/10.1016/j.jdeveco.2022.102880
11. Gibson, J. & Boe-Gibson, G. Nighttime lights and county-level economic activity in the United States: 2001 to 2019. *Remote Sens.* **13,** 2741 (2021). https://doi.org/10.3390/rs13142741
12. Zhang, X. & Gibson, J. Using multi-source nighttime lights data to proxy for county-level economic activity in China from 2012 to 2019. *Remote Sens.* **14,** 1282 (2022). https://doi.org/10.3390/rs14051282
13. Galimberti, J. K. Forecasting GDP growth from outer space. *Oxf. Bull. Econ. Stat.* **82,** 697–722 (2020). https://doi.org/10.1111/obes.12361
14. Chiovelli, G., Michalopoulos, S., Papaioannou, E. & Regan, T. Illuminating the Global South. *Econ. J.* ueaf134 (2026). https://doi.org/10.1093/ej/ueaf134

15. Cui, Y. et al. A global continuous 500 m nighttime light dataset (1992–2024) via NDVI-guided DMSP-OLS correction and U-TransNet cross-sensor harmonization. *Int. J. Appl. Earth Obs. Geoinf.* **148,** 105230 (2026). https://doi.org/10.1016/j.jag.2026.105230

16. Mendez-Guerra, C. & Qian, J. Harmonized nightlights and local economic activity: Better focus on levels, not growth. Preprint at SSRN https://doi.org/10.2139/ssrn.7299423 (2026).

17. Gotway, C. A. & Young, L. J. Combining incompatible spatial data. *J. Am. Stat. Assoc.* **97,** 632–648 (2002). https://doi.org/10.1198/016214502760047140

18. Openshaw, S. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography No. 38 (Geo Books, Norwich, 1984).

19. Ma, Y., Chen, S., Ermon, S. & Lobell, D. B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **301,** 113924 (2024). https://doi.org/10.1016/j.rse.2023.113924

20. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12,** 1620–1633 (2021). https://doi.org/10.1111/2041-210X.13650

21. Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40,** 913–929 (2017). https://doi.org/10.1111/ecog.02881

22. Fiaschi, D., Parenti, A. & Ricci, C. Aggregation bias in proxy measurement: Nighttime lights and local economic activity. Preprint at https://doi.org/10.48550/arXiv.2607.14825 (2026).

23. Kim, B., Gibson, J. & Boe-Gibson, G. Measurement errors in popular night lights data may bias estimated impacts of economic sanctions: Evidence from closing the Kaesong Industrial Zone. *Econ. Inq.* **62,** 375–389 (2024). https://doi.org/10.1111/ecin.13183

24. Bickenbach, F., Bode, E., Nunnenkamp, P. & Söder, M. Night lights and regional GDP. *Rev. World Econ.* **152,** 425–447 (2016). https://doi.org/10.1007/s10290-016-0246-0

25. Fontaine, I., Hermet, F. & Lucic, N. Are nightlight data a relevant proxy for economic activity in small island developing states? *Econ. Lett.* **253,** 112375 (2025). https://doi.org/10.1016/j.econlet.2025.112375

26. Glawe, L. & Mendez, C. Harmonized luminosity and economic activity across provinces in China: Cross-sectional differences, regional time series, and inequality dynamics. *Appl. Econ.* **57,** 10677–10693 (2025). https://doi.org/10.1080/00036846.2024.2439583

27. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. & Ghosh, T. VIIRS night-time lights. *Int. J. Remote Sens.* **38,** 5860–5879 (2017). https://doi.org/10.1080/01431161.2017.1342050

28. Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F.-C. & Taneja, J. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sens.* **13,** 922 (2021). https://doi.org/10.3390/rs13050922

29. Hengl, T. Annual time series of global VIIRS nighttime lights for 2000–2024 at 500-m spatial resolution extrapolated using logistic regression. *Zenodo*, version v0.4 (2025). https://doi.org/10.5281/zenodo.17294744

30. NASA VIIRS Land Science Investigator-led Processing System. VIIRS/NPP Lunar BRDF-Adjusted Nighttime Lights Yearly L3 Global 15 arc second Linear Lat Lon Grid, VNP46A4 Collection 2. *NASA LAADS DAAC* (2025). https://doi.org/10.5067/VIIRS/VNP46A4.002

31. Román, M. O. et al. NASA's Black Marble nighttime lights product suite. *Remote Sens. Environ.* **210,** 113–143 (2018). https://doi.org/10.1016/j.rse.2018.03.017

32. Conley, T. G. GMM estimation with cross sectional dependence. *J. Econom.* **92,** 1–45 (1999). https://doi.org/10.1016/S0304-4076(98)00084-0

Author contributions

Bole Tan: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing.

Xiaoyu Jiang: Conceptualization, Formal analysis, Investigation, Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing.

Xin Dai: Investigation, Data curation, Validation, Project administration, Writing – review & editing.

Zhong Li: Conceptualization, Supervision, Project administration, Validation, Writing – review & editing.

Yangbo Liu: Methodology, Data curation, Validation, Visualization, Writing – review & editing.

Additional Information

Supplementary Information

Supplementary Information and Supplementary Data 1 accompany this manuscript as separate files.

Competing interests

The authors declare no competing interests.

Funding

The authors received no funding for this research.

Tables

Table 1. Annual forward temporal-transfer validation against B0 through 2024

Product	Train → test	N	RMSE	Skill vs B0	OOS R^2 vs test mean
EOG	2014→2015	288	0.0322	−0.1362	−0.2104
EOG	2014–15→2016	288	0.0365	−0.0807	−0.1564
EOG	2014–16→2017	282	0.0283	−0.6606	−0.7216
EOG	2014–17→2018	294	0.0190	−0.0155	−0.1113
EOG	2014–18→2019	295	0.0205	−0.0547	−0.2965
EOG	2014–19→2020	296	0.0539	0.0058	−1.8093
EOG	2014–20→2021	295	0.0238	0.0816	−0.2071
EOG	2014–21→2022	296	0.0375	0.0584	−1.9337
EOG	2014–22→2023	296	0.0214	0.0178	−0.2078
EOG	2014–23→2024	296	0.0209	0.0621	−0.3239
Black Marble	2014→2015	288	0.0282	0.1287	0.0718
Black Marble	2014–15→2016	288	0.0353	−0.0064	−0.0770
Black Marble	2014–16→2017	282	0.0248	−0.2699	−0.3166
Black Marble	2014–17→2018	294	0.0203	−0.1673	−0.2774
Black Marble	2014–18→2019	295	0.0213	−0.1311	−0.3904
Black Marble	2014–19→2020	296	0.0536	0.0176	−1.7759
Black Marble	2014–20→2021	295	0.0225	0.1757	−0.0834
Black Marble	2014–21→2022	296	0.0424	−0.2003	−2.7394
Black Marble	2014–22→2023	296	0.0221	−0.0461	−0.2863
Black Marble	2014–23→2024	296	0.0207	0.0863	−0.2898

Note: B0 is the national training-period mean. Target-year GDP is excluded from calibration and target-year light is observed. Skill = 1 − SSE(light)/SSE(B0). Test-mean out-of-sample R^2 uses the realized target-year mean only for evaluation. Extension rows use the priority-source panel; source-defined reruns are in Supplementary Table S52A. Historical intervals remain in Supplementary Table S7.

Table 2. Conditional incremental skill by product benchmark and temporal domain

Product	B	Primary 2016–19 pooled	Primary 2016–19 equal	Stress 2020–21 pooled	Stress 2020–21 equal	Extension 2022–24 pooled	Extension 2022–24 equal
EOG	B0	−0.0296	−0.0426	0.0636	0.0808	0.1097	0.0763
EOG	B1	−0.0362	−0.0305	0.0418	0.0549	0.0766	0.0555
EOG	B2	−0.0168	−0.0130	0.0336	0.0448	0.0671	0.0493
EOG	B3	−0.0075	−0.0103	0.0352	0.0304	0.0319	−0.0225
Black Marble	B0	−0.0347	−0.0572	0.1105	0.1474	−0.2240	−0.1513
Black Marble	B1	−0.0259	−0.0228	0.0573	0.0801	−0.1379	−0.0901
Black Marble	B2	−0.0047	−0.0036	0.0371	0.0540	−0.1127	−0.0734
Black Marble	B3	−0.0016	−0.0016	0.0484	0.0378	−0.1759	−0.1191

Note: M1 adds observed target-year light to recursively constructed, intercept-recalibrated M0. Primary = 2016–2019; stress = 2020–2021; extension = 2022–2024. Pooled skill is 1 − pooled SSE(M1)/pooled SSE(M0); equal is the unweighted annual mean. The domains contain 1,159, 591 and 888 rows per product–benchmark cell. Extension entries use the priority-source panel on common 296-prefecture support. Table 3 reports outcome-source robustness and extension mean-loss intervals; historical intervals are retained separately.

Table 3. Outcome-source robustness and conditional loss uncertainty, 2022–2024

Product	B	Priority-source skill	Verified-panel skill	Mean ΔLoss $\times 10^{-4}$	95% interval $\times 10^{-4}$
EOG	B0	0.1097	0.1097	−0.623	[−1.065, −0.275]
EOG	B1	0.0766	0.0766	−0.498	[−0.878, −0.211]
EOG	B2	0.0671	0.0671	−0.444	[−0.807, −0.180]
EOG	B3	0.0319	0.0318	−0.443	[−1.004, 0.041]
Black Marble	B0	−0.2240	−0.2239	1.271	[0.447, 2.915]
Black Marble	B1	−0.1379	−0.1378	0.896	[0.360, 1.940]
Black Marble	B2	−0.1127	−0.1126	0.745	[0.291, 1.707]
Black Marble	B3	−0.1759	−0.1760	2.450	[0.851, 5.921]

Note: The priority-source and verified-enhanced panels each contain 888 rows on the same 296-prefecture support. Skill = 1 − SSE(M1)/SSE(M0). ΔLoss compares M1 with M0; negative values favor adding light. The single mean ΔLoss and 95% interval columns report the verified-enhanced panel (P4); priority-source (P1) intervals are provided in Supplementary Table S53A. Intervals use 9,999 province-panel refits conditional on the observed three target years and do not quantify uncertainty over temporal regimes. All-workbook and official-only sensitivities are reported in Supplementary Appendix J. The EOG B3 interval includes zero.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 02SupplementaryInformationScientificReports20261015.docx